The Meaning *of* Structure

Marijn Koolen

The Value of Link Evidence for Information Retrieval

The Meaning of Structure

the Value of Link Evidence for Information Retrieval

ILLC Dissertation Series DS-2011-03

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

The Meaning of Structure
the Value of Link Evidence for Information Retrieval

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op vrijdag 15 april 2011, te 14.00 uur

door

Marinus Henricus Aloysius Koolen

geboren te Hoorn.

Promotiecommissie:

Promotor: Prof.dr. J.S. MacKenzie-Owen
Co-promotor: Dr.ir. J. Kamps

Overige commissieleden:
Dr. N.E. Craswell
Prof.dr. M. Lalmas
Dr. M.J. Marx
Prof.dr. M. de Rijke
Prof.dr.ir. A.P. de Vries

Faculteit der Geestewetenschappen
Universiteit van Amsterdam

NWO
Netherlands Organisation for Scientific Research

The work in this thesis has been funded by the Netherlands Organization for Scientific Research (NWO) in the project Multipe-collection Searching Using Metadata (MuSeUM), grant number 640.001.501, which is part of the Continuous Access To Cultural Heritage (CATCH) research programme.

SIKS Dissertation Series 2011-15

SIKS

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover design by Donald Roos.

Printed and bound by Ipskamp Drukkers B.V.

Published by IR Publications, Amsterdam

ISBN: 978-90-814485-5-0

ACKNOWLEDGMENTS

I start this book with thanking people who helped me in the past five years in starting and finishing this thesis.

First and foremost, thank you Jaap, for all your advice, help, inspiration and enthusiasm through the years. It is incredible how often I managed to lose track of your great advisory one-liners and get stuck in trivial matters. It is equally incredible how you kept your calm every time I did, and helped my get back to the more interesting avenues. I'm also very grateful you offered me a Ph.D. position and not the position for scientific programmer that you brought to my attention several times. Given my programming skills, I think you made the right choice.

I thank John MacKenzie-Owens and my thesis committee for reading and approving my thesis and giving encouraging and helpful comments and suggestions. I'm grateful to the INEX community for providing a friendly surroundings in which I could safely make my first scientific blunders, and for creating the resources on which a large part of this research is based.

Thanks to Avi for inspiring discussions about theory and modelling in IR, pointing out many flaws in my mathematical work and providing the right material to mend it. Thanks to Tikitu for explaining the importance of typography and making me care about it, and helping me make this thesis look much better than my initial, uninformed and messy attempt. Thanks to Donald for designing the cover of this thesis and for further explaining the importance and beauty of typography. I am sure there are still plenty of typographic horrors to be found in here, for which I take full responsibility. I look forward to further education.

Frans, I've thoroughly enjoyed the countless useful discussions on IR, speech segmentation of infants, writing papers, writing a thesis, doing science, and I even vaguely recall some discussions on music. Also, thanks for proofreading the Dutch summary and giving tips, which greatly improved it.

Anna, thank you for very many things, including proofreading this thesis in exchange for taking you out for dinner a few times. I owe you plenty more.

ABSTRACT

How can search engines use the hyperlinks between documents to determine which documents are the most relevant for a search query? Some search engines use links to determine popularity, where the underlying idea is that the number of links pointing to a document (Web page) is a measure of its popularity. Search results are ordered or ranked based on their popularity and on similarity between the content of the document and the content of the search query. Another aspect of links is they provide a signal that two documents have related content. After all, a link is a reference. If a document A is relevant for a search query, then documents linked to A are possibly also relevant. Link information could possibly contain evidence for the *topical relevance* of a document.

This thesis describes an investigation of the value of those two aspects of link information—popularity and topical relevance—for ranking search results. This question has been addressed before in the field of *information retrieval*. Starting in the late nineties, researchers conducted large scale experiments to see if link information could help search engines to find as many relevant documents on a search topic as possible, and rank these documents as well as possible. The results were disappointing. No consistent improvements by incorporating link information in the search process could be reported. Representatives of search engine companies argued that users of Web search engines are not looking for as many relevant documents on a topic as possible, but for the home pages of specific Web sites and pages that provide a good starting point to further explore Web pages on a certain topic. The researchers changed their attention to these Web-centric search tasks, with immediate success. The home pages and other important pages that Web searchers are looking for tend to be popular pages, which can be more easily identified by using link information. The value of link information for information retrieval seemed clear: link information is useful for measuring popularity, but not for measuring the topical relevance of documents.

The question of why links are not useful for measuring topical relevance was never answered. The goal of this thesis is to give a more precise and complete account of the value of link evidence for information retrieval. This is first investigated using the English *Wikipedia*,

because it is obtainable in its entirety, including all the links between the Wikipedia articles. Because it is an encyclopedia, Wikipedia is a natural source for users to search for articles relevant to a certain topic, which makes it an appropriate starting point to measure link evidence for topical relevance. The findings are then validated on a much larger corpus of Web pages, to find out how they generalise to searching on the Web.

Evidence for popularity can best be measured by using all the links on the whole Web, and counting how many point to each Web page. We call this *global* link information, derived from the *global* link structure. For popularity, the direction of the link is important; a link from A to B makes B popular, but not A. The page with the most incoming links is the most popular. The order in which pages are ranked is determined partly by their popularity and partly by how well their content matches that of the search query.

Evidence for topical relevance can be derived from link information by first finding a list of Web pages in which the search terms frequently occur, and then using only the links between those pages. This way, links are selected based on a topic: the topic of the search query. This list of pages is called the set of *local* pages, and the links between those pages are called *local* links. For topical relevance, the direction of the link is not important; if page A discusses the same topic(s) as page B, then page B discusses the same topic(s) as page A. Search terms that occur in page A form text evidence that A is topically relevant to the search query. If the search terms occur frequently in page A as well as in page B, then a link from A to B is a signal that the text evidence for page A is also evidence for the topical relevance of B. Vice versa, the text evidence for B is also evidence for the topical relevance of A. The number of local links between A and other local pages represents the amount of evidence for A. More links between A and other local pages means more evidence for the topical relevance of A. The page with the most local links is considered the most relevant.

Although links are considered to be a signal that two linked pages have topically related content, this relation is not the same for all pairs of linked pages. To measure how strong the topical relation between two pages is, we use the category information in Wikipedia. In Wikipedia, the value of links for measuring topical relevance is dependent on the relation between the pages they link. A link between pages about the same topic is more effective evidence for topical relevance of those pages than a link between pages about unrelated topics. This finding

confirms that, in Wikipedia, link information can be used as evidence for the topical relevance of a page.

From these findings, we draw a number of conclusions. Global link information is independent of the search query and provides evidence for popularity, but not for topical relevance. Local link information is dependent on the search query and can provide evidence for topical relevance. For global link information, the direction determines its meaning. For local link information, the direction of the links has no impact on the relation with topical relevance.

Support for these conclusions is found in Wikipedia, which further clarifies the relation between local link information and topical relevance. First, the amount of local link evidence is related to the amount of relevant text in a document, regardless of the direction of the links. Second, the fraction of global links that is present in the local link structure (again, regardless of the direction of the links) is related to how specifically a document is about the search topic.

With these findings, the value of link information for ranking search results has become clearer and more complete. Link information can be evidence for both popularity and topical relevance. The meaning of information derived from the link structure is determined by the direction of the links, the topical relation between the linked documents and the selection of links that is used as evidence.

SAMENVATTING

Hoe kunnen zoekmachines de hyperlinks tussen documenten gebruiken om te bepalen welke documenten het meest relevant zijn voor een zoekvraag? Sommige zoekmachines gebruiken links om *populariteit* te meten, waarbij het onderliggende idee is dat een document (Webpagina) waar veel links naartoe wijzen populair is. Zoekresultaten worden geordend op basis van hun populariteit en op basis van de overeenkomst tussen de inhoud van een document en de zoekvraag. Een ander aspect van links is dat ze aangeven dat twee pagina's inhoudelijk iets met elkaar te maken hebben. Een hyperlink is tenslotte een verwijzing. Als pagina A relevant is voor een zoekvraag, dan zijn pagina's die gelinkt zijn aan A wellicht ook relevant. Linkinformatie bevat dus mogelijk bewijs voor de *inhoudelijke relevantie* van een document.

Dit proefschrift beschrijft onderzoek naar de waarde van die twee aspecten van linkinformatie—populariteit en inhoudelijke relevantie—voor het ordenen van zoekresultaten. Deze vraag werd eerder onderzocht binnen het vakgebied *information retrieval*. Vanaf eind jaren '90 werd op grote schaal geëxperimenteerd met het toepassen van linkinformatie om zoveel mogelijk relevante documenten te vinden voor een onderwerp en om deze documenten zo goed mogelijk te ordenen. De resultaten waren teleurstellend. Er waren geen consistente verbeteringen te meten door linkinformatie mee te nemen in het zoekproces. Vertegenwoordigers van zoekbedrijven gaven aan dat gebruikers van Webzoekmachines niet op zoek zijn naar zoveel mogelijk relevante informatie, maar naar de homepagina's van specifieke Websites, en pagina's die een goed startpunt vormen voor het verkennen van Webpagina's over een onderwerp. De onderzoekers besloten daarom hun aandacht te verschuiven naar deze Webspecifieke zoektaken, en hadden meteen succes. De homepagina's en andere belangrijke pagina's waar veel naar gezocht wordt zijn populaire pagina's, die met behulp van linkinformatie makkelijker te identificeren zijn. Hiermee leek de waarde van linkinformatie voor information retrieval duidelijk. Linkinformatie is nuttig voor het meten van populariteit, maar niet voor het meten van inhoudelijke relevantie.

De vraag waarom links niet nuttig zijn voor het meten van inhoudelijke relevantie werd nooit duidelijk beantwoord. Het doel van dit proefschrift is om de waarde van linkinformatie voor information re-

trieval preciezer en vollediger in kaart te brengen. Dit wordt eerst onderzocht met de Engelse *Wikipedia*, omdat deze in zijn geheel beschikbaar is, inclusief alle links tussen de Wikipediapagina's. Vanwege de encyclopedische aard is het zoeken naar informatie over een onderwerp in Wikipedia een natuurlijke taak. Dit maakt Wikipedia een geschikt startpunt voor het meten van linkbewijs voor inhoudelijke relevantie. De bevindingen worden vervolgens getoetst op een veel grotere collectie van Webpagina's, om vast te stellen in hoeverre zij generaliseerbaar zijn naar zoeken in het Web.

Bewijs voor populariteit kun je het best meten door alle links op het hele web te gebruiken en te tellen hoeveel er naar elke pagina gaan. Dit noemen we *globale* linkinformatie, afgeleid uit de *globale* linkstructuur. Voor populariteit is de richting van de link belangrijk; een link van A naar B maakt B populair, maar niet A. De pagina met de meeste links is het populairst. De volgorde waarin je de resultaten ordent wordt gedeeltelijk bepaald door de populariteit en gedeeltelijk door hoe goed de inhoud van een documenten overeen komt met de zoekvraag.

Bewijs voor inhoudelijke relevantie is af te leiden uit linkinformatie door eerst een lijst pagina's te zoeken waar de zoekwoorden vaak in voorkomen, en vervolgens alleen de links tussen die pagina's te gebruiken. Zo selecteer je links op een bepaald onderwerp: het onderwerp van je zoekvraag. Die lijst van pagina's noemen we de *lokale* pagina's en de links tussen die pagina's noemen we *lokale* links. Voor de inhoudsrelatie maakt de richting van de link niet uit; als A over hetzelfde gaat als B, dan gaat B ook over hetzelfde als A. Zoekwoorden die in pagina A voorkomen vormen tekstueel bewijsmateriaal dat pagina A relevant is voor de zoekvraag. Als de zoekwoorden vaak voorkomen in zowel pagina A als pagina B, dan is een link van A naar B een signaal dat het tekstuele bewijs voor A ook bewijs is voor de relevantie van B. Andersom zegt het tekstuele bewijs voor B ook iets over de relevantie van A. Het aantal lokale links tussen pagina A en andere lokale pagina's geeft aan hoeveel bewijs er voor A is. Hoe meer van die lokale links pagina A heeft, hoe meer bewijs er is voor de relevantie van A. De pagina met de meeste lokale links wordt beschouwd als het meest relevant.

Hoewel links een signaal geven dat twee gelinkte pagina's inhoudelijk aan elkaar gelateerd zijn, is die inhoudelijke relatie niet altijd even sterk. Om te meten hoe sterk twee pagina's inhoudelijk aan elkaar gerelateerd zijn gebruiken we de categie-informatie in Wikipedia. In Wikipedia blijkt de waarde van linkinformatie voor inhoudelijke relevantie afhankelijk te zijn van de relatie tussen twee gelinkte pagina's.

Een link tussen twee pagina's die over hetzelfde onderwerp gaan is effectiever bewijs voor de inhoudelijke relevantie van die pagina's dan een link tussen twee pagina's die over verschillende onderwerpen gaan. Deze bevinding bevestigd dat in Wikipedia linkinformatie als bewijs kan dienen voor de inhoudelijke relevantie van een pagina.

Uit deze bevindingen worden een aantal conclusies getrokken. Globale linkinformatie is onafhankelijk van de zoekvraag en geeft bewijs voor belangrijkheid (populariteit, authoriteit) maar niet voor inhoudelijke relevantie. Lokale linkinformatie is wel afhankelijk van de zoekvraag, en geeft bewijs voor inhoudelijke relevantie. Voor globale linkinformatie is de richting bepalend voor de betekenis ervan. Voor lokale linkfinformatie is de richting van minder belang.

Verdere ondersteuning voor deze conclusies vinden we in Wikipedia, waarbij de relatie tussen lokale linkinformatie en inhoudelijke relevantie verder verduidelijkt wordt. Ten eerste blijkt de hoeveelheid aan lokaal linkbewijs gerelateerd aan de hoeveelheid relevante informatie in een document, ongeacht de richting van de links. Ten tweede blijkt de fractie van het globale aantal links dat aanwezig is in de lokale linkstructuur (ongeacht de richting van de links) gerelateerd aan hoe specifiek het document over het zoekonderwerp gaat.

Hiermee is het antwoord op de vraag wat de waarde van linkinformatie voor het ordenen van zoekresultaten is, duidelijker en vollediger geworden. Linkinformatie kan bewijs vormen voor zowel populariteit als inhoudelijke relevantie. De betekenis van linkinformatie wordt bepaald door de richting van de links, de inhoudelijke relatie tussen de gelinkte documenten, en de selectie van links die worden gebruikt als bewijs.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part i

Introduction

1 INTRODUCTION

One of the most prominent characteristics of the World Wide Web (Web) is the ubiquity of hyperlinks. Each document in the Web can link to other documents using hyperlinks—like references in books or scientific articles—which are active links that allow the user to go straight from the source document to the documents that are referenced by the hyperlinks. Looking at the Web as a whole, the documents and hyperlinks form a huge interconnected network of data, in which the documents are the nodes of the network and the hyperlinks are connections that can be made between any of the nodes to form a trail of related information (see Figure 1). Researchers quickly realised that analysis of the structure of links that emerged as the Web evolved could provide valuable insights into how information on the Web is organised. Commercial search engine companies have heralded the use of link structure as one of their key technologies. How link information can be of value for information retrieval is an important question that merits investigation.

The hyperlinks on the Web seem invaluable for information access. Many popular search engines use hyperlinks to crawl the Web and discover new pages. Web surfers use them to navigate to the information they are looking for. Many Web pages have very little text, so search engines need other features to distinguish between billions of Web pages. Link information about a document can be directly observed in the link structure, such as the number of links originating from or pointing to the document. We can use this link information to derive other characteristics of the document, such as popularity or authority. Link *information* can thus serve as *evidence* for aspects of a document that cannot be directly observed. For instance, link information can provide evidence about the popularity of pages by considering the number of links pointing to those pages, and textual descriptions through the anchor text associated with links. On top of that, the context of a page can be considered by looking at the pages that are connected to it, which help interpreting the content of a page.

The value of link evidence for information retrieval (IR) is a large open problem. At least two aspects of links have been investigated. First, the fact that the author of a document can only link to documents that

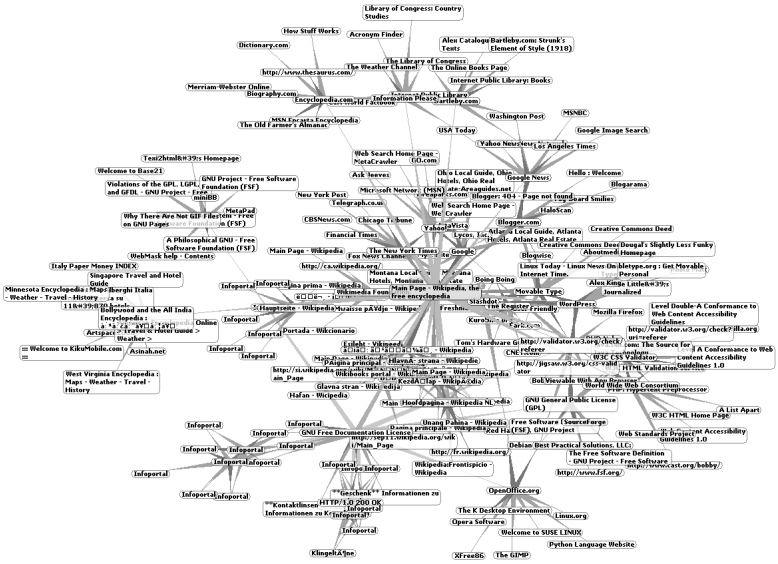

Figure 1: Hyperlink structure of a sample of Web pages (source: `http://worddork.blogspot.com/2010/01/hyperlink-addiction.html`).

he or she knows to exist. The number of references to a document are used to quantify how well-known a particular document is. Interpreting a link as an author's statement that the linked document is worthwhile, the number of links to a document can then be seen as a measure of how important or useful a document is. Well-known algorithms like PageRank (Page et al., 1998) and HITS (Kleinberg, 1999) use the link structure between documents to derive the importance or authority of each document, which is similar to ideas of status and prestige in social network analysis (Wasserman and Faust, 1994).

The second aspect is the reason that an author references another document. We assume that the author refers to other documents because they are in some way related to the content of the document that the author is writing. An example is given in Figure 2, which shows part of a Web page about the architecture of Robert Hooke. The underlined parts of the text represent hyperlinks to other Web pages about the architecture of Hooke. If this Web page is returned as a search result in response to the query *Robert Hooke architecture* because it contains all the query terms, the hyperlinks might be a signal that the referenced pages are also relevant to the query. IR researchers have tried to use this

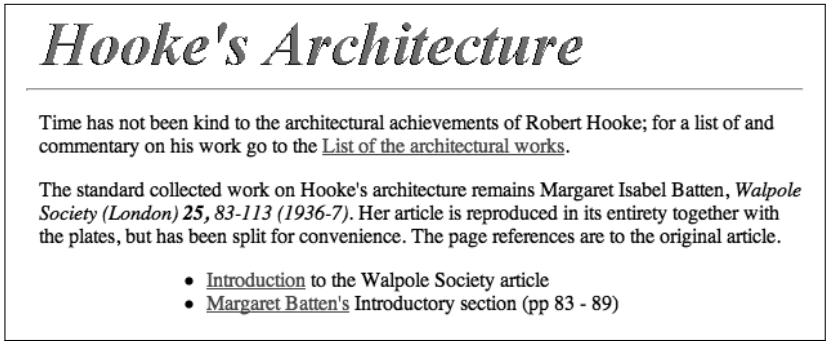

Figure 2: Example of hyperlinks in a Web page (source: `http://www.roberthooke.org.uk/arch1.htm`).

semantic aspect of links to find other documents related to the same topic.

What is the practical use of analysing the nature of links? If we have a better understanding of how link information affects retrieval performance, we can determine when and how to use it as evidence, whether to filter out low-quality search results or to identify the most topically related result or perhaps for something completely different. Search engine companies can use link information as evidence to improve the quality of search results, especially in very large collections with millions or billions of documents varying strongly in quality, distinguishing important entry pages and high-quality information on reliable Web sites from thousands or millions of uninteresting, low-quality pages. Many Web pages have very little text, which makes it hard for search engines to determine the topical relevance of such pages for the typically short queries posed to Web search engines. A short page might be easier to interpret if we take into account the pages it links to and the pages that link to it. Link information provides more context to determine the topical relevance of a short page.

1.1 THE VALUE OF LINK INFORMATION FOR IR

The question about the value of link evidence for retrieval can be viewed from a practical and a scientific perspective. With the rapid growth of the World Wide Web, IR researchers at TREC (Text REtrieval Conference, TREC 2009) thought hyperlinks would be a useful feature of Web pages to improve retrieval algorithms for Web search (Hawking and Craswell, 2005, p. 6). After all, search engine companies claimed

to use link information to help rank results, and were producing good results. The value of hyperlinks for information retrieval became one of the main points on the scientific agenda of the TREC Web Track. The assumption that hyperlinks would be beneficial for retrieval was then tested on a sample of Web data using the standard ad hoc retrieval methodology. This methodology was developed around the notion of a human searcher having a fairly precisely defined information need and a desire to find all documents relevant to this information need. A test collection was created with a set of information needs and relevance judgements based on the traditional assumptions that 1) a user wants to read text relevant to the topic of their information need and 2) the relevance of a document is based on its textual content alone.

Despite high expectations, the TREC experiments failed to establish the effectiveness of link evidence for general ad hoc retrieval (Hawking, 2001, Kraaij and Westerveld, 2001). The question about the value of link information was still unanswered.

Several internet search engine experts observed that on the Web, typical search is different from ad hoc search (Hawking and Craswell, 2005). Unlike the strict TREC Ad Hoc definition of relevance given above, they argued that "Web searchers typically prefer the entry page of a well-known topical site to an isolated piece of text, no matter how relevant" (Hawking and Craswell, 2005). Consider a Web searcher typing the query "Mercedes-Benz". The assumption in ad hoc retrieval is that the user is looking for text about Mercedes-Benz, such as an historical overview of the company or news articles about its financial situation. The entry page of the company Web site might be a typical portal with very little textual information, and therefore considered irrelevant according the above assumption, but for many Web users it might be more appropriate than any financial news items. As a consequence, the evaluation methodology of the Web retrieval task was changed towards more Web-centric tasks like home page and named page finding and topic distillation. Here, link information was a highly beneficial feature, as links often point to home pages of Web sites and other important pages within sites. From a practical perspective, it showed the value of links for actual Web search.

The value of link structure to find important or authoritative documents is well established. With the positive results of using link information for the new Web-oriented search tasks, the discrepancy between what search engine companies claimed—that link information is useful for ranking Web results—and what the TREC Ad Hoc participants found—that link information does not improve the ranking of ad hoc

search results—seemed resolved. As a consequence, the investigation of the value of link evidence for ad hoc retrieval was quickly abandoned. However, from a scientific perspective, the value of link information for IR remains an open issue.

Why is link evidence effective for typical Web search tasks but not for ad hoc retrieval? Is it because document importance is not useful for finding topically relevant text? Are the current link-based ranking methods not suitable to derive the semantic aspect of links? Or is link evidence strongly correlated to content evidence and therefore has nothing to add to content evidence? Are the links in the Web too heterogeneous and noisy to effectively derive useful semantic information? Is the Web link graph too sparse to usefully distinguish between relevant and non-relevant pages? The link sparseness issue has been addressed with some success by Gurrin and Smeaton (2004), but the document collection they created is a very small artificial subset of a larger Web collection and the improvements due to link-based methods are also very small. As part of a research agenda to properly study links for ad hoc retrieval, they present a list of prerequisites that a Web collection must satisfy.

Thus, in the Web, the value of links as indicators of document importance is well established, but their value as indicators of topical relevance is not clear.

One of the problems of Web retrieval evaluation is the vast size of the Web and the large amount of resources required to process the data. The test collections of the TREC Web Tracks of 1999-2004 have been criticised for being too small, and unrepresentative of the Web. Because these collections are only small samples of the entire Web, the link structures of these collections are incomplete. We cannot study links from Web pages outside those collections. Because of this, there are many of the above mentioned questions that we cannot answer.

A more controlled experiment can be conducted on Wikipedia. Wikipedia is a free Web encyclopedia that is collaboratively edited by countless individuals around the globe and presents an interesting case to study topical aspects of link information. It is a single Web domain with encyclopedic articles that are densely interlinked and is available in its entirety, including all hyperlinks. If we focus on Wikipedia, we can do a more thorough analysis of the link topology, because we have all the link information. Moreover, there are extensive and high-quality test collections from the INEX Ad Hoc Tracks (INEX, 2009). These test collections consist of large number of topics and relevance judgements on the English Wikipedia and allow a detailed study of the relation

between links and relevance. Because Wikipedia is part of the Web, general principles of link topology should hold in Wikipedia as well. However, Wikipedia links might be a special case and some aspects of the Wikipedia link graph might not hold for the Web, but simply be artefacts of Wikipedia. Therefore, any findings about the nature of links in Wikipedia should be validated on the larger Web to establish whether they are artefacts of Wikipedia or general aspects of hyperlinks.

Wikipedia is an important resource in its own right, so any finding can provide valuable insight. For several years now, it has consistently been one of the most popular Web sites on the internet[1] and one of the most important knowledge bases. On top of that, it is a natural resource for informational search, with both content and links being created, modified and removed in a collaborative fashion by millions of contributors around the world.

There are possible disadvantages of using Wikipedia to study hyperlinks in general. There are many aspects that might make Wikipedia very different from the Web. Wikipedia is much smaller than the Web and arguably suffers less from spam. Each Wikipedia page can be edited by anyone while on most Web sites, pages can often only be edited by a handful of people who maintain the site. Wikipedia also has guidelines on how and what information to add to Wikipedia, and when and how to create links. As an encyclopedia, its articles are written in an objective style, with little redundancy of information between articles. On top of that, the context in which users search Wikipedia and the Web might be radically different.

Some of these differences between Wikipedia and the Web could cause their link structures to be different, and might help us understand when and why link evidence is useful. There is a guideline stating that links between Wikipedia articles should only be created when they are relevant to the context (Wikipedia, 2010). There are so-called bots—small computer programs that automatically edit Wikipedia pages to conform to certain style guidelines—that automatically insert links serving a particular purpose (for instance, all dates are linked for presentational purposes). These processes are different from those that lead to the creation of hyperlinks on the Web. Whereas in Web documents an author can arbitrarily link his page to any other page, whether there is a topical relation or not, in Wikipedia links tend to be semantic: a link from page A to page B shows that page B is

1 At the time of writing (September 2010), around 13% of global internet users visit Wikipedia per day, and it is the sixth most popular site according to Alexa, (`http://www.alexa.com/siteinfo/wikipedia.org`).

semantically related to (part of) the content of page *A*. Arguably, there will be some fraction of links that do not denote an important topical relation between pages, and not all links will be equally meaningful in all search contexts, such as links to dates created by bots. But the linking guidelines provide a mechanism that results in links that are relevant to the context. Furthermore, its topical organisation makes it clear what information is there, and where to link to, further suggesting the special nature of links in Wikipedia.

From analysing Wikipedia links in an information retrieval context and comparing them to general Web links, we might be able to gain valuable insight into the nature of hyperlinks in general. Therefore, the main research question of this thesis is:

- What is the value of link evidence for information retrieval?

Because we want to study links for information retrieval, we have to work with IR test collections. We compare the INEX Wikipedia collection against and validate our findings on TREC Web collections. Although it is hard to establish how representative these collections are for the Web at large, they are the best publicly available Web test collections.

Of course, there are many different ways to look at link structures. To guide our investigation, we focus our work on addressing a number of questions based on intuition and the previous experience of others. Based on earlier findings described above, we can break down the main question into several more specific questions.

1.2 RESEARCH QUESTIONS

The more specific research questions can be bundled into four groups:

1. Links for Wikipedia and Web retrieval: Links in Wikipedia might differ from general Web hyperlinks in certain characteristics. Their impact on retrieval might be different.

- Can link information in Wikipedia be used as evidence to improve the ranking of ad hoc retrieval results?
- Is the value of links in Wikipedia different from their value in the Web?

2. Global and local link evidence: Link information can be derived from the entire link graph of the collection, or from a subset of query-dependent retrieval results.

- How is global, query-independent link evidence related to relevance?

- How is local, query-dependent link evidence related to relevance?

3. Importance and topical relevance: Links can be used as indicators of popularity or importance of documents, or as indicators of how topically relevant linked documents are to the search topic.

- Is link evidence for document importance useful for ranking ad hoc retrieval results?
- Is link evidence for topical relevance useful for ranking ad hoc retrieval results?

4. Quantity and semantic relatedness: The information conveyed by links is affected by the quantity of links and the semantic relatedness of linked documents.

- What is the impact of link density or link quantity on the value of link evidence?
- How does the semantic relatedness of linked documents affect the value of link evidence?

1.3 STRUCTURE AND OUTLINE OF THIS THESIS

To study the value of link evidence for information retrieval in general and answer the questions above, we need a test collection that allows a detailed analysis of the relation between links and relevance. This requires a document collection with a dense link graph and a semantic categorisation of the documents to study the semantic relatedness of linked documents. On top of that, to study the relation with relevance, we need a set of search requests and associated relevance judgements. At the time of writing, no such a collection of Web pages exists that is representative of the Web in general and meets these requirements. The collection best meeting these criteria is the INEX 2006 Wikipedia collection. Of course, Wikipedia might be different from the Web in general, which is why we validate our findings on a recently created Web test collection. Unlike the INEX Wikipedia collection, this new Web collection does not have the detailed information on which parts of a document are relevant, nor a fine-grained category structure to which the documents are assigned. However, we can validate our findings on the impact of global and local link evidence, document importance and link density. This thesis consists of five parts.

Part I: Introduction

This chapter and the next on related work form the introduction to the research problem addressed in this thesis.

Chapter 2: Related work
This chapter provides an overview of research on information retrieval in general, the analysis of link structure, and on how links have been used in information retrieval and more specifically in Web retrieval. Various link-based ranking and propagation algorithms are discussed, as well as the first large scale evaluation of link information for Web retrieval at TREC. Furthermore, the notion of relevance is discussed, as well as the distinction between the traditional ad hoc retrieval task and more Web-oriented search tasks.

Part II: The importance of link evidence in Wikipedia

In this part we look at the impact of link evidence in Wikipedia and compare that against its impact on a well-studied Web test collection used for the TREC Web tracks. We look at the relation between link degrees and relevance and the impact of query-independent and query-dependent link evidence. Our findings help towards answering sets 1 and 2 of the research questions.

Chapter 3: Link evidence for Wikipedia ad hoc retrieval
Can we use link information in Wikipedia as an indicator of topical relevance? The methodology, data and experimental set-up are described. We use the INEX Wikipedia collection and a large collection of ad hoc topics and relevance judgements to conduct experiments. Because we want to understand what meaningful information we can derive from structure, we look at the link degrees, that is, the number of links incident with each document, and consider the link structure on a global level—using all the links in the entire collection—and on a local level—using only the links between the documents retrieved for a given topic. Our main findings are that incoming link evidence in Wikipedia can improve retrieval performance. Documents with a higher in-degree have a higher probability of being relevant. However, using the global number of incoming links to re-rank documents is not as effective as local link evidence for improving the document ranking of a content-based approach. Local link evidence keeps much more focus on the topic at hand and leads to significant improvements over a text-retrieval baseline. The work in this chapter is based on Kamps and Koolen (2008).

Chapter 4: Wikipedia and Web link structure
In the Web, link evidence is an indicator of document importance. It helps Web-oriented tasks by identifying site entry pages and other

important or authoritative pages, but not ad hoc search tasks. On Wikipedia, it is effective for ad hoc retrieval. This difference leads us to investigate if and how Wikipedia link structure differs from the link structure in the larger Web and whether this affects the impact of link information on retrieval performance. Experiments are conducted on the INEX Wikipedia collection and the .GOV collection and the TREC 2004 Web Track topics. Our main findings are that, structurally, Wikipedia links are fairly similar to general Web links. The main difference is that in the Web, incoming links are more related to relevance than outgoing links, while in Wikipedia, there is little difference between incoming and outgoing links. Global link evidence is more effective for Web-centric tasks, while local evidence is more effective for ad hoc retrieval. The work in this chapter is based on Kamps and Koolen (2009).

Part III: The nature of link evidence

In this part we present a deeper analysis of the nature of link evidence. We use the detailed relevance information of INEX Wikipedia test collections—which tells us how much of the text of a document is relevant—to study the relation between query-independent and query-dependent link evidence on the one hand and document importance and topical relevance on the other hand. We also look at the impact of the density of the link graph by filtering links in various ways and use the Wikipedia category structure to see how the semantic relatedness of documents affects the impact of link evidence. Our findings help answer sets 2, 3 and 4 of the research questions.

Chapter 5: From document importance to topical relevance
To what extent are links in Wikipedia related to document importance or topical relevance? Here we take a closer look at the relation between query-dependent and query-independent link evidence and topical relevance. We study the overlap and differences between degrees of incoming, outgoing and undirected links and how they are related to the amount of relevant text in documents. Our main findings are that within the set of documents retrieved for a given query, the in- and out-degrees are more strongly correlated to each other than over the entire collection. However, over the documents with the highest degrees this correlation is substantially lower, indicating that in-degree and out-degree do promote different documents. All link degrees show a clear relation with the amount of relevant text. Documents with the highest local degrees tend to be the documents with the most relevant text. The work in this chapter is based on Koolen and Kamps (2009).

Chapter 6: Link evidence and semantic relatedness
Local link evidence can be used as an indicator of topical relevance while global link evidence seems ineffective, even though the links are all derived from the same link graph. This shows that not all links are equally effective. Links in the local graph seem better indicators of the semantic relatedness of linked documents than the links in the global graph. But the quantity of links must also play a role. With fewer links we have less information to distinguish between documents. In this chapter we want find out which links are effective. We use the category structure in Wikipedia to measure the semantic relatedness between linked articles and filter out less semantic links to study the trade-off between the quantity and semantic nature of links. We observe that local links are more semantic than global links, and that global link evidence cannot be made more effective by filtering out the less semantic links. Our main findings are that semantic relatedness determines the effectiveness of link evidence for ad hoc search. Links between semantically related documents are more effective than links between unrelated ones. The work in this chapter is based on Koolen and Kamps (2011).

Part IV: Generalising to the Web

In this part we test our findings from Wikipedia on the Web. In the course of writing this thesis, a new Web test collection became available through the TREC 2009 Web Track. We take advantage of this opportunity to see which aspects of link evidence in Wikipedia are aspects of hyperlinks in general and which aspects are particular for Wikipedia. Our findings help answer sets 1 and 2 of the research questions.

Chapter 7: From Wikipedia to the Web
A new, high-quality Web retrieval test collection is being developed, which should be a much better representation of the Web than earlier collections. We take advantage of this opportunity and use the first set of evaluation data to draw tentative conclusions on how our findings about link evidence in Wikipedia generalise to the larger Web. Our main findings are threefold. First, in the new Web collection, link evidence can improve ad hoc retrieval performance. Second, the presence of Wikipedia in the new Web test collection changes the nature of the collection and the impact of link evidence. Third, in the non-Wikipedia part the impact of link evidence is similar to the impact of link evidence in the Wikipedia part. Only when we combine Wikipedia with the rest of the Web, the special nature of Wikipedia means that global link

evidence becomes more effective because it promotes Wikipedia pages. The work in this chapter is partly based on Koolen and Kamps (2010).

Part V: Conclusions

In the final part of this thesis we draw conclusions.

Chapter 8: Conclusions
In this final chapter we describe the contribution of this thesis by addressing the main research questions of each chapter, and summarise the findings. We draw conclusions on the value of link evidence for information retrieval and discuss future research.

1.4 A NOTE ON TERMINOLOGY

One point on terminology. The IR research community typically speaks about *documents* as the units that are returned as search results, whereas the Web search community conventionally speaks about *pages* as search results. Wikipedia research often uses the term *articles* to refer to the encyclopedic entries that are returned as search results. As Wikipedia is part of the Web, and each encyclopedic entry has a unique URL (Uniform Resource Locator) as identifier, the articles are also Web pages. We use these terms interchangeably to indicate the retrievable units in the collection. That is, we consider pages, articles and documents to mean the same thing.

The term *ad hoc retrieval* can be interpreted in different ways. In this thesis, we adopt the TREC interpretation of ad hoc retrieval which assumes the user is a dedicated, experienced searcher who does ad hoc searches on an archived data collection for new topics and requires high precision and high recall, and who is "willing to look at many documents ... in order to obtain high recall" (Harman, 1993). The same model is assumed for the INEX Ad Hoc test collections built from Wikipedia.

2 RELATED WORK

In this chapter we review related work for the rest of this thesis. The first section (Section 2.1) presents the field of IR in general. The succeeding sections discuss research on link information in general (Section 2.2), Web retrieval and the value of link information (Section 2.3) and Wikipedia (Section 2.4).

2.1 INFORMATION RETRIEVAL

The field of information retrieval started in answer to an explosion of available information (Bush, 1945). Early research focused on the existing classification schemes and indexing languages, and the evaluation of these schemes and languages (Robertson, 2008). With evaluation came the notion of relevance, which turned out to be a difficult concept to employ. Because it also plays an important role in this thesis, we will start with a short overview of some attempts to get to grips with relevance.

2.1.1 *Relevance*

Relevance is an important notion in information retrieval, and one that has been extensively debated and studied. Documents that are considered irrelevant in the ad hoc retrieval methodology could in fact be relevant for tasks like home page finding. If different tasks lead to different relevance judgements, they must use different interpretations of what makes a document relevant, which urges us to look at these interpretations of relevance.

Kochen (1974) distinguishes between "relevance as a relation between propositions and the recognition of relevance on its judgement by a user, which resembles a utility or significance judgement." This can be interpreted as an *objective* relevance relation and a *subjective* relevance relation respectively. Cosijn and Ingwersen (2000) distinguish five manifestations of relevance: algorithmic, topical, cognitive, situational and socio-cognitive. Saracevic (1975) describes a framework for thinking about relevance and distinguishes several different views on what relevance means. He uses the intuition that relevance has to do with

the success of the communication process and describes it as a measure of the effectiveness of the contact between a source and a destination in a communication process. Mizzaro (1998) wrote a large overview of several decades of research related to relevance.

Some interpretations of relevance will be used in this thesis:

- *Topical relevance*: Topical relevance roughly corresponds to the subject knowledge view of relevance, which describes relevance as the relation between the *subject content* of the question or information request and the existing *subject knowledge* (Saracevic, 1975). This is also closely related to the notion of *aboutness* (Hutchins, 1977). Topical relevance is independent of the system and the user.

- *System relevance*: Sometimes referred to as *algorithmic* relevance, which is a relation between "*information or information objects retrieved by the system* and *the query*" (Saracevic, 2007). "Topical relevance certainly is the basis for system or algorithmic relevance ... word-based retrieval is based on trying to establish topical relevance" (Saracevic, 2007, p. 1931).

- *User relevance*: This follows from the user context. User relevance relates to changes in the cognitive state.

- *Pertinence*: Pertinence is the relation between the subject knowledge of documents and the underlying information need. The information need involves the knowledge state of the user, which the system has no access to and can only guess at. A document can only be relevant if the user can understand the content and if it contains information that changes the knowledge state of the user.

- *Utility*: According to Cooper (1971) "Utility is a catch-all concept involving not only topic-relatedness but also quality, novelty, importance, credibility and many other things."

- *Situational relevance*: A form of logical relevance bearing on a user's individual situation and personal view (Wilson, 1973). It involves the problem at hand. It is inferred from criteria such as "usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like" (Saracevic, 2007).

In Web-centric search tasks, the assumed user model is of a user first trying to locate the entry page to a particular Web site and use the links on this entry page to navigate to pages that satisfy her information need. The entry page itself might not contain the information to satisfy

the user, but gives access to the rest of the site and allows the user to browse, representing a first step in a longer session. The relevance of entry pages is based not only on topical relevance, but also on user relevance, utility or situational relevance. The traditional ad hoc retrieval methodology of TREC treats each search result as an individual document and assumes the user only wants pages that contain the information that satisfies (part of) her information need. This seems more restricted to the notion of topical relevance.

Some argue that topical relevance underlies all other types of relevance (Soergel, 1994), while others used examples to show there can relevance without there being any topical relation (Harter, 1992, Hersh, 1994). There is a lot more to relevance than presented here, but the above mentioned interpretations should suffice to show the difference between the interpretation of relevance for which link information has been found effective—namely, the quality, importance and credibility as aspects of *utility* that underlie the notion of relevance used to model Web search—and the notion of topical relevance underlying the ad hoc search methodology, for which the value of link information is still an open issue.

2.1.2 *Evaluation*

Establishing whether a phenomenon, such as the existence and structure of hyperlinks, is useful for information retrieval is often done through evaluation using test collections. This methodology uses a collection of documents, a number of information needs or requests, and relevance judgements indicating which documents in the collection are relevant to which information request. If we want to know whether or not link information is useful for IR, we create a baseline retrieval system S_1 that uses no link information and an alternative version S_2 of the same system that uses link information. The set of information requests, in the form of queries, is processed by the two different systems, and the returned results are compared with the relevance judgements, after which scores for both systems are produced indicating how well they performed at finding the right documents. This allows us to compare the performance of the two systems.

If we are interested in knowing whether links can help finding more relevant documents, we can measure the recall (the fraction of all relevant documents that are retrieved) of S_1 and S_2. If we are interested in the impact of link information on precision (the fraction of retrieved

documents that are relevant), we can measure the number of relevant documents in, for instance, the first 10 results.

There are many aspects of performance we can measure, but it is important to understand what we should be measuring. What is most important for the user? Does the user want as many relevant documents as possible, or to quickly find at least one relevant document that contains the required information? This depends on the particular context in which the user is using the retrieval system.

2.1.2.1 *Search Tasks*

People use IR systems for many different purposes. A person involved in a hefty argument about the cultural value of modern art might be looking for newspaper articles or text books supporting his or her perspective, while another person trying to book a flight to Bangkok might be looking for a Web site that shows which airlines offer cheap flights to the desired destination. The first person is probably more interested in a number of texts that discuss the interpretation of modern art at length, while the second person probably wants a single site that gives ticket prices for a large number of airlines. The first person is searching for information about a certain topic, modern art, while the second is looking for a good starting point to compare airline ticket prices. These are different search tasks with different goals and different criteria of what makes a search result useful. The search task of the first person is close to what has long been the dominant user model in IR research: searching for text on a certain topic. The search task of the second person was later adopted as part of a more appropriate model for how and why people search on the Web.

In fact, even the first person might still prefer a retrieved result that is the entry page of a whole Web site on modern art, instead of a page deep within that site with a lot of detailed information. The entry page might have no directly relevant text, but might give the user a better idea of what information about the topic is available on the same site and how the various parts are related to each other. Locating the entry page to such a topically relevant Web site is the task of home page or entry page finding. Locating the pages with detailed information on the topic is the task of ad hoc retrieval.

There are different stages in the search process, with differing degrees of clarity and structure (Vakkari, 1999). Initially, the problem is vague and the relevance criteria are loosely and partially defined. Vakkari (1999) argues that the complexity of a task is related to how well the problem is structured and understood. If the user has a clear

notion of the information requirements, process and output of the search problem, the task is perceived as simple and performance can be predicted more easily. In some cases, Web-centric tasks like entry page finding are simple tasks because the user has a clear understanding of the problem—she knows exactly what she is looking for—such as the case of finding a good site to compare air fares. In other cases, the search process is in an early stage, where the information need is still vague, and the user wants to navigate to a topically relevant Web site to explore the topic further to get a better idea of what she is looking for. It is not entirely clear whether the relevance criteria are the same in these different cases, but for all cases, the task is to locate a relevant entry page.

In the first TREC experiments using Web data—the Very Large Collection (VLC) Track—the organisers compared the performance of the systems of six TREC participants against five live Web search engines Hawking et al. (1999a). They adopted the ad hoc evaluation methodology to assess the relevance of the returned results. One of the surprising findings of the track was that the standard text retrieval systems used in the research community clearly outperformed the Web search engines in terms of both precision and recall.

One explanation given was that actual Web search engines use less effective retrieval algorithms for efficiency reasons. They need to process enormous amounts of data and have to respond to hundreds or thousands of queries at the same time. A complex algorithm that gives optimal results but takes half a minute per query to respond is not acceptable. Another explanation was offered by Craswell et al. (1999), who pointed out several problems with the assessments of web pages for the VLC. They offered a number of hypotheses why the experiments did not properly model Web search.

First, the abundance of hyperlinks allow a system to return a Web page that has no relevant text itself, but many links to relevant pages. In the ad hoc methodology, where pages are judged on their content alone, independent of any other pages, such a page would be judged irrelevant, even though it might be of value to a user.

Second, the TREC topics represent information needs of someone writing an article or report, while Web users might look for particular sites or pages, addresses and phone numbers of people and companies, and answers to all kinds of questions. These tasks require different responses. A user typing the query "Mercedes-Benz" in a Web search engine is probably not looking for a list of pages with as much text about Mercedes-Benz as possible, but might want to be pointed to the

home page of Mercedes-Benz as a good starting point, regardless of whether the information on the home page would be useful in writing a report.

A third important difference is the quality of pages. The ad hoc methodology does not take document quality into account, even though this might be important to the user. A page with a large amount of topically relevant information can still be of little value to the user if she does not understand or trust the information.

Broder (2002) describes a taxonomy of Web search queries. There are three main types of queries:

1. *Navigational*: The user is looking for a specific Web page or Web site.
2. *Informational*: The user is looking for information on a topic, and wants to read one or more Web pages on that topic.
3. *Transactional*: The user wants to make some kind of transaction, e.g., buying tickets, downloading a file, communicating with other people, playing games.

One of the main points he makes is that navigational and transactional queries—which constitute more than 50% of the queries sent to Web search engines—are best supported using not only on-page text, but also link analysis, anchor text, click-through data and semantic analysis, among others.

There have been other, more elaborate taxonomies to capture user intent of Web search queries, such as the one by Rose and Levinson (2004), and the rapid growth of social media and Web 2.0 applications have further broadened the range and nature of Web search. What Broder's analysis showed is that search behaviour on the Web is different from the user model assumed for traditional ad hoc topic search tasks on which the Cranfield (Cleverdon, 1997) and early TREC evaluations were based. The Cranfield experiments were designed to evaluate indexing languages for literature search. The early TREC test collections focused on ad hoc search: given a static collection of documents, find all documents containing information relevant to the topic of request. Each document is judged in isolation, and is only considered relevant if it contains some text relating to the user's information need.

This marks an important difference betweens the aims of different search tasks. The ad hoc task closely models Saracevic' subject knowledge perspective on relevance, that is, topical relevance, while Web-centric tasks are more modelled to capture the utility or pragmatic view of relevance, where quality, novelty, credibility and importance play a role.

2.1.2.2 *Test collections*

Test collections in information retrieval typically consist of a static set of documents, a large number of statements of information needs called topics and a set of relevance judgements. This design is based on the Cranfield paradigm (Cleverdon, 1997), more specifically, the Cranfield II experiments (Robertson, 2008, Voorhees, 2002).

The evaluation based on these test collections has three major underlying assumptions:

- Relevance can be approximated by topical similarity. The similarity of a document and a query can be represented in a binary judgement: a document d is similar (enough) to a query q such that d is relevant for the user stating q, or it is not, such that d is not relevant for the user. The similarity can also be represented by a graded judgement. For instance, a document d can be non-relevant, slightly relevant, mostly relevant or highly relevant for a user stating query q. One implication of adopting binary relevance judgements is that all relevant documents are automatically considered equally relevant. Other implications are that the relevance of a document is independent of the relevance of any other document, regardless of how many possibly relevant documents the user has already seen, and that the information need is static.
- A single set of relevance judgements is representative of the user population.
- The list of relevant documents for a topic is complete, that is, all relevant documents have been judged and judged relevant.

In general, these assumptions are not true. Assessor agreement studies of the TREC Ad hoc relevance judgements have shown that between a pair of assessors, the average agreement lies between 0.42 and 0.49—an agreement of 1.0 meaning assessors agree completely and an agreement of 0.0 meaning they completely disagree (Voorhees, 2002). Among three different assessors, the average agreement is 0.30. For big document collections, judging every single document to find out whether it is relevant for a given topic or not is prohibitively expensive, especially since we want to average evaluation results over a large number of topics. To work around this problem, a technique called pooling was introduced (Sparck-Jones and van Rijsbergen, 1975). A subset of the collection is created by combining the top results from a large number of different contributing retrieval systems. The intention is that the resulting *pool* of documents to be judged is relatively small compared

to the size of the collection, but contains most of the relevant documents. . In typical TREC test collections, such pools contain between 1000 and 2000 documents from a collection several orders of magnitude bigger.

These simplifying assumptions make performance scores hard to interpret in an absolute sense, but allows a comparative evaluation of systems. A system A that significantly outperforms a system B on a large test collection can be considered a better system for the particular retrieval task on which the relevance judgements are based.

2.1.2.3 *Effectiveness measures*

Based on Saracevic's interpretation IR, an IR system is successful if it successfully communicates to the user a list of relevant information and no irrelevant information (Saracevic, 1975). The measure of success is often expressed in terms of *precision* and *recall*. Precision is the fraction of documents retrieved that are relevant. Recall is the fraction of relevant documents that are retrieved.

In this thesis, we will use several retrieval effectiveness measure to evaluate retrieval methods. A definition and short description is given for each.

Precision at rank n (P@n) In IR, as well as in many classification tasks, precision is defined as the fraction of results that are classified correctly. Precision is a set-based measure. In IR, where results are typically in the form of a ranked list, precision is measured over a set of documents up to a certain rank. In the case of a results list of n retrieved documents, the precision over all documents up to that rank n is the fraction of documents that are judged relevant. More formally, it is computed as:

$$\mathrm{P@}n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Rel}(i)$$

where $\mathrm{Rel}(i) = 1$ if document i is relevant and zero otherwise. As an example, if five of the ten highest ranked documents are relevant, $\mathrm{P@10} = \frac{5}{10} = 0.5$. Averaged over Q different queries, we get:

$$\mathrm{P}_Q\mathrm{@}n = \frac{1}{Q} \sum_{i=1}^{Q} \mathrm{P}_i\mathrm{@}n$$

Mean Average Precision (MAP) Precision and recall are set-based measures. But when a ranked results list is returned, the order in which the results are presented should be considered. Intuitively, we want the relevant documents to be ranked higher than any non-relevant

documents. The average precision AveP expresses the average of the precision at each of the ranks of the relevant documents. If we consider multiple topics, each with a ranked list and an average precision AveP, the overall average precision is the mean of these AveP scores:

$$\text{AveP} = \frac{1}{\sum_{i=1}^{N} \text{Rel}(i)} \cdot \sum_{i=1}^{n} \text{P@}i \ \text{Rel}(i)$$

$$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AveP}_q$$

where N is the total number of documents in the collection, n is the number of results returned for query q and Q is the total number of queries.

*Mean Reciprocal Rank (*MRR*)* The reciprocal rank expresses how far a user has to go down the results list to find the first relevant document. It is the inverse of the rank of the first correct answer. For each query, the reciprocal of the rank at which the first relevant documents is returned, and MRR is the mean of the reciprocal ranks over all topics.

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{r_q}$$

where r_q is the highest ranked relevant document in the results list returned for query q.

2.2 LINK INFORMATION

Standard retrieval models used the textual content of documents to match documents against queries. But there are more document features that provide information, such as document length, the logical and physical document structure, metadata and links.

In this thesis we focus on the value of hyperlinks. Studying the structure of links between documents is a form of network analysis with the aim of identifying associations and relationships between documents. Before we turn to information retrieval on the Web and research on the value of link information for Web retrieval, we look at earlier work on using links. Before the Web was created, IR researchers were already looking at ways to exploit inter-document structure in the form of citations in scientific literature and this new type of document called hypertext.

2.2.1 *Bibliometrics*

The idea of using citation information to find documents related to each other was investigated well before the Web. Kessler (1963b) introduced a method for grouping scientific literature based on bibliographic coupling units. "We define a unit of coupling: Two papers that share one reference contain one unit of coupling" (Kessler, 1963b). In (Kessler, 1963a) he showed results of this method on a large number of papers and found that the resulting groups had a high degree of logical correlation.

2.2.2 *Hypertext*

One of the great advantages of digital documents is the possibility to create *hyperlinks*, which allow readers to jump directly from one document to another, related document, without having to search for a physical copy of the referenced document. Ideas about a large information networks of interlinked documents date back as far as the 1930s, when Paul Otlet envisioned a new form of globally accessible encyclopedia based on linked documents (Rayward, 1994).

Early on, people realised that the structure of hyperlinks in hypertext conveys information about the hypertext. To address the problem getting disoriented by jumping between bits of hypertext, the so-called "lost in hyperspace" problem, Botafogo and Shneiderman (1991) and Botafogo et al. (1992) analysed the structure of hypertexts and came up with measures such as the centrality of a node and the compactness of the hypertext. Hypertext authors can use these measures to improve the structure of a hypertext and make it more comprehensible for readers.

2.2.3 *Hypertext Retrieval*

Before the advent of the Web, there were many ideas about using hyperlinks for retrieval of hypertext media. Most of these approaches considered the topical relatedness of linked documents. In other words, they hoped to use links to determine the topical relevance of documents.

Cohen and Kjeldsen (1987) use constraint spreading activation (Anderson and Pirolli, 1984) on semantically linked network of research topics, funding agencies and proposals to find relevant agencies for a particular research proposal. They compare their linked network with associations in human memory. The main idea is to find *semantically*

related topics and determine the *likelihood of support* of an agency, which depends on the relationship of the topics.

Croft and Turtle (1993) used links to extend document representations with terms from the citing document. They compared citation links and nearest neighbour links and found that citation links results in greater improvement in retrieval performance on the CACM test-collection. Nearest neighbour links were generated using cosine similarity and tend to form clusters of documents similar to cluster-based search, which was shown earlier not to be effective (Willett, 1988).

Savoy (1994) argues that specialised mechanisms effective on small text collections are not necessarily effective on large, *unrestricted* text collections. He generated relevance links between documents relevant to the same query, based on relevance feedback. The main idea is to use learning through feedback. The value of a relevance link is based on how often the two documents are relevant to the same query.

Frei and Stieger (1995) used 4,341 hyperlinks between 962 Berkeley-UNIX manual pages, 15 queries and compared top 10 results. They distinguished between referential links and semantic links, and indexed links with terms from the linked documents (basically anchor text indexing). Semantic links have attributes like link type, creation time and author name. For their experiments they compare spreading activation with standard text retrieval. They consider referential links to be navigational in nature and therefore useless for retrieval.

Ellis et al. (1996) look at different types of relevance considerations: judge-relevant, navigator-relevant and searcher-relevant. They also discuss the difficulty of evaluating effectiveness of either browsing or querying if the user can do both. This is important with respect to link-based ranking: browsability and connectedness are important features for user behaviour and satisfaction (Bates, 2002).

Picard and Savoy (2003) explain the assumptions behind relevance propagation as follows: "if a document is cited by a relevant document, then it is possibly relevant itself." They propose a Probabilistic Argumentation System that uses propositional logic to propagate link evidence in a sound way, based on earlier work by Picard (1998).

From this overview, it seems it was generally accepted that hyperlinked text had something new to offer for IR experimentation. Links were seen as valuable evidence for identifying relevant documents, but they also introduced interesting problems for the IR community. The presence of hyperlinks puts a strain on the assumption adopted for the Cranfield experiments that the relevance of a document is independent

of other documents in the collection. Within a hypertext collection, the user is expected to follow the links to create their own trail and gather bits of information of their interest. IR researchers wondered how this aspect of hypertext retrieval should be evaluated (Agosti, 1993). Even for a topic search task, hyperlinks make a difference to the user experience. Savoy (1992) argues that for hypertext retrieval, it is important to find good starting points and thus precision is more important than recall. This is similar to the argument used later in Web retrieval evaluation that early precision is more important than recall.

2.2.4 *The World Wide Web*

Based on experience with early hypertext systems, Berners-Lee (1990) proposed a system to keep track of large amounts of information at CERN, that later led to the development of the World Wide Web. Instead of new people having to ask around about where to go for a particular piece of information or who to talk to for a certain task, he envisioned a large, linked information system where all the recorded information about the organisation and past projects is stored and can be search non-linearly. In his proposal, links between notes could be labeled to indicate the type of relation between the information objects.

However, the use of typed or labeled links never really took off in the WWW. Pirolli et al. (1996) describe a transition from closed hypertext systems to the World Wide Web: "In its current implementation, the World-Wide Web lacks much of the explicit structure and strong typing found in many closed hypertext systems. While this property probably relates to the explosive acceptance of the Web, it further complicates the already difficult problem of identifying usable structures and aggregates in large hypertext collections."

THE STRUCTURE OF THE WEB With the rapid growth of the Web, the global hyperlink structure was a popular object of study. Kleinberg et al. (1999) and Broder et al. (2000) studied the link structure in the Web as a graph. Both found that the in- and out-degrees of web pages follow a power law distribution. "A power law implies that small occurrences are extremely common, whereas large instances are extremely rare. (Adamic, 2007)" In terms of incoming link degrees, it means that many pages have few incoming links while very few pages have very many incoming links. Formally, the probability of having x incoming links is:

$$P(X = x) = x^{-a} = \frac{1}{x^a}$$

where $-a$ is the slope of the distribution.

Broder et al. (2000) also looked at the connectedness of the Web link graph. They found that there is a single large component of pages that can be reached from each other merely by following the link structure. In a set of 200 million web pages, this *Strongly Connected Component* (SCC) consisted of 56 million pages (28%). There are two other large sets of pages, the set IN of pages that can reach the SCC by following links, but that cannot themselves by reached from the SCC, and the set OUT of pages that can be reached from the SCC but from which the SCC cannot be reached. The IN and OUT sets are roughly of equal size, each containing around 44 million pages.

These power law distributions of the link degrees had been observed earlier in analysis of citations in scientific literature (Fairthorne, 1969). Several studies have tried to explain this phenomenon. Price (1976) came up with the notion of cumulative advantage as a mechanism to explain the occurrence of the power law distribution. Kleinberg et al. (1999) observed this phenomenon with the link structure in the Web, and suggested a copying process, in which a web page copies some of the links of a randomly picked other page. Barabási and Albert (1999) introduced the notion of *preferential attachment*, where newly added links tend to point to popular pages. Since popular pages are more well-known than unpopular pages, they attract more hyperlinks.

The link structure of the web also invites social network analysis (Wasserman and Faust, 1994), in particular notions of authority or importance (Katz, 1953, Seeley, 1949). Particularly intriguing is the question whether such a link-based notion of importance can help improve search results. This question has been addressed by using either the global link structure, PageRank (Page et al., 1998), or the local link structure, HITS (Kleinberg, 1999).

Links have been used to identify so-called 'cyber-communities' in the Web. These communities are "groups of content-creators sharing a common interest (Kumar et al., 1999)." They identify groups of web pages linking to each other by scanning the link structure for strongly connected bipartite graphs using co-citation information. Gibson et al. (1998) use the HITS algorithm on a set of results returned by an internet search engine to identify communities. The idea of using link structure to identify communities leans on the assumption that pages close to each other in the link topology are also topically related to each other.

Support for this assumption came from Davison (2000), who investigated whether web pages actually tend to link to other web pages with related content. His main finding is that the likelihood of linked pages

having similar content is high. He measured the textual similarity of linked and co-cited pages and observed that the similarity between pages linked to from the same source increases when the links are closer together on the source page.

(Chakrabarti et al., 2002) investigated the degree distribution of sets of pages focusing a single broad topic. They found that the link degree distribution of a set of pages on the same topic resembles that of the larger Web. They also showed that the topic distribution converges when starting from different topics. Random forward walks lose focus more slowly than undirected and backward walks. Within communities of different topics lose focus at different rates.

2.3 WEB RETRIEVAL

Where IR research was born out of an attempt to deal with the information explosion after the second world war, Web retrieval research, in turn, focused on dealing with another information explosion when the World Wide Web became popular.

With the advent of the Web and Web retrieval, the ideas about the value of hyperlinks gradually changed. Perhaps through the enormous popularity and explosive growth of the Web in its early years, the understanding of hyperlinks decreased as the Web became ever more heterogeneous.

Perhaps the quick growth was caused by the ease of use of HTML and the ease of creating hyperlinks without specifying their type. Whatever the reason, the hyperlinks of the World Wide Web are abundant, but created for many different reasons and without any semantic label.

The people who initially participated in the TREC Web tracks based their techniques on early research of retrieval in hypertext, which was conducted on collections very different from the World Wide Web. Back then, the semantics of links was considered useful for retrieval. Shakery and Zhai (2006) argued: "Given a query, intuitively, a good result document is one whose content is related to the query topic and which is surrounded by other good documents; i.e. located in the center of a subset of the collection relevant to the query. Thus in order to maximize ranking accuracy, we need to consider the relevance of the document to the query as well as the relevance of its neighbors." And according to Bharat and Henzinger (1998): "The goal of connectivity analysis is to exploit linkage information between documents, based on the assumption that a link between two documents implies that the documents contain related content (*Assumption i*), and that if the

documents were authored by different people then the first author found the second document valuable (*Assumption ii*)." When Web usage and search finally took off, the retrieval environment was radically different from the document collections used earlier. The Web was not a collection of high quality articles written by experts with carefully placed citations and hyperlinks, but an almost uncontrolled mess of Web sites and Web pages where countless individuals could share any kind of information they wanted, in any form.

2.3.1 *Large scale evaluation of link information: TREC Web Tracks*

Based on claims from commercial search engine companies, over the course of several years of Web search experiments at TREC (TREC, 2009), organisers and participants have tried to establish the effectiveness of link information, including anchor text, for retrieval. Despite the enthusiasm and effort of many participating groups, in the first two years, 1999–2000, participants failed to show any improvements due to link information (Hawking and Craswell, 2005).

At TREC-8, in 1999, participants could not show consistent improvements over content-only baselines using link information (Hawking et al., 1999b). This unexpected result led participants to believe that the collection had too few inter-server links for link evidence to be effective. In response, a new collection, named WT10g, was constructed focusing on inter-server link density (Bailey et al., 2003). In the TREC-9 Web Track, many different link-based methods were used, including attempts at exploiting anchor text for ad hoc retrieval, but again no one could show any improvements using link information (Hawking, 2000). Singhal and Kaszkiel (2000) raised doubts about the TREC evaluation methodology used to model Web search, as they found different results for anchor text when comparing TREC results against their in-house tests.

2.3.1.1 *Web-centric search tasks*

Several studies (Craswell et al., 1999, Singhal and Kaszkiel, 2001) pointed at the differences between traditional ad hoc search as evaluated at TREC and Web search behaviour. Web searchers tend to "prefer the entry page of a well-known topical site to an isolated piece of text, no matter how relevant" (Hawking and Craswell, 2005). Web users often have short-term information needs such as finding a particular Web site (Broder, 2002, Jansen and Spink, 2006) and rarely look beyond the first page of search results (Jansen et al., 1998, Silverstein et al., 1999).

For them, the quality of the first results page is far more important than what comes after the first page.

As new, more realistic Web tasks were introduced, the value of link information was finally shown Hawking and Craswell (2001). Craswell et al. (2001) and Kraaij et al. (2002) found anchor text to be very effective for site-finding, and home page finding tasks. Ogilvie and Callan (2003) and Kamps (2005) showed that document prior probabilities based on URL depth and link in-degree significantly improve performance on known-item search tasks. Craswell et al. (2005) study query independent evidence for a mixed query set of topic distillation, home page finding and named-page finding topics, and find that, in order of impact, PageRank, in-degree (both explained in Section 2.3.6), URL length and click-distance improve the effectiveness over the mixed query set. The click-distance is a metric for the distance in clicks needed to arrive a page from a certain root page. Nie et al. (2006) selected 12 top level categories from the Open Directory Project (ODP, 2010) to compute a content vector for all document/query pairs of the TREC .GOV collection and the TREC 2003 Web Track Topic Distillation topics. They adjust the PageRank and HITS algorithms to differentiate between following a link to a page on the same topic or following a link to a page on a different topic and found their technique outperforms text-based ranking functions. This seemed to close the gap between the general belief that links are useful for search and the contradictory findings at TREC. According to Hawking and Craswell (2005, p.215):

> Hyperlink and other web evidence is highly valuable for some types of search task, but not for others.

Although the switch to more Web-centric search tasks like home page and named page finding showed link information to be very effective for these tasks, no clear explanation was given why link evidence is not effective for ad hoc retrieval.

Gurrin and Smeaton (2004) pointed out that the inter-server link density of the WT10G collection was still very low, and extracted a subset of the collection, WT-*dense*, which has a much higher inter-server link density. Within this tiny subset they found that a combination of content and link information could improve precision on the ad hoc topics of the TREC-9 Web track. This led them to come up with a list of requirements that a representative test collection must satisfy to study the value of link information. A good Web collection needs to be sufficiently large and have sufficiently high inter- and intra-server link densities.

The size issue was addressed in the Terabyte Tracks of 2004–2006, which used the GOV2 collection, based on a crawl of the .gov domain in 2004, consisting of 25 million documents.[1] Again, anchor text was found to be highly effective for Web-centric tasks, but not for ad hoc search (Kamps, 2006b, Kamps et al., 2005). However, the .gov domain is very different in nature from the .com domain on which the WT10G collection is based, and the .GOV2 collection has fewer incoming links per page. Thus, although it is larger than the earlier Web Track collections, its link density is much lower, making it hard to investigate the impact of collection size.

At the TREC 2009 Web Track (Clarke et al., 2009) a new, large Web collection—ClueWeb09 (CMU-LTI, 2009)—was introduced and the traditional Ad hoc Task was paired with the new Diversity task. This new collection is much larger than the collections used at TREC 8 and 9, and was crawled to reflect the first tier of a commercial search engine index consisting of the most important pages, so should have a relatively dense link structure, allowing us to study both aspects of collection size and link density. If a large number of documents and a high link density are indeed requirements for link evidence to be effective, this new collection might finally reveal its potential.

Surely, the issue of having enough (inter-server) links is critical for any search task, but perhaps the link density needs to be higher to be effective for ad hoc retrieval than for entry page finding. Links within the same site are often navigational links, with anchor terms such as 'click', 'here' and 'next' (Eiron and McCurley, 2003). Therefore, it is generally assumed that links between sites are more meaningful, including their anchor text (Metzler et al., 2009).

2.3.2 *Types of Web Search*

There are many forms of Web search, a number of which have been explored by the TREC VLC/Web tracks.

- *Online service finding*: the user is looking for pages providing some online service, where the user can make a transaction, such as downloading an MP3, buying tickets or booking a hotel.
- *Home page finding*: the user is looking for the entry page of a Web site relating to some entity, be it a person, company or product.

1 Unfortunately, the crawl on the .gov domain was exhausted long before reaching the targeted 100 million pages, and plans to rectify this by crawling additional pages from the .edu domain were never realised.

- *Named-page finding*: the user is looking for a "single important document" that is not a site-entry page.
- *Topic distillation*: the user wants a list of key resources on a certain topic. These key resources often are pages in a topically relevant site, at the right level of the hierarchy. Although in some sense related to home page and named-page finding, in that the relevant pages are often entry pages, it is also related to traditional ad hoc search in the sense that the user is not looking for a single, known web page, but a list of pages that provide entry at the right level into a site with topically relevant information.

Home page and named-page finding tasks, which cover the navigational queries in the Web search taxonomy (see page 20), are easy to judge because the user is looking for one particular page. Although some pages have multiple URLs, this should not be a big problem. If one is retrieved, typically all variants are retrieved unless some kind of duplicate detection is used. As long as all variants are identified, evaluation can be done properly. Because the user is looking for one particular result, suitable evaluation measures are Mean Reciprocal Rank and Success at n documents retrieved (Success at n means there is at least one relevant document among the first n results).

Topic distillation tasks, which cover the informational queries in the Web search taxonomy, require a results list with many topically relevant entry pages in the top ranked results, so the early precision must be high. Mean Average Precision and R-precision (the precision at rank R where there are R relevant documents in total) are suitable measures.

2.3.3 *TREC Web collections*

Throughout the history of the TREC Web tracks, there has been discussion about what constitutes a good Web test collection. Not only the tasks and topics are important, but also the document collection. For comparison, some information of the test collections created for the TREC evaluations are given in Table 1. No link counts are known for the full VLC2 collection. The WT2G and WT10G are derived from the VLC2 collection, which is based on a crawl in 1997, and were created for the Web Tracks of 1999–2001.

An overview of the tasks, topics and collections used for the TREC Web Tracks is given in Table 2. 1999 was the first year in which the value of hyperlink information was investigated in the Web Track. Ad

Name	*Year*	*Domain*	*# Pages*	*# Links*
VLC2 (WT100G)	1997	.com	18,571,671	–
WT2G	1997	.com	247,491	1,166,702
WT10G	1997	.com	1,692,096	8,062,918
.GOV	2002	.gov	1,247,753	11,110,985
.GOV2	2004	.gov	25,205,179	82,711,345
ClueWeb09 (B)	2009	.com	50,220,423	1,180,631,904

Table 1: Information on the year, domain, size and number of links of the TREC Web Track collections.

hoc search task evaluations were conducted in 1999–2001, 2004–2006 and 2009.

2.3.4 *Crawling and page quality*

An important aspect of Web retrieval evaluation is the sample of the Web that is used to represent it. Several Web test collection have been made for the TREC Web Tracks, all based on crawls of parts of the Web.

A Web *crawler* is a program that traverses the Web by following hyperlinks to discover new pages and page content. Starting from a list of *seed* URLS, the crawler downloads the pages found at those URLS, stores the content and extracts all the hyperlinks of these pages pointing to new URLS. These new URLS are downloaded next and the process of extracting content and hyperlinks is repeated until all extracted URLS have been downloaded and no new URLS are found. This is a way for internet search engines to discover content on the Web and index the text of the web pages they find so that users can search for them. The process is called *crawling* and the resulting collection of downloaded pages is a *crawl*.

One important difference between the new ClueWeb collection and previous TREC Web collections is the quality of the pages in the crawl, which is related to the way it is constructed, and which directly affects the density of (inter-server) links. Several studies have looked at the impact of crawling policy on the quality (Baeza-Yates et al., 2005) and search effectiveness (Fetterly et al., 2009a,b) of the crawled collection. Page importance metrics can be used to schedule the most important or useful pages to be crawled first. Since page importance is usually derived using link-based measures such as PageRank (Page et al., 1998) or On-line Page Importance Computation (Abiteboul et al., 2003), which

Year	*Task*	*Name*	*# topics*
1999	Large Web (ad hoc)	VLC2	50
	Small Web (ad hoc)	WT2G	50
2000	Large Web (online services)	VLC2	50
	Main Web (ad hoc)	WT10G	50
2001	Web Topic Relevance (ad hoc)	WT10G	50
	Home page Finding	WT10G	145
2002	Topic Distillation	.GOV	50
	Named Page Finding	.GOV	50
2003	Topic Distillation	.GOV	50
	Named Page Finding	.GOV	150
	Home Page Finding	.GOV	150
2004	Topic Distillation	.GOV	75
	Named Page Finding	.GOV	75
	Home Page Finding	.GOV	75
	Ad hoc	.GOV2	50
2005	Ad hoc	.GOV2	50
	Named Page Finding	.GOV2	252
2006	Ad hoc	.GOV2	50
	Named Page Finding	.GOV2	181
2009	Ad hoc	ClueWeb09	50
	Diversity	ClueWeb09	50

Table 2: Types of search tasks and test collection information for the TREC Web collections

give a higher score to a page if it has more incoming links, the first part of a crawl based on such policies tends to have a high link density. One of the primary goals of creating the ClueWeb data set was "to approximate Tier 1 of a web search engine index" Callan et al. (2008). The category B data set, which we use in this thesis, consists of the first 50 million English pages of this crawl.

2.3.5 *Web retrieval outside TREC*

Outside TREC, other researchers have used link information to improve Web retrieval performance.

Carrière and Kazman (1997) used the links as "relationships between some set of nodes of interest", where the nodes of interest are documents retrieved in response to a query. Documents matching the query form the root set and are expanded with any document connected to one of the documents in the root set. The direction of links is ignored and documents are ranked in decreasing order by the number of incoming and outgoing links. Marchiori (1997) used links to propagate document scores through the document network, with a fading or damping function so that a document's score has less impact on documents that are further away from it in the link structure. He found that the precision of then-popular search engines could be improved.

Bharat and Henzinger (1998) adjusted the HITS algorithm by incorporating the relevance score of a page in the formula so that the highly ranked pages also have the biggest influence on the calculation, and removing documents that are not sufficiently similar to the query. Chakrabarti et al. (2002) classify Web pages according to a 482-class topic taxonomy based on the DMoz (DMOZ, 2010) structure and study the link structure within the sets of pages belonging to each topic. Chakrabarti et al. find that forward random walks lose the starting topic memory as quickly as undirected walks. Haveliwala (2003) precomputes topic-specific PageRank scores (PageRank is described in Section 2.3.6.2) using 16 top-level topics from DMoz. Class-probabilities for the 16 topics are computed for a given query, after which the query-sensitive importance score is computed by multiplying the class probability with the topic-specific PageRank score.

There has been an important attempt to bridge the gap between the scale of scientific IR test collections and the Web at large. Najork et al. (2007) study the effectiveness of link-based evidence on 463 million Web pages, 28,043 queries and evaluate on the top 10 results. They find that combining link-based features with the content-based scores lead

to substantial improvements, with features based on incoming links (PageRank, in-degree, HITS authorities) superior to features based on outgoing links (out-degree and HITS hubs).

2.3.6 *Link-based Ranking Algorithms*

Link-based ranking algorithms use the link structure to determine an ordering of the documents or nodes in a link graph. The best-known algorithms are degree-based and/or propagational algorithms.

2.3.6.1 *Degrees*

One of the simplest ways to derive information from the link structure is to count links incident to a page, called the link degree. Links can be counted for the pages to which the links point (the incoming link degree or in-degree), the pages from which the links originate (the outgoing link degree or out-degree) or the combination of the two (the undirected link degree).

2.3.6.2 *Propagation algorithms*

Propagation algorithms propagate some kind of score from one document to another document via links. The best-known propagation algorithms are PageRank, HITS, SALSA and relevance propagation.

PageRank is an algorithm that objectively and mechanically rates Web pages using the link structure as an approximation of the relative importance of individual pages. Intuitively, "a page has a high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks" (Page et al., 1998). The ranking algorithm is based on citation analysis of academic papers, but is tailored to take into account the diverse nature of Web pages in terms of quality, usage, citations and length. "Unlike academic papers which are scrupulously reviewed, Web pages proliferate free of quality control or publishing costs" (Page et al., 1998). The algorithm models a random surfer blindly clicking links. PageRank is computed as:

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in I(p_i)} \frac{\mathrm{PR}(p_j)}{L(p_j)}$$

where $I(P_i)$ is the set of pages linking to page p_i and $L(p_j)$ is the number of outgoing links on page p_j. The damping factor d is the probability

that the random surfer gets bored and jumps to a random page in the collection and is usually set to 0.85. The algorithm is executed iteratively until the PageRank scores converge and a stable distribution is reached. PageRank is usually computed on the entire document collection that is indexed. In other words, it uses link information on a *global* level.

HITS (Hyperlink Induced Topic Search) is an algorithm to compute the authority of a page for particular search topic. Intuitively, authoritative pages are linked to by many good hub pages, and good hubs have many links to good authoritative pages on a certain topic. Kleinberg distinguishes between broad and specific queries. For specific queries, only a few relevant pages exist, and the challenge is to identify them. For broad queries, on the other hand, thousands of relevant pages exist, and the challenge is to identify the most useful, reliable pages. HITS is designed for broad topics, and aims to identify the most authoritative pages among the large set of retrieved pages. It runs at query time and computes two scores for each page p_i in a small, *local* set S of pages: an *authority* score $x^{\langle p_i \rangle}$ and a hub score $y^{\langle p_i \rangle}$. The algorithm is thus query-specific. Like PageRank, it is also iteratively executed until convergence. The idea is that authorities and hubs are mutually reinforcing. Authority and hub scores are computed as:

$$x^{\langle p_i \rangle} \leftarrow \sum_{p_j \in I(p_i)} y^{\langle p_j \rangle}$$
$$y^{\langle p_i \rangle} \leftarrow \sum_{p_j \in O(p_i)} x^{\langle p_j \rangle}$$

where $I(p_i)$ is the set of pages linking to page p_i and $O(p_i)$ is the set of pages linked to by page p_i. Initially, all pages in the set have unit authority and hub scores $x_0^{\langle p_i \rangle} = y_0^{\langle p_i \rangle} = 1$. The local graph $\mathcal{G}$ based on S is formed by taking the top N (usually 200) retrieved results for a query q as a root set R and expanding this set by adding pages that link to or are linked to from pages in R. The expansion step is done to add possibly relevant pages that do not match the query.

The main difference between HITS and PageRank is that HITS is often used on a relatively small set of documents retrieved for a particular query—and is therefore query-dependent—while PageRank is often used to determine the importance of a page within an entire collection of documents and is thus query-independent. HITS works on a local level, while PageRank works on a global level. Xu and Croft (1996, 2000) compared global and local feedback techniques. They hypothesise that the top-ranked documents tend to form several clusters. Their

conjecture is that non-relevant documents in the top-ranked set also tend to cluster, because they are similar to the query. Within the top-ranked set, each cluster might represent a different topic, and the largest cluster might not necessarily cover the requested topic. This could pose a problem for local link evidence based on link counts. Borodin et al. (2005) discusses the problem of *Tightly Knit Communities*, explaining that HITS is known to favour nodes belonging to tightly interconnected components. The effectiveness of HITS is thus very dependent on the relevance of these components (Borodin et al., 2005, p. 277–286). This problem was also observed by Lempel and Moran (2001), and led them to develop the SALSA algorithm.

SALSA (Stochastic Approach for Link-Structure Analysis) is similar to HITS and is equivalent to a weighted in-degree analysis of the link structure. Using the notions of hubs and authorities, SALSA performs a random walk where transitions consist of traversing two links, one link forward and one link backward (or vice versa), effectively avoiding the Tightly Knit Communities problem described above. Hub and authority scores are then computed iteratively until they converge.

Relevance propagation is yet another propagation algorithm, but one that uses the retrieval score of the returned results as initial relevance weights (Shakery and Zhai, 2006) and thereby indirectly exploits the document content of linked documents to augment a document's own content. More relevant documents contribute more to a document's score than less relevant documents. They use incoming and outgoing links (forward and backward propagation) and find that both are effective and the combination of the two even more effective.

Tsikrika et al. (2007) use a K-step random walk to propagate relevance:

$$p_0(d_i) = p(q|d_i)$$

$$p_t(d_i) = p(q|d_i)p_{t-1}(d_i) + \sum_{d_j \in I(d_i)} (1 - p(q|d_j)) \frac{p_{t-1}(d_j)}{|O(d_j)|}$$

where $p(q|d_i)$ is the retrieval score in the form of a probability, $I(d_i)$ is the in-degree of document d_i and $O(d_j)$ is the out-degree of document d_j.

2.3.6.3 *Anchor text*

Anchor text is the text on a Web page in which a hyperlink is anchored and provides a textual description of the targeted page. The anchor text

in Web pages is used to create an extra document representation which can be retrieved in the hope of improving the textual representation (closing/narrowing the semantic gap). This has a direct impact on the textual retrieval model. Anchor text representations explicitly use external descriptions of a page, i.e., what others say about a document. An anchor text representation implicitly filters links on the search query. Although the generation of links is query-independent, the score of a retrieved anchor text document reflects how often query terms hit the document, and thereby how many incoming links are related to the topic. Thus, it uses all links related to the query.

Typically, anchor text is propagated only from the document with the link anchor to the linked document. However, for Web search there is additional value in propagating anchor text in multiple steps to reduce the sparseness of the link graph (Metzler et al., 2009). Within a single site, not all pages receive anchor text from site-external pages. Some pages are only linked to by other pages from the same site. The anchor text representation of such documents can be expanded by propagating the anchor text from site-external pages in multiple steps to pages within the site that have no site-external incoming links.

Eiron and McCurley (2003) found that anchor text behaves very much like real user queries. Web authors use the same labels to describe pages as Web searchers use to find pages. If that is the case then anchor text has the potential to bridge the gap between queries and pages and lead to high precision, if the anchors and pages in the collection are of high quality.

2.4 ANALYSIS OF WIKIPEDIA

Wikipedia has been a popular subject of study and has been as a knowledge base for many different tasks. Voss (2005) analysed the size, growth, content and quality of Wikipedia articles and found that Wikipedia can be modelled as a scale-free network. Capocci et al. (2006) analysed the statistical properties of several different language versions of Wikipedia and found that, like the Web, the growth of Wikipedia can be modelled by local mechanisms like preferential attachment even though users can make changes to the network on a global level. The incoming and outgoing link degrees follow a power law distribution. In Wikipedia, as in most technological networks—including the Web—the link topology exhibits disassortative mixing: pages with few incoming and outgoing links tend to be connected to pages with many incoming and outgoing links (Zlatic et al., 2006).

Bellomi and Bonato (2005) analyse PageRank and HITS on the Wikipedia link graph and provide lists of most authoritative pages, countries and cities, historical events, people and common nouns. The HITS authority ranking reveals space (geographic locations) and time (periods and historical events) to be the main organising categories in Wikipedia. The articles with the highest PageRank scores are dominated by concepts related to religion. They conclude that there seems to be a strong bias towards Western culture and history in the English Wikipedia.

Because all changes and contributions to Wikipedia are time-stamped, it is possible to study temporal aspects of the evolution of large document networks. Buriol et al. (2006) find that there is a strong correlation between PageRank and in-degree, "indicating that the microscopic connectivity of the encyclopedia resembles its mesoscopic properties." They also found that Wikipedia has matured in terms of the link degree distribution and connectedness, in the sense that, over time, the power law degree distributions and fraction of articles connected to the main component are stable.

Milne and Witten (2008) use Wikipedia links to compute the semantic relatedness of concepts. They find that using only link information is more effective than measuring semantic relatedness using the Wikipedia category structure, which was done by Strube and Ponzetto (2006), and almost as effective as the more complex Wikipedia-based Explicit Semantic Analysis algorithm by Gabrilovich and Markovitch (2007).

Ahn et al. (2004) were among the first to use Wikipedia as an external resource to improve retrieval performance. Test collections of Wikipedia have been built through the INitiative for the Evaluation of XML retrieval (INEX), where a snapshot of the English Wikipedia of early 2006 (Denoyer and Gallinari, 2006) was used for the Ad Hoc tracks of 2006–2008 (Fuhr et al., 2008a, Kamps et al., 2009, Malik et al., 2006). Since then, Wikipedia has been used to evaluate entity ranking techniques (de Vries et al., 2007, Pehcevski et al., 2008, Zaragoza et al., 2007) and link-detection (Huang et al., 2008). Kaptein et al. (2009) successfully used the Wikipedia category structure to improve ad hoc retrieval performance.

The link structure has also been used to measure semantic relations between pages. Milne and Witten (2008) derive the semantic relatedness of two Wikipedia articles from the link structure, and compare their technique against manually defined relatedness measures and find it to be very competitive. This link-based relatedness measure is used by Lizorkin et al. (2009) to evaluate the semantic relatedness of Wikipedia articles clustered by a link-based community detection algorithm. They

filter the dense link graph for computational reasons and retain only meaningful links, and find that the clustered articles show high levels of semantic relatedness. In a similar vein, Capocci et al. (2008) investigate the overlap between Wikipedia articles clustered using link information and those grouped by categories and find that link-based clusters show very little overlap with the categorical organisation of Wikipedia. Chernov et al. (2006) use the Wikipedia link structure to infer semantically important relationships between categories. Categories are closely related if there are many links between the documents in these categories. For example, Wikipedia pages about capitol cities often have links to pages about countries and vice versa. As a consequence, these links connect documents that have a semantically important relationship and could be labelled as semantically important links. An extensive overview of using Wikipedia as a knowledge base for many different tasks is presented by Medelyan et al. (2009).

In a sense, Wikipedia is the collaboratively developed universal encyclopedia that Paul Otlet envisioned (Rayward, 1994), although the lack of typed links and the openness to allow any person to edit any piece of information in Wikipedia places it closer to the uncontrolled and heterogeneous Web.

In this chapter we have seen how links have been used for information retrieval and more specifically in Web retrieval. Links in the Web have proven effective for identifying entry pages and other important pages, but so far have failed to show their value for identifying pages on a requested topic. In the next chapter we start our analysis of the value of link evidence for information retrieval, where we focus on Wikipedia ad hoc retrieval.

Part ii

Link Evidence in Wikipedia

3

LINK EVIDENCE IN WIKIPEDIA

From the initial conjecture that links in Wikipedia are more semantic than Web hyperlinks in general, we first set out to investigate whether we can use link evidence in Wikipedia to improve ad hoc retrieval effectiveness. Our main research question is:

- Can the link degree structure of a semantically linked document collection be used as evidence for the relevance of ad hoc retrieval results?

As mentioned in the previous chapter, there are many ways of exploiting link information, such as PageRank, HITS, SALSA and relevance propagation, and anchor text. One characteristic of link structure that is analysed by many link-based ranking methods is the incoming link degree of a document, that is, the number of links pointing to a document. As a consequence, these methods are highly correlated to link degrees, simply because they explicitly use the number of incoming and outgoing links in their score computation. However, they are computationally more complex and harder to interpret than link degrees. Since we are interested in the information conveyed by the link topology, a content-based relevance score propagation method would be inappropriate, as it muddles the impact of the purely structural link information by combining it with the text-based retrieval score. The same holds for link anchor text, which is used for text matching, and thus using more than just the link structure. Therefore, we will start our investigation by looking at the link degree structure of the Wikipedia link graph. Degree information can be obtained by simply counting links, even when used in a query-dependent way (which is discussed in Section 3.1.5), and is derived purely from the link structure. It is also easy to interpret: a document with an in-degree of m has m other documents linking to it.

We will first describe our experimental methodology, then analyse the link degree structure of Wikipedia and its possible relation to topical relevance. Finally, we will describe how we can incorporate link degree evidence into our retrieval model and discuss the impact of link evidence on retrieval effectiveness.

Description	Value
Documents	659,388
Topics	221
Judged documents	124,322
Judged per topic	563
Relevant documents	12,107
Relevant per topic	54.78

Table 3: Statistics of the INEX Wikipedia test collection

3.1 EXPERIMENTAL SET-UP

This section describes the experimental set-up used throughout this thesis.

3.1.1 *Test collection*

We use the INEX 2006 Wikipedia collection (Denoyer and Gallinari, 2006), which is a snapshot of the English Wikipedia of early 2006, and consists of 659,388Wikipedia articles transformed into XML format. This snapshot is taken from a full article dump provided by the Wikimedia foundation (Foundation, 2009). The Wikimedia dump contains only the editable article text, without any of the navigational frames that are automatically provided for all Wikipedia pages. The INEX Wikipedia collection contains only the encyclopedic articles of the English Wikipedia, without the main entry page, discussion, history and category pages. To evaluate retrieval effectiveness, we use a set of 221 ad hoc topics and relevance judgements from the 2006–2007 Ad Hoc Tracks (Fuhr et al., 2007, Malik et al., 2006). For these topics, assessors were asked to highlight all and only relevant text in articles in the judgement pools (Lalmas and Piwowarski, 2006, 2007). For our analysis we will transform these to article level judgements by assuming that an article is relevant for a topic if at least some of its content is highlighted by the assessor of that topic. This is similar to the TREC Ad Hoc methodology, where a document is considered relevant if it contains any text relevant to the search topic.

Some statistics of the test collection are shown in Table 3. On average, 563 out of 659,388 or 0.085%, (median 571, or 0.087%) Wikipedia articles are judged per topic, and almost 55 are judged relevant (median 34). We expect the Wikipedia collection to have little redundancy, because each

topic has a single, dedicated page. In this light, the pools are very deep. The task of highlighting relevant text ensured that assessors carefully read the whole document and made it hard for them to accidentally label a document relevant when it was not. The large number of topics ensures that observed results are fairly stable, even for early precision measures (Buckley and Voorhees, 2004). Pal et al. (2008, 2010) looked at the stability and error rates of system rankings using the INEX 2007 Ad Hoc topics and judgements and 62 runs for the Focused Task, and found that reducing the pool size of each topic by 20%—that is, randomly removing 20% of the judged relevant text—would lead to system rankings that are very similar to those using the full judgements. Even for extremely early precision measures, which are less stable than measures taking a larger part of the ranking into account, the system rank correlation between using 100% and 80% of the judgements is still above 0.9 over 62 runs and 107 topics. In other words, the pools could have been less deep and still led to the same system ranking. This means that the INEX Ad Hoc pools are deep enough to obtain reliable evaluation results. We use 221 topics, including the 107 used by Pal et al., and therefore expect the evaluation results to be very stable and reliable even for early precision measures.

3.1.2 *Index*

Although the INEX 2006 Wikipedia collection was created for the purpose of evaluation focused retrieval techniques, we want to look at the impact of link evidence on effectiveness of retrieving whole documents. To this end, we stripped all XML markup from the documents and removed stopwords before indexing. No stemming was done. For indexing we use our own language model extension (ILPS, 2005) of Lucene (Lucene, 2010) (see Section 3.1.4).

3.1.3 *Links*

We only use the links between the encyclopedic articles; all other links to external Web pages are ignored. The main reasons are that these external Web pages are not part of the collection, and they are not open to the same collaborative editing environment. The link graph consists of 17,014,573 collection-internal links. Some of these are repeated links, that is, there are multiple links going from page A to page B. We treat repeated links as a single directed connection between two articles, which gives us a total of 13,602,613 links. With 659,388 articles, this

means each article has on average $\frac{13{,}602{,}613}{659{,}388} = 20.63$ incoming links. Because each link has a source article and a destination article, the average number of incoming links is the same the average number of outgoing links: the average outgoing link degree is also 20.63.

3.1.4 *Retrieval model*

In this thesis we run retrieval experiments using the language modelling framework Ponte and Croft (1998). In this framework, a separate language model of each document is used to calculate the probability that the language model of a documents generates the query. The assumption is that the probability of generating the query typed by the user reflects how well a document matches a query. Documents are ranked in descending order of their probabilities. We use a language model extension of Lucene (an open source document retrieval toolkit, see ILPS (2005)), i.e., for a collection D, document d and query q:

$$P(d|q) = P(d) \cdot \prod_{t \in q} \left((1 - \lambda) \cdot P(t|D) + \lambda \cdot P(t|d) \right),$$

where $P(t|d) = \frac{\text{freq}(t,d)}{|d|}$, $P(t|D) = \frac{\text{freq}(t,D)}{\sum_{d' \in D} |d|}$ and $P(d)$ is a document prior probability. For ad hoc retrieval, where document length was found to have a linear relationship with relevance (Singhal et al., 1996), we use a document length prior, $P(d) = \frac{|d|^\beta}{\sum_{d' \in D} |d'|^\beta}$. where β controls the impact of the document length. For our experiments we have used $\lambda = 0.15$ and $\beta = 1$ throughout. Our implementation of the model calculates ranking-equivalent logs of the probabilities (Hiemstra, 2001). We take the exponent to get a score resembling a probability.

3.1.5 *Global and local link evidence*

Throughout this thesis we will discuss link evidence on two levels:

Global: at this level, we consider the entire link graph of the collection for evidence: for instance, the global in-degree of a document d is the total number of links in the collection pointing to d.

Local: at this level, we consider the link graph of a subset of the collection for evidence: for instance, the local in-degree of a document d is the total number of links from document in the subset pointing to d. This subset contains the highest ranked documents retrieved for a given query.

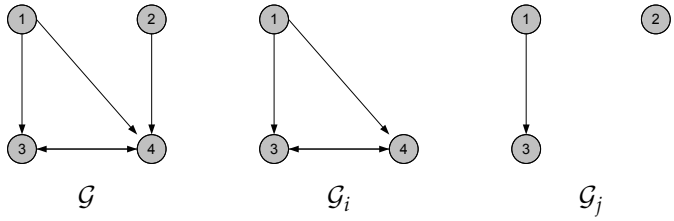

Figure 3: Example global and local link graphs.

The most important difference between global and local link evidence is that global evidence is query-*independent* while local evidence is query-*dependent*. For a document retrieved for two queries, the global link evidence will be the same, but the local link evidence is (probably) different for the two queries. As an example, consider the *global link graph* $\mathcal{G}(\mathcal{N}, \mathcal{L})$ with nodes or documents $\mathcal{N} = (1, 2, 3, 4)$ and edges or links $\mathcal{L} = ((l_{1,3}), (l_{1,4}), (l_{2,4}), (l_{3,4}), (l_{4,3}))$ where $l_{i,j}$ means there is a link from document d_i to d_j (see left side of Figure 3). The full collection consists of documents 1 to 4. For a query q_i, documents 1, 3 and 4 are retrieved. The *local graph* $\mathcal{G}_i$ for query q_i consists of the documents 1, 3 and 4, and the links between them (middle of Figure 3). For query q_j, documents 1, 2 and 3 are retrieved, leading to the *local graph* $\mathcal{G}_j$ (right of Figure 3).

The two queries lead to very different local graphs. In global graph $\mathcal{G}$, documents 1 and 2 have an in-degree of zero, document 3 has an in-degree of 2 and document 4 has an in-degree of 3. In local graph $\mathcal{G}_i$, document 1 has an in-degree of zero, while documents 3 and 4 both have an in-degree of 2. In local graph $\mathcal{G}_j$, documents 1 and 2 have an in-degree of zero and document 3 has an in-degree of 1. Obviously, documents with zero in-degree in the global graph will also have zero in-degree in any local graph. The local graph is a subset of the global graph. This means the local degree can never exceed the global degree. However, for documents with a global in-degree of at least 1, the local in-degree for a given query is determined by the subset of documents retrieved for that query.

The degrees are based on a set of documents; the local link degrees are a product of the set of documents that form the local set, and is not affected by how those documents are ordered. The local graph is the same whether we ranked the documents for q_i as 1, 3, 4 or as 3, 4, 1. Only when the composition of the local set changes—by adding, removing or substituting documents in the set—can the degree structure change.

There are three factors that determine the composition of the local set. We already mentioned the query. Different queries lead to different document rankings, thus to different local sets. The second factor is the retrieval model that produces the ranking. The document ranking produced by a standard language model can be very different from the ranking of, say, a BM25 model (an alternative retrieval model, described in Robertson et al. (1994)). Any change made to a model—by changing a parameter, switching from unstemmed to stemmed indexes, expanding the query—can and often will result in a different ranking and, as a consequence, to a possibly different local graph with different degrees. Although the choice of model will affect the degrees, we want to study the nature of links irrespective of the chosen model, and will limit our experiments to using the model described above.

The third factor that affects the composition of the local set and the associated local link graph is the number of documents we consider for the local set. Since we are concerned with using link information to improve relevance ranking, we might want to focus our local link evidence on relevant pages. By taking into account only the links between documents retrieved for a given query, we effectively filter the link graph to retain only links related to the topic of the query.

Kleinberg (1999), argues that the local set on which HITS (described in Section 2.3.6.2) is used should have the following three properties: it should i) be relatively small, ii) be rich in relevant pages and iii) contain most (or many) of the strongest authorities. Larger sets have more links but less focus on the search topic, and tend to require more computation. The HITS algorithm usually starts with a base set of the top 200 documents, which is expanded by adding documents that are connected to those top 200 documents, resulting in a local set of 1000-5000 documents.

Kleinberg also makes a distinction between broad-topic queries and specific queries. For specific queries, there are only a few pages that contain the required information, whereas for broad topics the Web contains many thousands of relevant pages. The HITS algorithm was designed to identify the most authoritative pages among those thousands of relevant documents. In typical ad hoc retrieval tasks, the topics tend to be more specific, and the number of relevant documents is much lower: tens or hundreds, instead of thousands. The challenge is to identify all relevant documents, or at least as many as possible. We already saw that the INEX Ad Hoc topics have 55 relevant documents on average (Table 3). Clearly, a set of 1000-5000 Wikipedia articles would not be rich in relevant articles.

We choose as the local set the top 100 retrieved documents. Although this set will have some non-relevant documents, it will be relatively rich in relevant documents and some of the non-relevant ones might still be tangentially related to the search topic and have links to the relevant documents. If we make the set much smaller, say, top 10 or 20, there will be hardly any links in the set and the impact will be small. If we use more results, we get more links, thus more evidence. But the quality of the local set (how closely the documents in the set are related to the topic) converges towards to quality of the global set as the size of the local set grows. Of course, there will be differences between individual topics. Some topics have only a handful of relevant documents, while others might have several hundreds. Based on experiments, we found the top 100 results to be a suitable level of focus for our research questions. Our goal is not to find the optimal size of the local set, but to understand the nature of local and global links.

3.2 ANALYSIS OF WIKIPEDIA LINK STRUCTURE

In this section, we analyse the link structure of Wikipedia. More specifically, we look at the incoming and outgoing link degree of Wikipedia articles.

3.2.1 *Degree distribution*

Is the link structure of Wikipedia different from the link structure of the Web? Recall that the mechanisms of generating links in Wikipedia are unlike those of the larger Web. Does the encyclopedic organisation, where there is little redundant information, put a bound on the number of incoming links? Does the organisation in mono-topical entries or lemmas restrict the number of outgoing links? We look at the number of different incoming links (in-degree) and the number of outgoing links (out-degree).

Figure 4 shows the incoming and outgoing link degree distributions of Wikipedia. Because few pages have a high in-degree (data sparsity), we use the complementary cumulative distribution function (CCDF) to obtain a smooth curve. The CCDF shows at each particular degree x, the prior probability $P(X \geq x)$ of an article having a degree of at least x. For example, if 25 out of 100 articles have 10 or more incoming links, the prior probability of an article having at least 10 incoming links is $P(X \geq 10) = \frac{25}{100} = 0.25$. All articles necessarily have an in-degree of $X \geq 1$, so $P(X \geq 0) = 1$. Note that strictly speaking, the CCDF is a

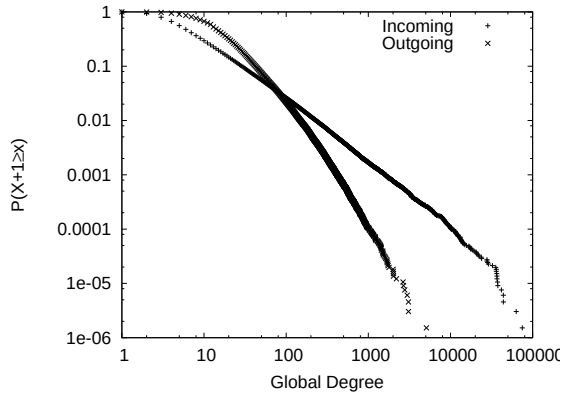

Figure 4: Cumulative distribution of link in-degree and out-degree distribution over 659,388pages in the INEX Wikipedia collection

continuous function while the degrees form a discrete distribution. The resulting probability distribution is therefore also discrete. Because we plot on the log-log scale, $x = 0$ cannot be shown (the log of zero is $-\infty$). For all figures with degree distribution we therefore assume that each page has a self-referencing link. From a navigational perspective, link degrees reflect the accessibility of pages (the number of pages that can be reached from a particular page or the number of pages from which a particular page can be reached). Since any page can be reached from itself, all pages can be said to have a minimum in- and out-degree of one. We use this assumption and plot $X + 1 \geq x$ so that articles with no incoming links are visible in the figures.

The in-degree is shown on the top, and the out-degree is shown on the bottom. Both curves approximate straight lines on the log-log scale, suggesting a power law distribution—the number of documents with a certain degree are a power of that degree[1]—that is familiar from the Web at large (Faloutsos et al., 1999). This is also in line with earlier studies of the link structure of Wikipedia (Bellomi and Bonato, 2005, Capocci et al., 2006, Voss, 2005) and the idea that Wikipedia link generation adheres to the notion of preferential attachment (Barabási and Albert (1999), see Section 2.2.4). In an encyclopedic collection, preferential attachment can be linguistically motivated through term statistics; encyclopedic entries about very general concepts will have

1 The power law distribution is actually given by the Probability Distribution Function $P(X = x)$ which often has a 'messy' tail. For reasons mentioned above, we plot the CCDF $P(X \geq x)$

a high in-degree because these concepts are mentioned often. The flat part of the distribution at the low out-degrees can be explained by the general rule in Wikipedia that articles with few outgoing links should either be extended or deleted. Zlatic et al. (2006) found power law exponents of $\gamma = 2.21$ for the in-degree distribution and $\gamma = 2.65$ for the out-degree distribution of the English Wikipedia of January 2005. The higher exponent of the out-degree means the out-degree distribution follows a line that falls faster than that of the in-degree, which is the same in our data. Buriol et al. (2006) analysed the in-degree distributions of 17 snapshots of the English Wikipedia between January 2002 and April 2006 and found that the power law exponent is remarkably stable, around $\gamma = 2.00$. The in-degree distribution shows a straight line from the the lowest degree all the way down, while the out-degree distribution is flat at the low degrees and starts to fall rapidly at degrees higher than 10. This means that most articles have at least a handful of outgoing links. The highest out-degree is much lower than the highest in-degree. There are two aspects of the guidelines that can help explain this difference. First is the guideline on long articles, which states that very long entries should be split up into multiple shorter entries.[2] This puts a natural curb on the number of outgoing links an article can have. The second guideline warns against *overlinking*—the creation of obvious, redundant and useless links—and *underlinking*, in which case an article is not linked to related articles that help readers understanding the article and its context.[3] Incoming links on the other hand have no such restrictions. An article that provides relevant context for many other articles can and often will be linked from all those articles. In the rest of this chapter, we focus on in-degrees.

What is the degree distribution of relevant pages? Figure 5 plots the CCDF for the whole collection and for the subset of articles relevant for any of the 221 INEX 2006–2007 topics. We see a straight line again, suggesting another power law distribution. The relevant pages are also spread across the degree distribution, showing that both pages with low and high in-degree can be relevant. However, the slope of the distribution of relevant pages is less steep. This means the set of relevant pages is less dominated by pages with low in-degree.

2 See `http://en.wikipedia.org/wiki/Wikipedia:Splitting`.
3 See `http://en.wikipedia.org/wiki/Wikipedia:Linking`.

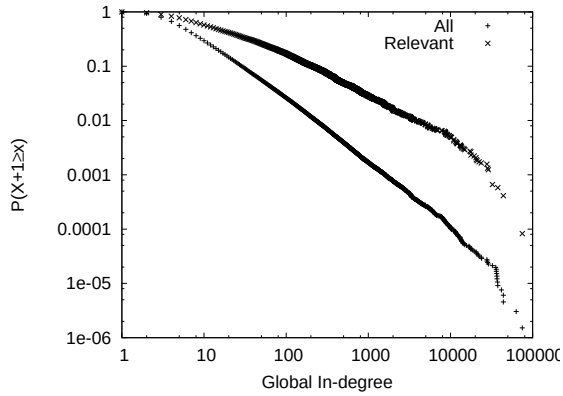

Figure 5: Link in-degree CCDF of the entire Wikipedia collection and the 12,107 "relevant" pages.

3.2.2 *Local degree distribution*

So far, we have looked at global evidence provided by the absolute number of links. We now zoom in on local evidence provided by the number of links among a subset of local pages. We used our baseline text retrieval system (discussed in detail in Section 3.1) to find the top 100 matching Wikipedia articles for each of the 221 topics. We treat these pages as local context, as they are more focused on the topic, and only consider links between pages in this subset and ignore all further links. By restricting our view to the local context, a large fraction of these local links should relate to the topic at hand. Is this local structure different from the global link structure investigated above?

For the 221 topics, a total of 22,016 articles are retrieved, of which 4,796 are relevant. The local degree distribution is shown in Figure 6. Again, the plot suggests a power law distribution, similar to the finding of Dill et al. (2002). They selected subgraphs of the Web based on domain restrictions or the occurrence of keywords and found that topically focused subsets have similar degree distributions as the overall set of Web pages in their collection. Chakrabarti et al. (2002) used the DMOZ (DMOZ, 2010) classification to study the degree distribution of pages in topic-specific subsets of Web pages and found that the degrees followed a power law distribution. The Wikipedia link structure shows no fundamental difference with the Web in this respect.

In the same Figure 6, we zoom in on only those top 100 retrieved articles which are relevant for their respective topics. Here we see

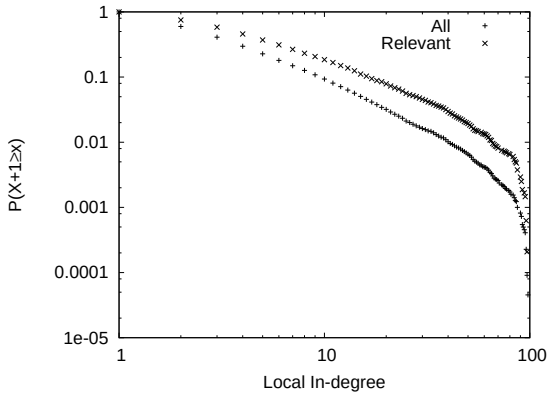

Figure 6: Wikipedia local link in-degree CCDF of 22,016 local pages and of 4,796 local relevant pages.

a similar distribution which also shows that both low and high in-degree pages can be relevant: local in-degree is not absolute evidence of relevance. Again, the degree distribution of the relevant pages diverges from the the distribution over all pages at higher degrees. Within the set of retrieved relevant pages, there are relatively few pages with low in-degree.

3.2.3 *Prior probability of relevance*

Above, we saw that neither global nor local in-degree provides absolute evidence of relevance. But can global or local in-degree be used as a (possibly weak) indicator of relevance? That is, if we would know nothing more of a page than its global or local in-degree, can we make an educated guess about the relevance of the page?

For a page of a given in-degree, we can calculate the prior probability that it is relevant (with respect to at least one of the INEX topics). We do this as follows. We used the cumulative degree distribution of the relevant pages to determine the data points. At each data point (x, y), we divide the number of relevant pages having a degree $X \geq x$ by the total number of pages having a degree $X \geq x$:

$$P(\mathrm{R}|X \geq x) = \frac{|\{X_{\mathrm{rel}} \geq x\}|}{|\{X \geq x\}|}$$

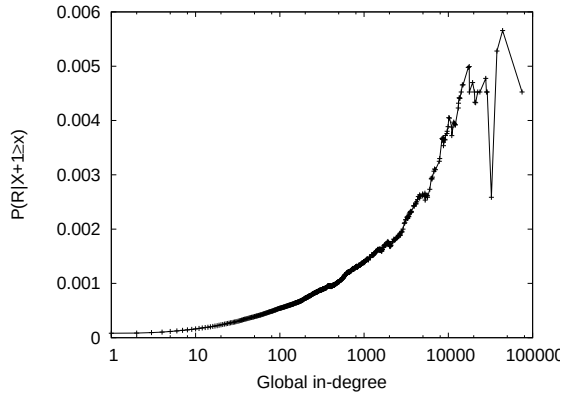

Figure 7: Prior probability of relevance of Wikipedia global in-degree.

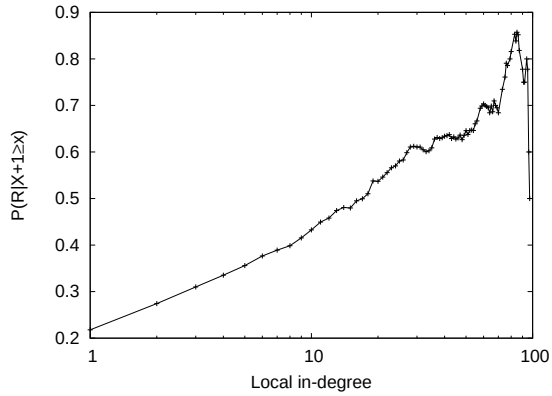

Figure 8: Prior probability of relevance of Wikipedia local in-degree.

where X_{rel} is the degree of relevant documents. If in-degree is positively related to relevance, we expect to see the prior probability of relevance increase as the in-degree increases.

In Figure 7 we see the prior probability of relevance of global in-degree. We see a clear increase in the prior probability of relevance with increasing global in-degree. Although there are more relevant pages with a low in-degree than with high in-degree (as was shown in Figure 5), this number is dwarfed by the total number of pages with a low in-degree (as shown in Figure 4), leading to a relatively low prior probability of relevance—around 0.0001. Conversely, although the number of relevant pages with a high in-degree is modest, this is still

a substantial fraction of all the pages with a high in-degree—around 0.004.

We do the same analysis for the local in-degree, shown in Figure 8. The prior probability of relevance also clearly increases with local in-degree. Again, although the absolute number of relevant pages with a low local in-degree is higher (as shown in Figure 6), a larger fraction of pages with a high local in-degree is relevant. The prior probability of relevance rises from 0.22 to around 0.85 for pages with a high local in-degree.

Of course, the local degree probabilities of relevance are much higher than the global degree probabilities. The local set is based on the text-based relevance score and is therefore much richer in relevant pages—22% of all pages in the local set is relevant while only 0.008% of the pages in the entire collection is relevant for a single topic.

3.2.4 *Naive reranking*

We selected one topic to look in detail at what happens to the top results when naively reranked by in-degree. Topic 339 has title *Toy Story*, and is about the computer animated movie from 1995. We compare the top 10 articles of the baseline run with the top 10 articles ranked by global and local in-degrees in Table 4. The top 10 results of the baseline (top left) and global all (top right) are based on all retrieved results (all documents that match the query to some extent). The bottom two lists are based on the top 100 results ranked by the baseline, and are reranked by global in-degree (bottom left) and local in-degree (bottom right). The top 10 articles of the baseline are clearly focused on the topic. Three articles are about the *Toy Story* films, one about a character from the films (*Buzz Lightyear*) and a few about people and companies involved in making the films (*Joe Ranft*, *Andrew Stanton*, *John Lasseter* and *Pixar*). Although the *Toy Story 2* article is judged as not relevant, it is closely related to the search topic.

The top 10 articles based on global in-degree are all articles with extremely high in-degrees that are completely off-topic. There are many articles containing either of the terms *Toy* and *Story* and most of them will be irrelevant. This top 10 gives a good insight into what type of articles have high in-degrees. Many Wikipedia articles mention dates and locations. Link bots automatically add links to articles about these dates and countries, resulting in high in-degrees for the date and country pages. If we use the in-degree to rank all results matching the

Title	Baseline	Title	Global all
Toy Story 2	7.84e-08	2005	44,025
Toy Story	7.07e-08	2002	27,780
Buzz Lightyear	6.18e-08	Japan	22,280
Toy Story 3	6.14e-08	Australia	20,405
List of Disney animated features' titles in variouslanguages [sic]	5.44e-08	1980	13,700
Pixar	3.07e-08	New York City	13,661
Joe Ranft	1.66e-08	India	13,318
Andrew Stanton	1.51e-08	1983	13,105
Little Bo Peep	1.21e-08	1975	12,209
John Lasseter	1.14e-08	1969	11,429
Title	Global top 100	Title	Local top 100
Toy	331	Toy Story	34
Computer-generated imagery	310	Toy Story 2	26
Pixar	177	Pixar	20
G.I. Joe	164	Toy	15
Transformers series	121	Monsters, Inc.	9
Toy Story	109	Buzz Lightyear	9
Aladdin (1992 film)	106	John Lasseter	7
Monsters, Inc.	79	Cars (film)	6
Toy Story 2	61	John Ratzenberger	6
Tim Allen	54	Computer-generated imagery	6

Table 4: Top 10 Wikipedia articles for topic 339 "Toy Story" ranked by content baseline (top left), global in-degree over all retrieved results (top right), global in-degree over baseline top 100 (bottom left) and local in-degree over baseline top 100 (bottom right)

query, the top of the ranked list is *infiltrated* by important but off-topic articles about countries and dates.

With *infiltration* we mean that articles for which there is little query-dependent evidence are pushed high up the ranking because the query-independent evidence indicates these articles are very important or of high quality. This happens when query-independent evidence dominates query-dependent evidence. If the score distribution for some query-independent feature is highly skewed, straightforward implementation of that feature will radically alter the ranking. This is the case with the global degrees, where many articles have no or only few incoming links and some articles have thousands of incoming links. Even if a completely off-topic article has a very low text-based score, if it has a very high in-degree, the in-degree prior will push this article to the top of the ranking.

If we restrict the reranking on global in-degrees to the top 100 results of the baseline run (bottom left of Table 4), we keep more focus on the topic. The top 10 ranked by global in-degree now contains several on-topic and related articles, but there is still a strong widening of scope; the top ranked article *Toy* shows infiltration is still a problem.

If we instead use only the links between the articles in the top 100 to rerank results (bottom right of Table 4), the top 10 results are much more focused on the topic. The first result, *Toy Story*, is clearly on-topic (and judged relevant) and has swapped position with *Toy Story 2* which is ranked first by the baseline. Several other articles in the list are closely related as well. However, we still see some infiltration. The article *Toy* has quite a few incoming links from other articles in the top 100 results, but the effect is much smaller than with the global degrees.

This qualitative analysis suggests that global and local in-degree are weak indicators of relevance. Therefore, in re-ranking, their weight should be small compared to the weight of the content-based retrieval score.

Summarising, our analysis of the link structure reveals that the Wikipedia link structure is a (possibly weak) indicator of relevance of Wiki pages. A naive re-ranking based on only global or local in-degree is not effective: it leads to infiltration by important but off-topic pages.

3.3 INCORPORATING LINK EVIDENCE

In this section, we discuss how link evidence can be incorporated in our retrieval model.

3.3.1 *Link degree priors*

Recall from Section 3.1 that we use a document length prior in our retrieval model. We can treat the link in-degree as another prior probability and incorporate this as a second prior in computing the final relevance score:

$$P(d|q) = P_{\text{link}}(d) \cdot P_{\text{length}}(d) \cdot \prod_{t \in q} ((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d))$$

where $P_{\text{link}}(d)$ is the in-degree (global or local) of document d. In Section 3.2.3 we computed the prior probability of relevance for in-degrees, which we could also use as probabilities instead of the degrees themselves. Experiments have shown their impact is similar to using degrees, so there is no clear advantage of using the probabilities of relevance over using the degrees themselves as priors. A disadvantage of using the probabilities of relevance is that in many situations, relevance information will not be available, whereas degree information is. On top of that, if we use probabilities trained on relevance information, it is harder to interpret the impact of link evidence on retrieval performance. We will for convenience refer to the link evidence as a prior probability, even though we do not transform the degrees into a probability distribution. Note that we can turn any prior into a probability distribution by multiplying it with a constant factor $\frac{1}{\Sigma_{d \in D} Prior(d)}$, leading to the same ranking.

The qualitative analysis in the previous section suggests we need to be careful when incorporating link evidence. We do not want to retrieve pages that only have a high link score, i.e., pages that may be important overall but unrelated to the topic of request. Hence, we experiment with using the link degrees as real priors, i.e., apply them to all retrieved results, as well as using them to re-rank the top 100 results alone. The latter is a safe-guard against infiltration of important pages from far down the ranking.

3.3.2 *Baseline*

Our baseline is the retrieval model described in Section 3.1.4 without using link evidence. To explain the impact of the link evidence, we look again in detail at Topic 339 and the effects of the priors on the top 10 articles. In the upper left corner of Table 5 the titles of the top 10 retrieved Wikipedia articles for the baseline run are repeated for easy comparison.

Baseline run	Global in-degree prior
Toy Story 2	Toy Story
Toy Story	2002
Buzz Lightyear	Pixar
Toy Story 3	Toy Story 2
List of Disney animated features' titles in various languages	2005
Pixar	1980s
Joe Ranft	1970s
Andrew Stanton	Television
Little Bo Peep	1975
John Lasseter	1990s
Global top 100 in-degree prior	**Local top 100 in-degree prior**
Toy Story 2	Toy Story
List of Disney animated features' titles in variouslanguages	Toy Story 2
Toy Story	Pixar
Pixar	Buzz Lightyear
Buzz Lightyear	Toy Story 3
Toy Story 3	John Lasseter
Joe Ranft	Andrew Stanton
Secret Wars	List of Disney animated features' titles in variouslanguages
G.I. Joe	Joe Ranft
Modern animation of the United States	Cars (film)

Table 5: Top 10 Wikipedia articles for topic 339 "Toy Story"

3.3.3 *Global in-degree*

The *global in-degree prior* is proportional to the global degree of an article:

$$P_{\mathsf{Glob(d)}} \propto 1 + global(d)$$

Our qualitative analysis of a single topic showed that even in combination with the text-based retrieval score, the global degrees still lead to infiltration. Therefore, we also experiment with a conservative *log global in-degree prior*:

$$P_{\mathsf{LogGlob(d)}} \propto 1 + \log(1 + global(d))$$

Inspecting our running example immediately confirms that we need to be careful when incorporating global link evidence. In the upper right corner of Table 5, we see that the combination of retrieval score and global in-degree prior has pushed the article *Toy Story* to the top of the ranking. But it has also pushed up many important but off-topic articles. As expected, using the global degree on all retrieved results leads to infiltration. In the bottom left corner we see the impact of the global in-degree prior on the ranking of the top 100 results. The top of the ranking suffers much less from infiltration, and the ranking stays more focused on the topic. But the article *Toy Story* is pushed down one rank in favour of the only vaguely related *List of Disney animated features' titles in variouslanguages* [sic] and there are still some unrelated articles pushed to the top, such as *Secret Wars* and *G.I. Joe*. The impact of the global in-degree is still too big.

3.3.4 *Local in-degree*

The *local in-degree prior* is proportional to the local degree of an article:

$$P_{\mathsf{Loc(d)}} \propto 1 + local(d)$$

Alternatively, we use a conservative *log local in-degree prior*:

$$P_{\mathsf{LogLoc(d)}} \propto 1 + \log(1 + local(d))$$

The local link graph and the local degrees are query dependent. Therefore, the prior is $P_{\mathsf{Loc(d)}}$ is not actually a prior, but another $P(d|q)$.

When we combine the retrieval scores and local in-degrees (bottom left of Table 5), the non-relevant article *Toy Story 2* and the relevant article *Toy Story* swap places. The relevant article *Pixar* is also pushed up. The reranking based on local in-degrees alone suffered from infiltration, but the combination of the local in-degree prior and the content-based score pushes the articles *List of Disney animated features' titles in variouslanguages* and *Little Bo Peep* down in favour of articles about the creators of the film. Average precision increases from 0.1434 to 0.2545.

3.4 EXPERIMENTAL RESULTS

In this section we discuss the results of using the degree priors for an ad hoc document retrieval task.

We report Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision at rank 10 (P@10) and Precision at rank 30 (P@30).

Run id	MAP	MRR	P@10	P@30
Baseline	0.3157	0.8119	0.4937	0.3621
All results:				
Global	0.2306$^{\bullet}$	0.7737°	0.4041$^{\bullet}$	0.2786$^{\bullet}$
Log Global	0.3180	**0.8341**$^{\circledcirc}$	0.4986	0.3627
Top 100:				
Global	0.2871$^{\bullet}$	0.7899	0.4439$^{\bullet}$	0.3376$^{\bullet}$
Log Global	0.3192°	**0.8341**$^{\circledcirc}$	0.4986	0.3626
Local	**0.3272**$^{\bullet}$	0.8249	**0.5081**$^{\circ}$	**0.3769**$^{\bullet}$
Log Local	0.3243$^{\bullet}$	0.8288°	0.5009°	0.3727$^{\bullet}$

Table 6: Results of using link evidence on the 221 ad hoc topics of the INEX 2006-2007 Ad Hoc tasks. Best scores are in bold-face. Significance levels are 0.05 ($^{\circ}$), 0.01 ($^{\circledcirc}$) and 0.001 ($^{\bullet}$), bootstrap, one-tailed.

To determine whether the observed differences between two retrieval approaches are statistically significant, we used the bootstrap method, a non-parametric inference test (Efron, 1979, Savoy, 1997). We take 100,000 resamples and look for significant improvements (one-tailed) at significance levels of 0.95 ($^{\circ}$), 0.99 ($^{\circledcirc}$) and 0.999 ($^{\bullet}$). The results are shown in Table 6.

As expected, the global in-degree prior has a disastrous effect on the relevance ranking when applied to all retrieved results. The extreme differences in degree seem to dominate in the final ranking and lead to a highly significant decrease in performance. The log of the degrees has a much flatter distribution and therefore a much smaller impact. At the top, the ranking even improves slightly, with a significant improvement in MRR. Gently pushing up the most cited article in the top clearly has a positive effect. The increase in performance becomes smaller as we go down the ranking. At rank 30 the impact on precision is negligible.

If we use the normal (non log) global degrees to only re-rank the top 100 retrieved results, performance drops significantly for all measures. The amount of infiltration is reduced compared to when applying the degree priors on all retrieved results, but is still severe. With the log degree priors we again see a significant improvement in MRR—in fact, the improvement is exactly the same as that of the log priors used on all retrieved results. The improvements in the top of the ranking now lead to a small but significant improvement in MAP.

The local degree priors improve the ranking according to all reported measures. Although the improvements in MRR are small and not significant, the improvement in precision up to rank 10 is larger and significant. Up to rank 30, the improvement in precision is even bigger and more significant, leading to a significant improvement in average precision. The combination of text-based score and local in-degree gives a better relevance ranking than the text-based score alone. This shows that local link evidence is complementary to text-based evidence. Curbing the local in-degree prior by taking the log reduces this complementary, positive impact. The relationship between local in-degree and relevance is strong enough to be used linearly as evidence.

This is in direct contrast to the results of the TREC Web Tracks of 1999 and 2000, but more in line with the early expectations of the Web Track participants. Links in Wikipedia seem to better reflect the topical relation between pages than hyperlinks between Web pages in general. These findings finally provide us with a new and possibly more fruitful way to analyse when and why links are effective and ultimately with a better understanding of the value of hyperlinks for retrieval.

3.4.1 *Per topic analysis*

The qualitative analysis of the *Toy Story* example is only anecdotal, to illustrate the problem of infiltration. The question remains whether it is illustrative of the general behaviour of link degrees. In this section, we analyse the impact of the link degrees on the scores reported in Table 6. For how many topics does the global degree affect early and overall precision? How often does the ranking improve with local degrees? Table 7 shows the number of topics for which average precision (AveP, see Section 2.1.2.3) and P@10 goes up, down or stays the same as a consequence of using link evidence.

We first look at the normal degree priors. The average precision per topic (AveP) almost always changes. Only for a few topics the average precision stays the same. The global degrees affect the average precision of all topics. Over all results, they improve performance on 43 topics (19%) and hurt performance on 178 topics (81%), hence, over all topics, MAP decreases. Using global degrees on all results leads to major infiltration. Over the top 100 results, the results are slightly better, but still performance drops for the majority (65%) of the topics. The local degrees improve the ranking for 66% of the topics and thereby the MAP. Infiltration is much less of a problem using local degrees, although still present, given that performance drops for 74 topics.

	AveP			P@10		
Measure	↑	↓	=	↑	↓	=
Global all	43	178	0	42	120	59
Global top 100	77	144	0	46	100	75
Local top 100	146	74	1	66	48	107
Log Global all	125	94	2	43	33	145
Log Global top 100	127	87	7	43	33	145
Log Local top 100	163	54	4	41	27	153

Table 7: Per topic comparison of baseline and link evidence runs, showing the number of topics for which in-degree evidence scores are better, worse or tied with the baseline scores.

Up to rank 10, the impact of the degrees is smaller. The number of topics for which P@10 changes is bigger for the global degrees than for the local degrees, and bigger when re-ranking all results than when re-ranking only the top 100. This is not surprising given the much larger variation in the global degree priors. The highest global degree is much higher than the highest local degree, while they have the same minimum. The more results we re-rank, the more impact on the top of the ranking. Apart from that, the pattern is the same as for AveP. The problem of infiltration becomes smaller as we make the link evidence more sensitive to the topical context.

The log of the degrees curbs the impact of link evidence, which is especially clear for P@10. For the majority of topics, the P@10 score stays the same. The log compresses the range of the degree priors, so that the ranking changes only subtly. Interestingly, the global degrees lead to more improvements (43) than the local degrees (41). For AveP, global and local degrees improve performance for the majority of the topics, but the local degree does substantially better. It is also clear that the log of the global degree is insensitive to the context in which it is used. Whether we re-rank all results or only the top 100, the impact is very similar for AveP and exactly the same for P@10.

We analyse the top results for three topics with the largest drop in average precision due to the local degree priors.

- Topic 381 is titled *ubiquitous computing and application*. The average precision drops from 0.4340 to 0.3163. In the top results, the relevant articles change from ranks 1, 2, 3, 6, 7, 8 to 1, 4, 6, 7, 8, 15 because

Computer and *Wearable computer* infiltrate from ranks 10 and 9 to rank 2 and 3 (some other changes occur as well, but have a smaller impact). A user might consider the article *Computer* irrelevant because it is too general, but the article *Wearable computer* does not seem like a bad result.

- Topic 471 is titled *Three greatest river +Japan*. The AveP drops from 0.4167 to 0.2562. The relevant articles change from ranks 2, 3, 6 to 4, 5, 8 because *Japan* and *River* infiltrate from ranks 22 and 12 to ranks 1 and 2. Again, these infiltrations are not obvious. When a user looking for the three greatest rivers of Japan sees the title *Japan* in a list of search results from Wikipedia, it is plausible that she considers it a promising result and clicks through to see if it contains information on Japan's three greatest rivers. An encyclopedia article titled *River* is more easily considered too general.

- Topic 344 is titled *xml database*. The AveP drops from 0.3262 to 0.2260. The first relevant article changes from rank 1 to rank 2 because *Database* infiltrates from rank 8 to rank 1. From the title, it is not clear that the article *Database* contains no relevant information on XML databases.

The failure analysis of these topics shows that local degrees still lead to infiltration, although the infiltrating articles are often tangentially or even closely related to the search topic. This illustrates the fragility of relevance ranking. The textual evidence also ranks closely related but irrelevant documents highly. Although the local link evidence seems to focus fairly well on the search topic, the link structure remains blind to the precise content and nature of the query and sometimes fails to pick out only the right documents.

3.5 DISCUSSION

In Section 3.2.3 we saw that both global and local link in-degrees are related to relevance, yet global link evidence is hardly effective for ad hoc retrieval. In fact, Figures 7 and 8 suggest that global in-degree has a stronger relation with probability of relevance than local in-degree. Why then is this not reflected by their impact on the relevance ranking? The most important documents, that is, the documents with the largest number of incoming links, have the highest prior probability of being relevant, but in a search task that requires finding all topically relevant documents, many important but non-relevant documents will also be pushed to the top of the ranking by the global in-degree. From

the results in Table 6 it seems clear that link evidence must be made sensitive to the topic to be made effective for identifying topically relevant documents.

The global degree distribution showed that most documents have a very low in-degree and only a few have a very high in-degree. Because of the sparsity at the high end, only a few articles have to be relevant for one or more topics to attain a high prior probability of relevance, but the signal is very weak. The signal of the local degrees is stronger, as reflected by the higher absolute probabilities. Close to local in-degrees of 100, the prior probability of being relevant is close to 1. In other words, a very high local in-degree means we can be almost certain that an article is relevant.

3.6 CONCLUSIONS

In this chapter we described our experimental set-up and ran initial experiments to address the question:

- Can the link degree structure of a semantically linked document collection be used as evidence for the relevance of ad hoc retrieval results?

We first looked at the degree distributions of incoming and outgoing links and found that both global and local degree distributions adhere to a power law. The in-degrees of the relevant pages have similar distributions, but with a less steep slope. We then analysed the relation between link in-degree and relevance. From the degree distributions over all pages and only the relevant pages, we computed the prior probability of relevance over in-degrees and found that both global and local in-degrees are related to relevance. A higher in-degree means a document has a higher prior probability of being relevant. This relation could be exploited in the ranking of retrieval results.

By looking at the impact of ranking by in-degree on an example topic, we identified the problem of infiltration of pages with high in-degree that are not related to the search topic. This problem is larger for global degrees than for local degrees, because local degrees are more sensitive to the topical context.

The second part of this chapter described a retrieval experiment using the INEX Ad Hoc topics, with the aim of establishing the effectiveness of link evidence for ranking retrieval results. We used the in-degrees as prior probabilities in our language model and found that global in-degrees lead to a decrease in performance unless curbed by taking

the log of the in-degree. The local degrees keep more focus on the topic and lead to significant improvements on early and overall precision. Because they are filtered on the search topic, there is no need to curb their impact by using the log of the degrees.

A more in-depth analysis of a few topics for which the local degrees hurt performance showed that some articles with high local in-degree contain no relevant text (and are therefore judged irrelevant) even though they are closely related to the topic. Even the local, query-dependent link structure is blind to the precise content of the query, and as a result, promotes some articles that are related to many of the top ranked articles, but not relevant to the topic of request.

Our main finding is that we have found link evidence to be effective for improving ad hoc retrieval in Wikipedia. This positive impact of link evidence matches the initial expectations of the TREC Web Track organisers and participants but lives in contention with their findings. In the Web, links were found ineffective for ad hoc retrieval, but very effective for other search tasks. This discrepancy urges us to compare the link structure of Wikipedia with that of the larger Web to see if the difference in impact on retrieval can be attributed to structural differences in their link graphs.

4

IS WIKIPEDIA LINK STRUCTURE DIFFERENT?

In the previous chapter we saw that the link structure of Wikipedia can be used to improve the ranking of ad hoc retrieval results. Although Wikipedia is part of the larger Web, similar experiments on Web test collections failed to show the effectiveness of link information for ad hoc retrieval. This immediately poses the question:

- Is Wikipedia link structure different from the link structure of the World Wide Web?

Both the Wikipedia and Web link structures have been extensively studied, but never compared for the purpose of information retrieval. Although links on the Web tend to connect pages with topically related content (Davison, 2000), this signal of general Web links might be too weak to improve the subtle topical relevance ranking of content-based retrieval for topic search (Kraaij and Westerveld, 2001). However, for more Web-centric tasks, the value of link information has been demonstrated (see Section 2.3). Link information seems to have different roles for Web and Wikipedia retrieval.

To study these different roles, we should look at the search scenarios where link evidence has a positive impact. Therefore, we compare the impact of link information on Wikipedia topic search and on a Web collection using Web-centric search tasks.

Our aim is to analyse the value of link evidence for information retrieval. Therefore, we want to approach the main question in this chapter from an information retrieval perspective and look at the relation between link evidence and relevance. This means we need sets of search requests and associated relevance judgments. We already described the Wikipedia test collection in Section 3.1 on page 46. For the Web, we use the TREC 2004 Web Track collection, consisting of 225 topics and the 1.2 million documents .GOV collection. The 2004 Web Track consisted of three differenet tasks. Home Page finding, Named Page finding and Topic Distillation (all described in Section 2.3.2 on page 31) topics are a mix of 75 Named Page, 75 Home Page and 75 Topic Distillation topics. Although this collection provides us with the necessary topics and relevance judgements, and is reasonably comparable in size and number of topics to the Wikipedia collection, it is

a relatively small crawl of a specific domain. We make no particular claims on the representativeness of this data set for the current Web, which is infinitely large and highly heterogeneous, but expect it to be a close enough approximation for our purposes (Soboroff, 2002).[1]

Our main research question breaks down into two parts. We start by investigating the Wikipedia link structure with an extensive comparative analysis of the two IR test collections, Wikipedia and .GOV. Specificially, we want to know:

- What is the link density of the Wikipedia and the .GOV collections?
- Are there differences in connectedness?
- What is the degree distribution of Wikipedia and the .GOV collections?
- Are there differences between distributions of incoming and outgoing links?
- And, in particular, how does the link topology relate to the relevance of retrieval results?

The second part of our main research question is about the effectiveness of link-based evidence. There are several, more complex link-based ranking algorithms, such as HITS and PageRank, that try to model aspects of popularity and authority. On top of that, HITS uses an expansion step to rank documents that link to the top retrieved documents, but were not retrieved themselves. These algorithms could provide additional insight into the difference between Web and Wikipedia links. More specifically, we want to know:

- What is the impact of link evidence on .GOV and Wikipedia retrieval?
- How do more complex link-based ranking algorithms compare against simple link degree counts in terms of retrieval effectiveness?

The rest of this chapter is structured as follows. Next, in Section 4.1 we discuss differences between Web and Wikipedia pages that could have consequences for their respective link structures. In Section 4.2, we perform a comparative analysis of the link structure of Wikipedia

1 The more recent ClueWeb09 collection is much larger and based on a more recent crawl than the .GOV and therefore arguably better represents the current Web. Unfortunately, this collection was not yet available at the time of writing. Since the .GOV collection was the first collection where the effectiveness of link-based methods was shown for Web-centric tasks and makes it an appropriate collection to study the difference between Wikipedia and Web link structure. However, the introduction of the ClueWeb09 collection urges us to revisit the importance of link evidence for ad hoc search in Chapter 7.

and .GOV and the relation between the link topology and the relevance of retrieval results. Then, in Section 4.3, we perform a range of retrieval experiments, investigating the impact of link evidence on retrieval effectiveness. We end this chapter in Section 4.4 where we summarise our findings and draw conclusions.

4.1 THE NATURE OF WEB AND WIKIPEDIA DOCUMENTS

It is tempting to speculate about differences between the internal link structure of Wikipedia and the link structure of the Web at large, and how these may affect the value of link based methods. There are a number of aspects of Wikipedia that make link generation different from link generation in the Web:

Encyclopedic organisation: As an encyclopedia, Wikipedia has low redundancy, where each topic has its dedicated page.

Style: Wikipedia entries are written in an informational, objective style with the aim of describing a particular topic. The guidelines urge authors to keep a neutral point of view. This fits the task of ad hoc retrieval which models informational or topic search and focuses on topical relevance. Web pages can be and are written in any possible style, to inform, explain, show, argue, discuss, prove, entertain, tempt, sell and even insult.

Shared authorship: In Wikipedia, everyone can create, modify and remove links, the content of the linking page and the content of the linked page. If an obvious or useful link is "missed" by one author, there are thousands of others, including link bots, to add the link. In the Web, document authors are often on their own. With hundreds or thousands of Web pages on the same topic, an author feels no need to link to all of them, and moreover, probably has no knowledge about the existence of most of them.

Linking guidelines: There are guidelines in Wikipedia explaining what should be linked, when links should be made and where links should point to. In the Web, document authors can create links for any reason, to any page and can omit or miss obvious, useful links.

Single domain: Wikipedia is a single domain, essentially giving all pages equal authority. Within a single domain or site, the site owner(s), or the author(s) creating the content, is in full control of the site-internal links and links pointing to pages on other sites, but has

no or very little control over the external links pointing *to* pages in their own site. Site-internal links are often considered not to confer authority (Kleinberg et al., 1999) because the source and destination pages are written by the same (group of) people.

Flat domain: The encyclopedic articles of Wikipedia have no explicit hierarchical structure. Only the main page is a clear entry page, all other pages are stored in the same directory and thus have the same URL depth. In the Web, individual Web sites often have a hierarchical structure reflected by the directory structure of the file system of the servers running those Web sites. Within a Web site, a single broad topic might be broken down into several aspects and facets using individual pages for each of these aspects and facets, in separate directories on the server. Because of this hierarchy, a site-external link to the entry page of a Web site might be an endorsement for the content of the entire Web site, or a signal that the content of the entire Web site is topically related to the content of the source page.

Size: The English Wikipedia is tiny in comparison to the vast Web, which has tens of billions of static Web pages and an infinite number of dynamically generated pages (Baeza-yates and Castillo, 2005).

How could these differences lead to differences in the link structures of Wikipedia and the World Wide Web in general?

First, as suggested by the *linking guidelines*, links seem to signal a semantic relation between pages rather than serve purely navigational purposes, and may therefore provide a strong source of evidence for the relevance of a given page.

Second, due to the *shared authorship*, *single domain* and *encyclopedic organisation* of Wikipedia, we may expect a far more complete link graph where all (or a large fraction of all) relevant links are present, leading to higher link density and connectedness of the link graph, and promoting the effectiveness of the link evidence. Another major consequence is that Wikipedia can be and is organised on a global scale while the larger Web can only be organised locally.

Third, due to the *encyclopedic organisation* and *shared authorship* the Wikipedia has relatively little redundant information. Topics have a single dedicated page, with a preferred title. Pages with alternative titles and duplicate information are quickly spotted by one of the many authors, and redirected or marked for deletion or merging.

Fourth, because of the low redundancy, the *encyclopedic organisation* also affects the size of Wikipedia. The already huge Wikipedia is dwarfed by the size of the Web at large, which may have a number of

consequences such as bounding the number of directly related incoming and outgoing links, as well as causing a quick loss of topical focus when traversing the link graph.

Fifth, given that we search within a single domain, the authoritativeness of individual pages is essentially the same, and the value of link evidence is primarily to signal topical relevance. On the highly heterogenous Web, link evidence may be used to signal other aspects of relevance, such as the general importance or authoritativeness of a site compared to other sites, or to indicate the best entry-page or entry-pages of the site.

Sixth, Wikipedia pages possibly have a higher average quality, due to *shared authorship*, *linking guidelines* and the *encyclopedic organisation*. Wikipedia is probably more spam-free, suffering less from nefarious tricks like link farming[2] and jokes like Google bombs.[3]

All these aspects point to a marked difference in linking between Wikipedia and the Web. In Wikipedia, authors are in control of both the incoming and outgoing links of an article, whereas in the Web, authors typically only control the outgoing links (at least the cross-site links). But because of the shared authorship, anyone and everyone can control both incoming and outgoing links of an article and any irrelevant, redundant or nepotistical link can and often will be removed. From this, we hypothesise that *in Wikipedia, outgoing links are very similar to incoming links and are therefore just as important.*

The different natures of the Web and Wikipedia also affect the types of search tasks typically performed on them. That is, in terms of the (Broder, 2002) taxonomy (see page 20), there are navigational queries (with the intent to reach a particular site), informational queries (with the intent to acquire some information present in some Web pages), and transactional queries (with the intent to perform some Web-mediated activity). Consider again the query "Mercedes-Benz". In the Web, this query is probably meant to locate the entry page of the Mercedes-Benz Web site. In this sense, the query is navigational. If typed in the Wikipedia search box, the underlying information need is probably different. The user might be looking for certain facts about the brand name or car models, or for a historical overview of the company's activities. We note that the latter could also be considered a case of known-item search (i.e., navigational) if the user knows or assumes the page to exist. Perhaps the difference between navigational and informational topics is less relevant in Wikipedia. At the same time, it

2 See `http://en.wikipedia.org/wiki/Link_farm`.

3 See `http://en.wikipedia.org/wiki/Google_bomb`.

strengthens our choice of using the .GOV collection to represent Web retrieval because it comes with a set of navigational topics, with the Named Page topics mixing navigational with informational aspects.

4.2 COMPARATIVE ANALYSIS OF LINK STRUCTURE

In this section, we look in close detail at the link structures of the Wikipedia and Web collections. We base our analysis on two IR test collections, consisting of a collection of documents, a large set of search requests and relevance judgments. For the Web, we take the .GOV collection used at the TREC Web Tracks of 2002-2004, which is based on a crawl of the `.gov` domain in early 2002.

Our analysis consists of four parts. We first give some statistics of the link graphs, then compare the connectedness of the two graphs. After that, we compare the degree distributions of the .GOV and Wikipedia collections and look at incoming and outgoing link degrees, both globally and locally. Finally, we look at the relation between link evidence and relevance.

4.2.1 *Web and Wikipedia graph statistics*

The .GOV collection contains 1,247,753 documents and 11,110,985 unique links between these pages (we ignore links which point to or from pages outside the collection). The Wikipedia collection contains 659,388 documents and a total of 13,602,613 unique links between these pages. We have also looked at how many of these links are reciprocal, i.e., a link from page A to B in combination with a link from page B to A. This is important with respect to our hypothesis that in Wikipedia incoming and outgoing links are similar to each other. A high number of reciprocal links would mean that the link graph is highly symmetrical and that pages have roughly the same number of incoming and outgoing links. There are 1,269,987 (11.4%) reciprocal links in the .GOV collection, and 1,182,558 (8.7%) reciprocal links in the Wikipedia collection. The Wikipedia link graph is far from symmetrical. Possible similarities between incoming and outgoing links are not caused by a tendency to create bidirectional links. The higher fraction of reciprocal links in the .GOV collection is likely due to the presence of navigational links within Web sites.

Table 8 gives some statistics on the incoming (in-degree) and outgoing (out-degree) links and document lengths of both collections. In .GOV the average number of incoming and outgoing links per document is 8.90, in

		min	max	mean	median	stdev
GOV	Indegree	0	44,228	8.90	1	126.00
	Outdegree	0	653	8.90	4	16.61
	Length	2	102,069	6,345	1,892	13,377
Wiki	Indegree	0	74,937	20.63	4	282.94
	Outdegree	0	5,098	20.63	12	36.70
	Length	16	281,150	2,473	1,309	4,238

Table 8: Statistics of the .GOV and Wikipedia collections

Wikipedia 20.63. Recall that, here, we are only using within-collection links, so every outgoing link is also an incoming link. The median number of incoming links is 1 in .GOV and 4 in Wikipedia and the median number of outgoing links is 4 in .GOV and 12 in Wikipedia. Also the maximal out-degree in Wikipedia (5,098) is much higher than in the .GOV collection (653). Given that Wikipedia has guidelines on article length and when and where to link, one might expect the maximum number of incoming and outgoing links to be relatively low compared the larger Web, which has no such guidelines and restrictions. However, no such bound is observed. Again, we make no particular claims on the .GOV collection being a good representative of the Web at large. On the one hand, the in-degrees should increase if we would consider a larger set of pages (since we cannot detect incoming links from pages outside the collection) leading us to underestimate the in-degrees. On the other hand, the limited crawl will likely have favoured pages with larger numbers of incoming links (e.g., how can a crawler find pages with no incoming links?) leading us to overestimate the mean in-degree. To put these numbers in perspective, Najork et al. (2007) use a Web crawl of 464 million pages and 18 billion hyperlinks, and find a mean in-degree of 6.10 and a mean out-degree (not limited to pages in the crawl itself) of 38.11.

We have also included character-length of pages in Table 8. The pages in .GOV have mean length 6,345 characters (median 1,892) and the pages in Wikipedia are shorter with a mean length of 2,473 characters (median 1,309).

The Wikipedia collection is thus more densely linked. This is surprising in the sense that the .GOV domain is much older, and link density tends to increase over time (Leskovec et al., 2005, 2007). There are at least two effects which help explain why the Wikipedia link graph is more "complete" than the .GOV link graph. First, due to the strongly

structured nature of Wikipedia and the existence of author guidelines, it is much clearer for Wikipedia authors where to link to and when. Second, due to peer editing and automatic link detection, "missed" links will be added in a matter of time.

4.2.2 *Connectedness*

The link density of graphs determines to a large extent the *connectedness* of the graph. Connectedness is a concept expressing the degree to which documents are reachable from each other via links. A graph is *connected* if there are no unconnected nodes. In an undirected graph, each node can be reached from any other node by following the links. A *component* is a subset of the graph that is *connected*. The connectedness of a graph is an important aspect for link-based propagation algorithms like PageRank, HITS and relevance propagation, and affects how visible pages are. A page that can be reached via hyperlinks from many other pages is in a sense more visible than a page that cannot be reached by following links at all.

How well-connected are the documents in the two collections? The connectedness of (random) graphs goes through phases (Janson et al., 2000). In the first phase, most nodes are isolated and only a few are connected to a few others in the form of trees. With a few more links, many more trees form and the first cyclic components start appearing. As more and more links are present, a single dominating component emerges, swallowing up many of the small trees and isolated nodes. With the equal numbers of nodes and links, such a *giant component* is already present. In the last phase, with N nodes and more than $\frac{1}{2}NlogN$ links, the graph is almost fully connected, with only a few nodes here and there that are unconnected.

For the Wikipedia collection the threshold for the final phase is 4,417,592 links, for the .GOV collection the threshold is 8,757,264 links. The actual numbers of links in Wikipedia (12,401,667) and .GOV (9,840,998) are higher than these thresholds. The Wikipedia collection is much further beyond this threshold than the .GOV. In terms of the phases of the connectedness of random graphs, we expect both collections to have a huge giant component containing almost all documents in the collection.

With the high link densities, we see a single giant component, i.e., a large set of connected pages. The giant strongly connected component (SCC) of the .GOV collection contains 912,877 (or 73.16%) documents and the giant weakly connected component (WCC) contains 1,209,325 (or

Collection	SCC	WCC
.GOV	912,877 (73.16%)	1,209,325 (96.92%)
Wikipedia	606,030 (91.91%)	657,601 (99,73)

Table 9: The size of the connected components of .GOV and Wikipedia

96.92%). The giant SCC of the Wikipedia collection contains 606,030 (or 91.91%) and the giant WCC contains 657,601 (or 99.73%). The WCC and especially the SCC of the Wikipedia collection contain a much larger part of the entire collection than the SCC and WCC of the .GOV collection. For both the .GOV and Wikipedia collections, the percentage of pages in the SCC is considerably larger than in the large crawl of (Broder et al., 2000). Soboroff (2002) finds that the .GOV collection structurally resembles much larger Web crawls but is very closely connected due to either starting the crawl from a small number of seeds or to the .GOV domain being more densely linked than the Web in general. The Wikipedia collection is a complete dump and the high link density cannot be a crawling artefact, as also observed by Buriol et al. (2006). The high connectedness might be explained by the shared authorship and encyclopedic organisation of Wikipedia. An author writing a new article can link to existing Wikipedia articles and can also add links to the new article by updating related pages. What does this mean for the link degree structure of the Wikipedia and .GOV collections?

4.2.3 *Global degree distributions*

In this part, we look at the degree distributions of the global and local links and particularly the differences between incoming and outgoing links. Our hypothesis is that in Wikipedia, incoming and outgoing links are very similar in nature. We again look at the cumulative in-degree and the out-degree distributions. We repeat some of the Wikipedia degree distributions shown in the previous chapter for ease of comparison. We use the same method as described on page 51.

The .GOV and Wikipedia in-degree distributions (left side of Figure 9) are surprisingly similar. Both show clear power laws, with Wikipedia having slightly higher probabilities because of its higher link density. In terms of graph evolution, both Wikipedia and the Web are scale-free networks. There are bigger differences between the out-degree distributions of the two collections (right side of Figure 9). In Wikipedia, the out-degrees follow a power law distribution above 10 outgoing

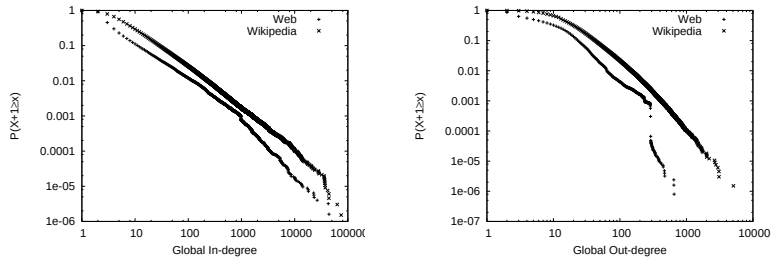

Figure 9: CCDF of the global incoming (left) and outgoing (right) link degrees of all pages for .GOV and Wikipedia.

links, suggesting that Wikipedia is also scale-free in its out-degree distribution. This might be caused by the presence of many list pages with links to all Wikipedia articles related to a particular topic, such as the article *List of environment topics*, which is extended when new articles on environmental topics are created. The .GOV out-degrees adhere less to a power law. Above 100 outgoing links, the curve becomes less steep and suddenly drops above 500 outgoing links. This might be an artefact of the collection.

In fact, the in- and out-degree distributions of Wikipedia show little difference apart from the slope of the distributions. This suggests that outgoing links indeed behave much like incoming links. This could be a consequence of having many reciprocal links—if a link from page A to B tends to be reciprocated by a link back from B to A, the graph would be symmetrical and in-degree and out-degree highly correlated—but we have already seen above that the Wikipedia collection has fewer reciprocal links than the .GOV collection. Note that this is also consistent with the idea mentioned in Section 4.1 that links in Wikipedia signal a semantic relation: if a link from A to B means that B is relevant (in some sense) to A, then it is also likely A is relevant (in some sense) to B.

4.2.4 *Relevant link distribution*

Is the degree distribution of relevant pages different from non-relevant pages? If there is a difference, we could possibly exploit this to separate relevant from non-relevant pages. For both collections we have available sets of search requests and associated sets of relevant pages. How are the link degrees of these relevant pages distributed?

Tasks	# topics	# rel. docs	# rel./topic
Home Page	75	83	1.17
Named Page	75	80	1.07
Topic Distillation	75	1,600	21.33
Mixed	225	1,763	7.84

Table 10: Statistics of the relevance judgements of the TREC 2004 Web Track tasks.

In the previous chapter we saw that the incoming link degree distribution over the relevant pages in Wikipedia follows a power law with a less steep slope than the distribution over all pages. Do the different natures of Wikipedia and the Web lead to different distributions of relevant pages?[4]

For .GOV, we use the TREC 2004 Web Track data consisting of 225 retrieval topics and in total 1,763 relevant pages. This is a mix of Home Page, Named Page and Topic Distillation topics. Statistics about the number of topics and relevant pages are shown in Table 10. For the Home Page and Named Page tasks, the target is a single entry page (although some entry pages have multiple URLS, giving rise to multiple relevant pages per topic). For the Topic Distillation topics, there are 21.33 relevant pages per topic on average.

The in-degree distribution of all and relevant pages for .GOV is shown in the top left of Figure 10. The relevant pages show a similar distribution as the total set of pages, but with a less steep slope. Low in-degree pages are relatively less frequent among the relevant pages. If we compare this with the in-degree distribution of the (relevant) pages of Wikipedia (top right in Figure 10), we can see almost no difference. Incoming link patterns are apparently very similar in Wikipedia and the Web, despite the many differences mentioned in Section 4.1.

We now turn to the out-degree distributions of (relevant) pages for the .GOV (bottom left of Figure 10). The out-degree distribution of the relevant pages in .GOV is almost the same as that of all pages. There is a difference between 5 and 100 outgoing links, which seems largest somewhere between 10 and 50. Above 100 outgoing links, the set of relevant pages is similarly distributed to the full collection. The set of

4 Of course, the different tasks target different types of pages. Web-centric search tasks often target entry pages, which are often not relevant within the ad hoc methodology. A difference in the degree distribution of relevant pages might be ascribed to the different natures of the search tasks. We would expect any such differences to be reflected by the impact of link evidence on retrieval.

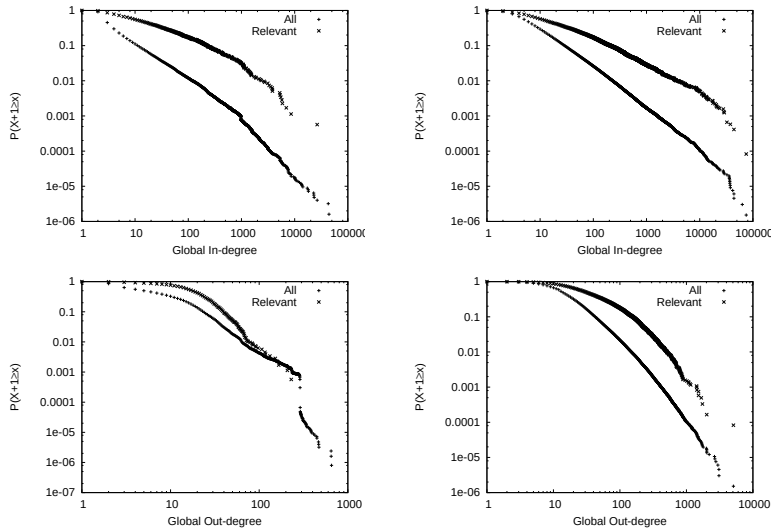

Figure 10: CCDF of all pages and relevant pages for the global incoming link degrees in .GOV (top left) and Wikipedia (top right) and for the global outgoing link degrees in .GOV (bottom left) and Wikipedia (bottom right).

relevant pages has relatively many pages with 5–100 outgoing links. For Wikipedia (bottom right of Figure 10), there is a marked difference between the distributions of relevant and all pages. Above 10 outgoing links, the distribution of the relevant pages falls less quickly than the distribution of all pages. The gap increases up to around 500 outgoing links, then seems to remain stable. Again, the out-degrees show similar behaviour to the in-degrees, which further supports our hypothesis that in Wikipedia there is little difference between incoming and outgoing links.

To sum up, on a global level the Web and Wikipedia link structures show a lot of similarities. The Wikipedia in-degrees seem very similar to in-degrees in the Web. Out-degrees in Wikipedia behave very similarly to in-degrees, and by transitivity should also behave similarly to in-degrees in the Web. For Web-centric search tasks, the out-degree distribution of relevant pages stays close to the out-degree distribution of non-relevant pages. Do these observations still hold when we zoom in on a small set of documents retrieved in response to a particular query?

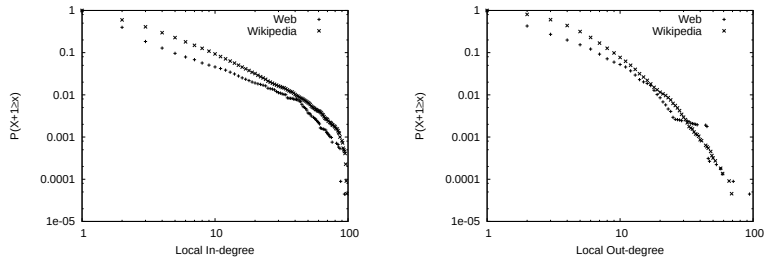

Figure 11: Cumulative distribution of the local link in-degrees (left) and out-degrees (right) for .GOV and Wikipedia.

4.2.5 *Local degree distributions*

On the left side of Figure 11 we compare the local in-degree distributions of the top 100 retrieved results for .GOV and Wikipedia collections. The Wikipedia distribution lies slightly higher than the .GOV distribution, meaning the local link graphs for the Wikipedia topics are more densely interlinked than the local graphs for the .GOV topics. This is to be expected with the higher global link density of the Wikipedia collection. Apart from that, the distributions are very similar, which is in line with the observations of the global in-degrees.

For the out-degree distributions (right side of Figure 11) there is a bigger difference between global and local degrees. On a global level, the out-degree distributions of the .GOV and Wikipedia collection showed bigger differences than the in-degree distributions, but in the local graphs, the out-degree distributions seem more similar. The Wikipedia distribution stays above the .GOV distribution up to around $x = 30$, which is where the .GOV distribution starts fluctuating and crosses the Wikipedia distribution.

Next, we compare the distribution over relevant and all pages for .GOV, in the top left of Figure 12. The gap between the two distributions grows with increasing in-degree. The high local in-degree pages are relatively more frequent among the relevant pages than among all pages. The same holds for the relevant pages in Wikipedia (top right of Figure 12). In the local sets the in-degrees behave in the same way as in the global sets, but the difference between relevant and all pages is not as pronounced as for the global in-degrees.

For the local out-degrees in .GOV, the bottom left of Figure 12 shows a difference between relevant and all pages that grows from a few outgoing links up to 35 or so. Above that, no difference is visible.

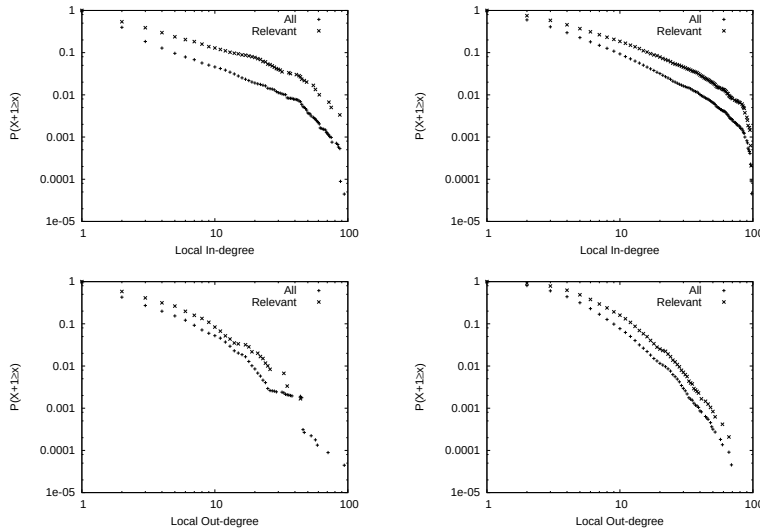

Figure 12: Cumulative distribution of all pages and relevant pages for the local incoming link degrees in .GOV (top left) and Wikipedia (top right) and for the local outgoing link degrees in .GOV (bottom left) and Wikipedia (bottom right).

This is again very similar to the global out-degree distributions. For Wikipedia (bottom right of Figure 12), the difference between relevant and all pages slowly grows from 1 to 10 outgoing links, then seems to stabilise.

In sum, from the degree distributions, there seems to be no big difference between the global and local link degrees. So far it seems the main difference between Web and Wikipedia hyperlinks is the distribution of the outgoing degrees. Wikipedia local in- and out-degrees show similar behaviour to each other and to their global counterparts. Does this also mean that in their relation with relevance, the incoming and outgoing link degrees are similar?

4.2.6 *Prior probability of relevance*

In this final part of the comparative analysis of the Web and Wikipedia link structure, we want to find out how the link degrees are related to the relevance of retrieval results. As in the previous chapter, we again analyse the prior probability of relevance of a page with a particular degree (see Section 3.2.3 on page 55 for details on how we derive the

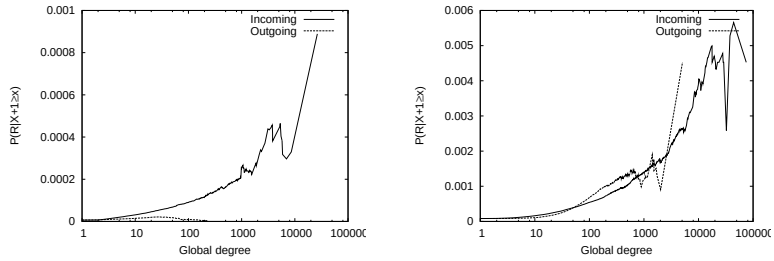

Figure 13: Global link degree prior probability of relevance for .GOV (left) and Wikipedia (right)

prior probabilities). This time we look at both incoming and outgoing link degrees for global and local graphs of the Wikipedia and .GOV collections.

On the left side of Figure 13 we see the probability of relevance over the global incoming and outgoing degrees for the .GOV collection. As the different shapes of the relevant and overall distributions suggested in Figure 10, the pages with high in-degree have a higher probability of being relevant than pages with low in-degree. Above 1,000 incoming links the data becomes sparse and the curve becomes erratic, but overall the probability of relevance seems to increase monotonically with in-degree. In other words, in the Web, global in-degree is an indicator of the type of pages the Web users are looking for. The prior probability of relevance of the outgoing degrees is at the bottom and only visibly in the middle range of the distribution. At a different scale—a log-log scale for instance—the full range of the curve would be visible. However, we use this scale to clearly show the difference between incoming and outgoing link evidence. Web pages with out-degrees in the middle range (peaking somewhere between 10 and 100) have a higher probability of being relevant than pages with lower or higher out-degrees.

In the Wikipedia collection (right side of Figure 13) both in- and out-degree seem to be good indicators of relevance: a higher degree corresponds to a higher probability of relevance. Recall from above that the fraction of reciprocal links in Wikipedia is actually lower than that of .GOV; it is not a result of pages linking back-and-forth. This, again, signals the difference in the link structure of Wikipedia and the Web at large. For the semantic links of Wikipedia, the difference between incoming and outgoing links seems to disappear and both can be used as indicators of relevance.

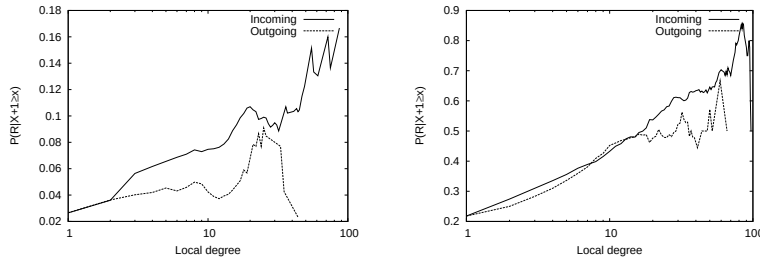

Figure 14: Local link degree prior probability of relevance for .GOV (left) and Wikipedia (right)

The prior probability of relevance for the local degrees is shown in Figure 14 for .GOV on the left side and for Wikipedia on the right side. The most striking difference between the local and global degrees in .GOV is that the local out-degrees show a stronger relation with relevance than the global out-degrees.

In this section we looked in detail at the .GOV and Wikipedia link structures to see if and in what way the Wikipedia link structure differs from general Web link structure. Wikipedia has a denser link structure, which might be a consequence of the shared authorship, encyclopedic organisation and the fact that Wikipedia is a single domain. This also results in greater connectedness, with over 90% of the pages belonging to the giant SCC.

The degree structures of Wikipedia and .GOV are fairly similar. Both collections have power law distributions of incoming and outgoing links, although in Wikipedia the outgoing link degrees are very similar to the incoming link degrees, which might again be attributed to the shared authorship aspect of Wikipedia. If all contributors to Wikipedia can control both incoming and outgoing links, we expect them to behave in similar ways.

The relevant pages have a slightly different distribution, adhering less to a power law. We also found that the global degree distributions are similar to the local degree distributions, which was also observed by Chakrabarti et al. (2002).

Finally, we looked at the relation between link degrees and relevance. Because the relevant pages tend to have a different degree distribution from the non-relevant pages, degrees are related to the relevance of retrieval results, with pages with higher degrees in general having a higher probability of being relevant. A major finding is that, because of the similarity between incoming and outgoing links in Wikipedia, both

in- and out-degrees are related to relevance. In the Web, where authors only have control of the outgoing links of their own pages, outgoing link degree are only somewhat related to relevance on a local level. Incoming link degrees show a stronger relation with relevance, both globally and locally, and are therefore more important.

From the previous chapter we know that the global in-degrees are not very effective for Wikipedia ad hoc retrieval, while the local in-degrees are. Does the similar behaviour of the in- and out-degrees mean that the same holds for the effectiveness of out-degrees? Does the similar behaviour of the Wikipedia and .GOV in-degrees mean that global degrees are also less effective than local degrees for Web-centric tasks? In the next section we compare the impact of link evidence on Web and Wikipedia retrieval.

4.3 EXPERIMENTS

We now turn to the second part of our set of research questions. What is the impact of link evidence on Web and Wikipedia retrieval? And how do more complex link-based ranking algorithms compare against simple link degree counts in terms of retrieval effectiveness? For the impact of link evidence on the Web, we focus on Web-centric retrieval tasks where link evidence is known to be effective (the impact on ad hoc search on the Web is discussed in Chapter 7). The .GOV collection was used for the TREC 2004 Web track data and comes with a mixed query set of 225 topics divided equally between Topic Distillation, Home Page and Named-Page topics. Here the known-item search topics tend to have a single relevant document (possibly more due to duplicates in the collection), and the distillation topics tend to have a larger set of key results.

As in the previous chapter (see Section 3.2.4, page 57), we first illustrate the impact of global and local link evidence by discussing in detail one of the TREC Web Track topics. We then describe the baseline systems used for the Web and Wikipedia experiments and the impact of link evidence on retrieval performance, including the HITS and PageRank algorithms.

4.3.1 *Naive Link-based Ranking*

Topic 119 of the TREC 2004 Web Track has as title *Groundhog day Punxsutawney* and is about a celebration day in Punxsutawney, where

people watch whether a groundhog leaves its burrow and sees its own shadow, to determine how long winter will last.

We use global degrees, i.e., the total number of incoming links, outgoing links, or combined in- and outgoing links for a page. To illustrate the effect of global degrees, we took the top 1,000 articles from the baseline run described below in Section 4.3.2, and list the 10 articles with the highest global indegree in Table 11 for .GOV. For comparison, the naive link-based ranking of Wikipedia topic 339 is shown on page 58.

What we see is that the content-only run has mostly pages in the top ranks with the word *Groundhog* in the title, and several of them are related to weather forecasts. The global degree mainly has pages that seem to have no bearing on the topic at all, but instead are navigational pages such as *Site Map*, *AMS Search* and *Metadata Records By Catalog Title*. As in Wikipedia, the global in-degree in .GOV leads to infiltration of important but off-topic pages.

We see that the local degrees keep slightly better focus on the topic of request. The fifth and eighth results are about weather forecasts and the latter is about Groundhog Day. We also observe that local links are sparser and just a few local links are all it takes to infiltrate the lower ranks.

Although the global in-degree showed a clear relation with relevance, the link-only ranking has mainly off-topic results in the top. Because the local links keep more focus on the topic, we expect that local link evidence is more effective for the .GOV topics as well.

4.3.2 *Baselines*

The baseline for Wikipedia is the same as in the previous chapter (see Section 3.1.4, page 48). For the Web collection, we use a mixture language model over three document representations: document text, incoming anchor texts and title field. This provides a much better baseline than using text alone (Kamps, 2005). For a collection D, document d and query q:

$$P(q|d) = \prod_{t \in q}((1 - \lambda_1 - \lambda_2 - \lambda_3) \cdot P(t|D) + \lambda_1 \cdot P_{\text{doc}}(t|d) + \lambda_2 \cdot P_{\text{anchor}}(t|d) + \lambda_3 \cdot P_{\text{title}}(t|d))$$

where each of the document language models is estimated as described in Section 3.1.4 on page 48.

For the mixture model run, all three models are weighted the same with $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$.

Title	Content
NCDC: Climate-Watch, Groundhog Day Special Report	1.89e-06
NOAA Puts Groundhog to the Test	2.76e-08
EOS Highlights Archive – NASA Satellite Saw More U.S. Snow In Early Winter; Groundhog May See More Coming (February 2 2001)	4.98e-09
FTWSWSTX/1	1.15e-09
Hewitt celebrates Groundhog Day with 'shadow' from Columbia High School	1.83e-10
Groundhog Job Shadow Day	1.53e-10
Groundhog Day in Souda Bay	1.47e-10
EO News: A Snowy Winter for Western U.S. - February 2, 2001	1.02e-10
Title	**Global in-degree**
Site Map	3,119
Online Library - HUD	2,119
Bureau of Labor Statistics Home Page	1,119
AMS - Search	730
The United States Mint	722
NHGRI: In The News	518
Metadata Records By Catalog Title	448
FCC Universal Licensing System	348
Title	**Local in-degree**
Bureau of Labor Statistics Home Page	61
NTP Meetings & Events	58
Recalls and other Press Releases	5
What's New	3
NCDC: Climate of 2001 - Climate Perspectives Reports	3
Youth Opportunity Movement Highlights	3
California Department of Motor Vehicles home page	2
Hewitt celebrates Groundhog Day with 'shadow' from Columbia High School	2

Table 11: Titles with the highest in-degrees in the .GOV collection for TREC topic 119, '*Groundhog day Punxsutawney*'

	.GOV			Wikipedia		
Variables	In	Out	Length	In	Out	Length
In	–	0.10	-0.01	–	0.19	0.16
Out		–	-0.07		–	0.65
Length			–			–

Table 12: Correlation between length and degrees for Web and Wikipedia collections.

The way the link priors are incorporated into the retrieval model is described in Section 3.3.1 (page 60). The degree score for a page may be based on either *local* or *global*, and either *in-degree* or *out-degree* (leading to four logical cases).

4.3.3 *Length prior*

Document length is related to relevance for ad hoc retrieval (Singhal et al., 1996), but not for Web tasks (Kamps, 2005). For Wikipedia, the baseline system uses a length prior to promote longer documents. For Web retrieval, a length prior might be detrimental. We first need to sort out the impact of document length on the relevance of retrieval results. Moreover, document length might actually be correlated to the link degree. Intuitively, we would expect that a document with many links going out is longer than a document with very few links going out. How is the length of documents related to the link degree? Moreover, we have seen above that the in- and out-degrees in Wikipedia show similar behaviour: how do these correlate?

Table 12 gives the correlation between the in-degrees, the out-degrees and the document length for both collections. For .GOV, we see a low correlation between in-degree and out-degree and no correlation between length and the degrees. For Wikipedia, we see a low correlation between in-degree and out-degree and between length and in-degree. However, there is a strong correlation between out-degree and document length in the Wikipedia collection. This makes sense, since pages containing more textual content will naturally give rise to more links inside Wikipedia.

We choose our baseline runs based on experiments with the document length priors. The best run for the .GOV collection uses no length prior (Table 13)—MAP drops from 0.3970 to 0.3419 when using the length prior, MRR drops from 0.4662 to 0.3868. The best run for the

	.GOV		Wikipedia	
Collection	MAP	MRR	MAP	MRR
Standard	**0.3970**	**0.4662**	0.2561	0.6969
Length prior	0.3419	0.3868	**0.3157**	**0.8119**

Table 13: Impact of length prior on Web and Wikipedia retrieval. Best scores are in bold.

Wikipedia collection uses a document length prior—MAP goes up from 0.2561 to 0.3157, MRR goes up from 0.6969 to 0.8119. We choose these best runs as baseline runs for the experiments with the link priors. That is, without length prior for .GOV and with length prior for Wikipedia. Note that the higher MAP for the Web data can be attributed to differences in the tasks, where there is a large fraction of known-item search topics for the Web data. The Mean Average Precision score is an average over the total number of relevant documents. With very few relevant documents per topic, it is easier to obtain a high MAP than when there are many relevant documents.

First we will discuss the experiments with the link evidence on the Web collection, then on the Wikipedia collection.

4.3.4 *Web*

We measure performance in terms of Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) scores. For topics with multiple relevant documents, the MAP gives a better idea of the overall performance, while MRR gives an indication of early precision. For topics with only one relevant document, the MAP score is the same as the MRR score: with only one relevant document, the average precision is the same as the precision of the first relevant document. In our discussion we will mainly focus on MAP.

Recall from the previous chapter that in Wikipedia, the global in-degree hurt performance unless we used the log of the degree, and even then give little improvement. The local in-degree was much more effective and became less effective when we used the log of the degree. We will compare the Web and Wikipedia results in more detailed in the next subsection. The top half of table 14 shows the results for the link prior runs on the Web track collection. We tested all the runs for significance of the increase or decrease in performance over the baseline

	MAP		MRR	
Run id	Global	Local	Global	Local
Baseline	0.3970		0.4662	
All In-degree	**0.4701**$^{\bullet}$	0.4544$^{\circledcirc}$	**0.5843**$^{\bullet}$	0.5462$^{\bullet}$
All Out-degree	0.4261°	0.3978	0.5031°	0.4819
All Log In-degree	0.4449$^{\bullet}$	0.4410$^{\bullet}$	0.5209$^{\bullet}$	0.5148$^{\bullet}$
All Log Out-degree	0.4082$^{\circledcirc}$	0.4181$^{\bullet}$	0.4789°	0.4879$^{\circledcirc}$
External In-degree	0.3278$^{\circledcirc}$	0.2178$^{\bullet}$	0.4603	0.3142$^{\bullet}$
External Out-degree	0.3176$^{\bullet}$	0.0820$^{\bullet}$	0.3863$^{\bullet}$	0.1641$^{\bullet}$
External Log In-degree	0.4361$^{\bullet}$	0.4127$^{\bullet}$	0.5130$^{\bullet}$	0.4823$^{\bullet}$
External Log Out-degree	0.3952	0.3988	0.4621	0.4669

Table 14: Results of the link degree priors on the 225 topics of the .GOV collection

using the bootstrap test, one-tailed, using 100,000 resamples. We report three significance levels, $p<.05$ ($^{\circ}$), $p<0.01$ ($^{\circledcirc}$) and $p<0.001$ ($^{\bullet}$).

Using the normal degree priors, global degrees are more effective than local degrees. Apparently, Web search tasks do not require the more topical focus of the local link degrees. The normal priors are more effective than the log priors, except for the local out-degrees. Global, query-independent link evidence is so important that no toning down is needed. Note that the global degrees are applied on all retrieved results. There is no need to limit the impact of global link evidence on the top 100 results as in the previous chapter.

With log degree priors, the local out-degrees are more effective than the global out-degrees. All global link evidence leads to significant improvements upon the content-only baseline. Overall, the global in-degrees are the most effective.

The local degrees also lead to significant improvements—apart from the non-log out-degree priors—but in general at lower significance levels.

These results are in line with the analysis of the prior probability of relevance over degrees in Section 4.2.6. For Web-centric search tasks, global link evidence is the most effective, and in-degrees are more effective than out-degrees. In other words, link evidence is most effective for finding entry pages when derived independent of the query. Entry pages must have more incoming links than other pages of the same Web site.

4.3.4.1 *Site-external links*

We can try to distinguish between site-internal Web links (for example, navigational links within a site) and site-external Web links (for example, a link to related content on a different site). Site-internal links are often considered to be less useful (Kleinberg, 1999) because they serve purely navigational purposes. We first identified the site of a page as its base URL, with the removal of any prefix starting with `www` and excluded links between pages within the same domain. We further reduced the set by removing links between base URLs when either is a substring of the other. For example, a link between `www.nih.gov` and `www.nlm.nih.gov` is regarded as site-internal, while a link between `www.nlm.nih.gov` and `www.nichd.nih.gov` is regarded as site-external. The resulting set of site-external links contains 1,693,477 links (or 15% of all links). For the Web collection, we use either the full link graph or the site-external links.

If we consider only site-external links, the in-degrees are more effective when used as log priors, but the improvements are much smaller than when using all links. Site-internal links are thus also important for locating entry pages. The site-external out-degrees are ineffective and the global in-degrees are more effective than the local in-degrees.

All normal degree priors have a negative impact. The local site-external degrees even lead to disastrous drops in performance. They seem to promote the wrong documents. This is in stark contrast to the highly beneficial impact of the local degrees over all links. Not only that, it is also in complete contrast with the claim that the query-dependent HITS algorithm benefits from ignoring site-internal links (Kleinberg, 1999). But the main difference between the set of pages on which HITS is used and this set of top ranked results is that, for HITS, the set of top ranked results is expanded with pages connected to those results. We will address this issue in Section 4.3.7.

4.3.4.2 *Sub-tasks*

We take a closer look at the difference between global and local, incoming and outgoing and site-internal and site-external links by zooming in on the three sub-tasks of the TREC 2004 Web Track: Home Page finding, Named Page finding and Topic Distillation. For Home Page finding, links are highly effective. Global degrees are more effective than local degrees and incoming links are more effective than outgoing links. With the external links, the outgoing links are only effective when used as log priors.

	Home Page		Named Page		Topic Dist.	
	MRR		MRR		MAP	
Run id	Global	Local	Global	Local	Global	Local
Baseline	0.4438		0.6595		0.0973	
All In-degree	**0.6575**$^{\bullet}$	0.6239$^{\bullet}$	0.6399	0.6596	**0.1320**$^{\bullet}$	0.0997
All Out-degree	0.5272$^{\circledcirc}$	0.5264°	0.6466	0.5749°	0.1137$^{\circledcirc}$	0.1100
All Log In-degree	0.5524$^{\bullet}$	0.5466$^{\bullet}$	0.6758	0.6805	0.1143$^{\bullet}$	0.1044$^{\circledcirc}$
All Log Out-degree	0.4733$^{\bullet}$	0.4738°	0.6543	**0.6847**$^{\circledcirc}$	0.1071$^{\bullet}$	0.1053$^{\bullet}$
External In-degree	0.5876$^{\circledcirc}$	0.4032	0.3182$^{\bullet}$	0.1847$^{\bullet}$	0.1029	0.0774°
External Out-degree	0.3286$^{\circledcirc}$	0.0726$^{\bullet}$	0.5515$^{\circledcirc}$	0.0915$^{\bullet}$	0.0892°	0.0843
External Log In-degree	0.5341$^{\bullet}$	0.4823$^{\bullet}$	0.6737$^{\circledcirc}$	0.6645°	0.1112$^{\bullet}$	0.1010°
External Log Out-degree	0.4519	0.4385	0.6429°	0.6688°	0.0997°	0.1002$^{\circledcirc}$

Table 15: Results for the degree priors over the different tasks.

For Named Page finding, links add very little to the content-only baseline. But the baseline is much higher than for home page finding, indicating that content-based retrieval is better for named paged finding. It is harder to improve upon this baseline. The log priors are more effective than the normal priors and the local degrees are generally more effective than the global degrees, except for the normal external link degree priors. The effectiveness of the log priors suggests that link evidence is somewhat noisy and needs to be curbed. It is not clear whether this is because of infiltration of important but off-topic pages or because the link priors push up pages that are topically relevant but not the desired ones. A possible explanation could be that pages with high in-degree are often home pages, which are considered irrelevant for the Named Page finding topics. The global in-degrees have a large impact on the ranking, but hardly affect the MRR. This suggests that global in-degree is unrelated to the relevance of named pages. Pages targeted by Named Page topics are equally spread over global in-degree distribution.

The scores for the Topic Distillation topics (columns 6 and 7) are much lower. Note that here we show the MAP scores, because there are multiple relevant pages per topic. However, the low scores indicate that the relevant key resources are not ranked highly. The link priors can improve performance. Global degrees are more effective than local degrees. In-degrees are more effective than out-degrees for the global priors, and vice versa for the local priors.

	MAP		MRR	
Run id	Global	Local	Global	Local
Baseline	0.3157		0.8119	
In-degree	0.2871$^{\bullet}$	0.3272$^{\bullet}$	0.7899	0.8249
Out-degree	0.2943$^{\bullet}$	**0.3276**$^{\bullet}$	0.7691$^{\bullet}$	0.8265
Log In-degree	0.3192°	0.3243$^{\bullet}$	**0.8341**$^{\bullet}$	0.8288°
Log Out-degree	0.3169	0.3218$^{\bullet}$	0.8311$^{\bullet}$	0.8254°

Table 16: Results of the link degree priors over the top 100 results for the INEX Wikipedia collection.

The global in-degrees are very effective for Home Page finding and Topic Distillation, and have no impact—thus no negative impact—on the Named Page finding topics. In general then, for Web-centric search tasks, global in-degrees are highly effective, as has been established before (Craswell et al., 2001, Kamps, 2005, Kraaij et al., 2002). Outgoing link degrees are also effective but less so. Although they are probably less related to popularity, they do seem to identify entry pages. Entry pages tend to have outgoing links to various parts of the site. Local link evidence can be effective as well, but its positive impact is usually smaller than that of global link evidence, suggesting that its relation to relevance comes from its correlation to global link evidence—a page can only have a high local degree if it has a high global degree—and not from its relation to the topical context of the query.

One would expect that introducing topical focus to link evidence would make it even more effective. After all, the requested pages might not have relevant text but still be topically related to the query. The main difference between Web-centric search tasks and ad hoc retrieval seems to be that there is no need to make link evidence sensitive to topical context of typical Web search queries. Instead, it is used to distinguish entry pages and other important and popular Web pages from the large bulk of low-quality pages and pages deeper in the hierarchy of sites.

4.3.5 *Wikipedia*

The results for Wikipedia are shown in Table 16. The results of the in-degree are repeated from Chapter 3 for ease of comparison. The global in- and out-degrees hurt performance when used directly as priors, with in-degree worse for MAP and out-degree worse for MRR.

If we curb their impact by taking the log of the degree, the global out-degree has an insignificant impact on MAP but a significant positive impact on MRR. The log global in-degree significantly improves both MAP and MRR. Global in- and out-degree are different from each other. Locally, however, the difference between incoming and outgoing link evidence has almost completely disappeared. As the probability of relevance analysis already suggested (see Section 4.2.6 on page 82), in Wikipedia the difference between incoming and outgoing links is small, and the local outgoing links are just as informative as the local incoming links. The direction of the links does not affect their impact. Their contribution is symmetrical, thereby conflating the notions of authorities and hubs.

If we compare this to the impact of links on the Web, there are two important differences: the *nature of the set of documents* and *the direction of the links* from which the link degrees are derived. This relates to our hypothesis that in Wikipedia there is little difference between incoming and outgoing links. With shared authorship, there is no difference between incoming and outgoing links in Wikipedia, while in the Web, the incoming links of a page are often authored by someone other than the page author herself and are less biased than the outgoing links, which are all authored by the page author.

For Web-centric search tasks, global degrees are more effective than local degrees, showing that for entry page finding, the importance of link evidence is not related to the query, but to the type of pages that tend to be desirable results for Web searchers. Evidence is best derived from the entire collection. For ad hoc search in Wikipedia, this is different. First, there are no entry pages within the set of encyclopedic articles. Each page stands on its own.[5] Second, the search task requires a different set of documents that cannot be identified properly by looking at query-independent link evidence. For ad hoc retrieval, link evidence is best derived from a set of document related to the query.

The other important difference of link impact is the link direction. For typical Web search, incoming links are more useful or informative than outgoing links. The link direction matters for finding entry pages, while for finding articles on a certain topic, the link direction is less valuable. Incoming and outgoing links are equally informative.

One structural difference between Web and Wikipedia links that might make links in Wikipedia more useful for ad hoc retrieval is

5 Although the guidelines state that long Wikipedia articles should be split up into a main article and several articles discussing sub-topics or aspects of the main topic, where the main topic article could be interpreted as an entry page for that topic. As the encyclopedic entries grow over time, more and more of them are split up.

that Wikipedia pages are meant to stand on their own. Whereas in some Web sites individual pages might be hard to understand or even be meaningless without the context of their surrounding pages, each Wikipedia article is meant to be interpretable independent of others.

4.3.6 *Beyond degrees: HITS and PageRank*

We briefly move away from the degree analysis and look instead at two propagation-based algorithms, HITS and PageRank. As discussed in Section 2.3.6, there are many link-based ranking algorithms that are more complex and try to derive information about quality and authority of pages. Most of these algorithms use the links to propagate scores through the network of pages and are iterative in nature to let the scores converge to a stable distribution. Here, connectedness plays a role. Through propagation, pages affect the scores of other pages that are part of the same graph component. The more connected the graph, the larger the components and the more pages are affected by each other.

PageRank and HITS are obvious choices to further examine the difference between global and local evidence. They are probably the two most well-known link-based ranking algorithms, are typically used on a global and local level respectively, and both algorithms use only the link topology for ranking. Descriptions of the PageRank and HITS algorithms can be found in Section 2.3.6.2.

For PageRank, we set the damping factor to the default value of 0.85 and use the resulting score as a document prior probability similar to the link degrees. Because all pages in the collection necessarily have a positive PageRank score, we use the score directly as a prior:

$$\begin{aligned} P_{\mathrm{PR}}(d) &\propto \mathrm{PR}(d) \\ P_{\log \mathrm{PR}}(d) &\propto \log\,(1 + \mathrm{PR}(d)) \end{aligned}$$

where $\mathrm{PR}(d)$ is the PageRank of document d. For the log prior we use $1 + \mathrm{PR}(d)$ to ensure the prior is always positive. PageRank scores below 1 would otherwise turn into negative scores.

We computed HITS authority and hub scores, both with and without expanding the initial root set of top retrieved results. Instead of iterating until the scores converge, we stopped after 5 iterations. For the topological ranking this hardly has any impact, but for the resulting score distribution the impact is much bigger. The more iterations we use, the larger the gap between low and high scores. The shape of

Rund id	MAP	Δ	MRR	Δ
Baseline	0.3970	0.00%	0.4662	0.00%
PR All	0.4861$^{\bullet}$	22.44%	**0.5988**$^{\bullet}$	**28.44**%
PR External	0.4252	7.10%	0.5360$^{\circledcirc}$	14.97%
Log PR All	**0.5021**$^{\bullet}$	**26.47**%	0.5877$^{\bullet}$	26.06%
Log PR External	0.4582$^{\bullet}$	15.42%	0.5588$^{\bullet}$	19.86%
HITS All 100 Auth.	0.3900	-1.76%	0.4711	1.05%
HITS All 100 Hub	0.3467°	-12.67%	0.4190°	-10.12%
HITS External 100 Auth.	**0.4090**	**3.02**%	**0.4978**$^{\circ}$	**6.78**%
HITS External 100 Hub	0.3498$^{\bullet}$	-11.89%	0.4267°	-8.47%

Table 17: Results of combining content-based and HITS scores on the .GOV topics.

the distribution is important when combining the HITS scores with the content-based scores. To combine the authority and hub scores with the content-based scores, we use the score directly as a prior:[6]

$$P_{\text{Auth}}(d) \propto \text{Auth}(d)$$
$$P_{\text{Hub}}(d) \propto \text{Hub}(d)$$

We will first discuss the experiments using only the top 100 results. Expansion of this set is discussed in Section 4.3.7. The results of using PageRank and HITS for re-ranking results on the .GOV collection are shown in Table 17. PageRank is far more effective than HITS. The normal PageRank scores are the most effective for MRR while the log PageRank scores are more effective for MAP. The improvements are comparable to those of the global degrees, although the PageRank scores are slightly more effective in general and especially when we use the log of the scores.

For HITS, the site-external links are more effective than using the full link graph, showing that for Web-centric search, authority is indeed better measured by ignoring links between pages written by the same author(s). The authority scores are more effective than the hub scores, much as the local in-degrees are more effective than the local out-degrees. However, the local degrees are more effective than the HITS

6 We experimented with both $1 + \text{HITS}(d)$ and $\text{HITS}(d)$ and found the latter to give the best overall performance.

Rund id	MAP	Δ	MRR	Δ
Baseline	**0.3157**		0.8119	
PR	0.2320$^{\bullet}$	-26.51%	0.7790°	-4.05%
PR top 100	0.2845$^{\bullet}$	-9.88%	0.7940	-2.20%
Log PR 100	0.3096	-1.93%	**0.8351**$^{\circ}$	*2.86%*
Log PR 10000	0.2855$^{\bullet}$	-9.57%	**0.8351**$^{\circ}$	*2.86%*
HITS 100 Auth.	0.3131	-0.82%	0.8015	-1.28%
HITS 100 Hub	0.3140	-0.54%	0.7910	-2.57%

Table 18: Results of combining content-based and HITS scores on the Wikipedia topics.

scores. This might be a result of the number of iterations used for the HITS algorithm. If we start with computing authorities and update the hubs afterwards, the first iteration authority scores are similar to the local in-degrees. After each subsequent iteration, the high authority scores remain relatively stable while the low authority scores rapidly drop to zero. This stretches the distribution more and more, and results in the authority scores dominating the content-based score such that it has almost no impact on the final ranking.

The impact of HITS and PageRank on the Wikipedia topics can be seen in Table 18. Similar to the global degrees, the query-independent PageRank scores hurt performance, especially when used over all retrieved results. If we use the log of the scores only the MRR improves, but overall precision drops, where the global degrees could slightly improve MAP. The HITS scores hurt performance on Wikipedia, which fits with our conjecture in Section 4.1 that the links in Wikipedia do not confer authority.

In Table 19 we see the impact of PageRank on the different topic types of the .GOV collection. The Home Page (HP) and Topic Distillation (TD) topics benefit most from the normal PageRank scores. The log priors are also effective but to a lesser extend. The reverse is true for the Named Page (NP) topics. The normal PageRank scores hardly affect the baseline performance, but the log version significantly improves it. Relevant named pages are not the most important pages overall—otherwise the normal PageRank scores would be more effective—but page importance does play a role. Named pages are more important than the bulk of the pages on the Web.

The site-external links alone also lead to improvements, but PageRank is more effective when using all links. The site-external link graph is

Run	HP MRR	NP MRR	TD MAP	Mix MAP
Baseline	0.4438	0.6595	0.0973	0.3970
PR All	**0.6616**$^{\bullet}$	0.6789	**0.1411**$^{\bullet}$	0.4861$^{\bullet}$
PR Ext.	0.5807$^{\bullet}$	0.5846$^{\circledcirc}$	0.1243$^{\bullet}$	0.4252
log PR All	0.6590$^{\bullet}$	**0.7281**$^{\bullet}$	0.1334$^{\bullet}$	**0.5021**$^{\bullet}$
log PR Ext.	0.5972$^{\bullet}$	0.6598	0.1260$^{\bullet}$	0.4582$^{\bullet}$

Table 19: Results for PageRank on .GOV using Home Page (HP), Named Page (NP), Topic Distillation (TD) and the mixed (Mix) topics.

much sparser, and far fewer pages have incoming site-external links. With fewer links, the graph is less connected and pages have less impact on each other. In a Strongly Connected Component, all pages affect each other's PageRank. When there are many small islands in the link graph, only pages on the same island affect each other's PageRank and the resulting PageRank score is thus less "global". Arguably, higher connectedness better reflects the flow of popularity and importance.

4.3.7 *Expanding the HITS root set*

We experimented with using only the top 100 or 200 results and with expanding the top results with pages connected to those top results. The original algorithm expands the root set R of the top 200 results with all pages linked from R and up to 50 pages linking to a page in R. This incoming link limitation is set because some pages have tens of thousands or even millions of incoming links and would come to dominate the local link graph. Because some Wikipedia pages have thousands of outgoing links, a similar problem could occur if we set no limit on the number of outgoing links used to expand the root set. Therefore, we use the limit d to determine the maximum number of incoming and outgoing links per page used for expansion. For $d = 50$, up to 50 incoming links and 50 outgoing links of a page p are used for expansion. We use the same limitation parameter d for .GOV and Wikipedia.

In Figure 15 we see the impact of the expansion parameter d on the MAP of authority and hub rankings of the full link graph of .GOV (top), the site-external link graph (middle) and Wikipedia (bottom).

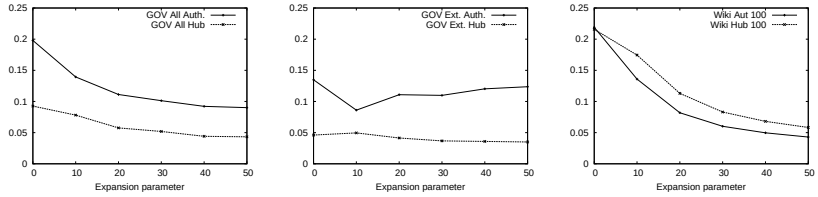

Figure 15: The impact of expanding the root set on MAP for HITS authority and hub scores.

On .GOV, the authorities perform much better than the hubs, both when using all links and only site-external links. Using the full link graph, expanding the root set only hurts performance. Expanding the root set using all links introduces many pages from the same site, which might be densely interlinked. When using only site-external links, the best performance is observed when d is set to zero, but from $d = 10$ onwards, authorities perform better as d increases. If any pages from other sites are added, a larger number of them better captures the authority relations between sites. The full link graph without expansion is much more effective for HITS than the site-external link graph with expansion. This means that site-internal links are useful to find entry pages. They are also useful to confer authority to the most authoritative pages within that site.

In Wikipedia, expansion hurts performance: it introduces many irrelevant pages. Especially highly connected pages are easily introduced into the final set of pages, and lead to heavy infiltration. Without expansion, authority and hub scores are equally useful for ranking. With expansion, hub scores are more effective than authority scores.

In sum, there seems to be little benefit in expanding the root set of top retrieved results. Although expansion can potentially bring in missing entry pages, and help identify the most authoritative pages, it also brings in many irrelevant pages and cause loss of topical focus.

4.4 CONCLUSIONS

In this chapter, we investigated the difference between Wikipedia and Web link structure. We first performed a comparative analysis of Wikipedia and .GOV link structure and then investigated the value of link evidence for improving search on Wikipedia and on .GOV.

In our comparative analysis of Wikipedia and Web link structure we hoped to find out:

- Are there differences between the Wikipedia and .GOV collection in terms of link density and connectedness?

The .GOV collection has a giant strong connected component larger than earlier general Web crawls, but the giant strong connected component of Wikipedia covers an even larger part of the collection. Wikipedia is more densely linked, but both link structures are in the final phase of connectedness. In terms of navigation, and the accessibility and visibility of pages in the two collections, both link graphs are relatively complete.

- What is the degree distribution of Wikipedia and the Web at large? Are there differences between distributions of incoming and outgoing links?

Although there are some striking differences between the Wikipedia and .GOV link structures, in many ways, they are very similar. Both have power law degree distributions where the out-degree distribution falls steeper than the in-degree distribution. Considering link direction, the small difference between Wikipedia and .GOV in-degrees suggest that incoming linking patterns are guided by the same principles. The most striking difference between the Wikipedia and .GOV link structures are the outgoing link degrees. In Wikipedia, the out-degree distribution is more similar to the in-degree distribution than in the .GOV collection. As we mentioned in Section 4.1, this might be the effect of the shared authorship and encyclopedic organisation of Wikipedia. Each topic has a dedicated Wikipedia page, so it is mostly clear where to link to, and the contributors can modify both the incoming and outgoing links of a page, which conflates the notions of hub and authority and essentially gives all pages equal authority.

- How does the link topology relate to the relevance of retrieval results?

For the .GOV collection, the global in-degree is a good indicator of relevance. Pages with many incoming links have a higher probability of being relevant than pages with few incoming links. The prior probability of relevance of the number of outgoing links first increases but then drops again, making the out-degree a less reliable indicator of relevance than the in-degree. For the Wikipedia collection, both in-degree and out-degree are good indicators of relevance. More generally, we observe that Wikipedia incoming and outgoing links are similar in character, again suggesting that in Wikipedia the notions of authority and hub are conflated.

In our retrieval experiments, we hoped to find an answer to the following question:

- What is the impact of link evidence on Web and Wikipedia retrieval?

The main difference between Web-centric search tasks and Wikipedia ad hoc retrieval seems to be that there is no need to make link evidence sensitive to the topical context of typical Web search queries. Instead, it is used to distinguish entry pages and other important and popular Web pages from the large bulk of low-quality pages and pages deeper in the hierarchy of sites. It is useful to identify the type of pages that tend to be desirable results for Web searchers. Local link evidence keeps more focus on the topic, but promoting any relevant page is not necessarily helpful, because the task assumes the user is only interested in the entry-page of a topically relevant site. The global link graph is much richer and better reflects the importance of documents.

For Wikipedia ad hoc retrieval, global link evidence is nowhere near as effective as local link evidence. Query-independent aspects of link evidence are less useful for ad hoc search. Global incoming link evidence seems slightly more useful than global outgoing link evidence. In Wikipedia, local link evidence is effective regardless of the direction of the links; incoming and outgoing link evidence is equally effective. As mentioned before, this makes sense if links in Wikipedia signal a topical relation between linked documents. Topical relatedness is a symmetrical relation.

The greater effectiveness of local link evidence implies that document importance cannot be the sole explanation for the effectiveness of link evidence for Wikipedia ad hoc retrieval. Local link evidence must be related to topical relevance as well. In the next chapter we will investigate how link evidence is related to both document importance and topical relevance.

Part iii

Links And Topical Relevance

5

FROM DOCUMENT IMPORTANCE TO TOPICAL RELEVANCE

Global link evidence is by nature query-independent, and is therefore no direct indicator of the topical relevance of a document for a given search request. As a result, link information is usually considered to be useful to identify the query-independent aspects of relevance which we refer to as aspects of the importance of documents. Incoming link evidence can be used as an indicator of authority or popularity. Outgoing link evidence can be used as an indicator of document length or 'hubness'. The direction of the link is assumed to determine the specific nature of the indicator. Our first conjecture is that *global link evidence is not related to topical relevance but to document importance*.

Local link evidence, in contrast, is query-dependent—based on our definition on page 48—and could in principle be related to topical relevance. Links are assumed to be a signal that linked documents are topically related to each other. The textual evidence for the relevance of a page is assumed to provide evidence for the relevance of its neighbours as well. Our second conjecture is that *local link evidence is related to topical relevance*.

Local link evidence is still derived from the global link structure, and the local link degree is bounded by the global degree. A page cannot have a higher local degree than its global degree. At the same time, pages with many links have a higher a priori probability of having links in the local set. Therefore, the local degree partly depends on the global degree and might also reflect the importance of documents.

This leads to our main research question:

- To what extent is link evidence related to the importance of documents, and to the topical relevance of documents?

We are mainly interested in the relation between link evidence and topical relevance. The relation between link evidence and document importance has been extensively studied and successfully exploited for Web-centric search tasks (see Section 2.3.1.1). The relation between link evidence and topical relevance has, to our knowledge, not been shown before and is still poorly understood.

Because local degrees depend on the global link structure, we first look at this dependence and address the questions:

- To what extent are local degrees dependent on global degrees?
- Can we make local link evidence less dependent on the global degree structure?

We can take both the global and local degrees into account and look at the fraction of all links that are present in the local set. We are mainly interested in the relation between link evidence and topical relevance, but to understand this relation, and because local degrees are dependent on the global degrees, we need to consider the relation between link evidence and document importance as well.

Although we cannot test our conjectures directly—we have no direct way of measuring topical relevance and document importance—they have several implications that should be observable.

Insofar as local link evidence is related to topical relevance, we would expect this evidence to be independent of the link direction and therefore symmetric. In the previous chapter we saw that local incoming and outgoing link degrees have a similar impact on overall retrieval performance. We argued that this makes sense if local link evidence is related to topical relevance. If a link between documents A and B is an indicator that A and B are topically related to each other, then the textual evidence of A for a given query also provides evidence for the relevance of B. But because topical relatedness is a symmetrical relation. The textual evidence for the relevance of B also provides evidence for the relevance of A.

We can test several aspects of link evidence for symmetry. First, there is the degree structure. Incoming and outgoing link degrees are derived from the same link graph. Perhaps pages with high local in-degree also have a high local out-degree and vice versa? If the degree structure is not symmetric, in- and out-degrees promote different documents. This could mean that if we use the evidence in both directions, essentially treating links as undirected links, we have more evidence to distinguish between pages.

Our second set of research questions addresses the relation between directed and undirected link evidence:

- Is local link evidence symmetric in its degree structure?
- Is global link evidence asymmetric in its degree structure?
- Does the symmetry of the degree structure increase as we make link evidence more sensitive to the search topic?

Another aspect of link evidence we can test for symmetry is the ability to distinguish relevant from non-relevant documents. So far we have only used link evidence in combination with text evidence. But we can also evaluate the document rankings of link evidence in isolation. We can compare incoming and outgoing link evidence in their ability to distinguish relevant from non-relevant documents. The undirected link evidence could also possibly further improve the degree-based ranking as it takes all neighbours into account. This idea was used by Carrière and Kazman (1997), who ranked Web search results based on their "connectivity". If local link evidence is related to topical relevance and global link evidence is not, we would also expect that local link evidence is more effective in isolation than global link evidence.

Our third set of questions is:

- Is the ranking based on local link evidence in isolation better than the ranking based on global link evidence in isolation?
- Is local link evidence symmetric in its ability to distinguish relevant from non-relevant documents?
- Is global link evidence asymmetric in its ability to distinguish relevant from non-relevant documents?

Finally, we will consider the possibility that some documents are more relevant than others. A good retrieval system ranks documents according to how relevant they are. The relevance judgements of INEX Wikipedia test collections contain detailed information on which parts of the text of relevant documents are relevant. Relevance assessors have highlighted those parts of the text that are relevant. This allows us to study the relation between link evidence, and the amount and fraction of relevant text in documents. We want to know:

- How are global and local link evidence related to the amount of relevant text in articles?
- How are global and local link evidence related to the fraction of relevant text in articles?

In the rest of this chapter, which consists of four parts, we will discuss how our three hypotheses hold up against our findings. In the first part (Section 5.1) we discuss how local link evidence can be made more independent from the global link structure and in the second part (Section 5.2) we analyse how different degree structures are related to each other. In the third part (Section 5.3) we compare the different link

directions and levels of link evidence in a retrieval setting and in part four (Section 5.4) we study the relation between the degrees and the amount of relevant text in articles. We draw conclusions in Section 5.5.

5.1 FROM QUERY-INDEPENDENCE TO QUERY-DEPENDENCE

The global degrees affect the local degrees in the sense that they determine the upper bound for the local degrees.[1] A page with n incoming links in the entire collection can have a maximum local degree of n for local sets of at least $n+1$ pages. The local set is query-dependent so the resulting link evidence is more focused on the query. However, the global degrees still play a role for the above mentioned reason. The local degree to some extent expresses the "local importance" of a page. Can we make link evidence more independent from the query-independent link structure? If two pages A and B have the same local link degree but different global link degrees, are they equally related to the local context or is one more related than the other? If page A has a much higher global degree than page B, we *expect* A to have a higher local degree as well, purely based on the a priori probability of having links in the local set. Intuitively, the page with the lower global link degree yet equal local link degree has stronger evidence of being related to the search topic. We want to make the a priori probability of local link evidence more uniform. A simple solution would be to normalise the local link degree by weighting it down with the global link degree. This should reduce the query-independent component in the query-dependent degrees.

Pages with a very high global degree but a low local degree 'lose' many of their links in the local graph, which could be interpreted as a signal that the link evidence of this page is related to document importance but not to topical relevance. In other words, the few links present in the local set are less meaningful. On the other hand, pages that have a low global degree but lose no links in the local set are supported by all their neighbours. It is highly unlikely that a page with low global degree has most of its links present in the local set by accident. Such a page "belongs to" this part of the link graph. If we weight link evidence by the ratio of local to global degrees, that is, the fraction of global links present in the local set, we make the evidence more sensitive to the topic and less sensitive to document importance.

1 Of course, the size of the local set also provides an upper bound. If the global degree of a page is higher than the number of pages in the local set, the upper bound is the determined by the local set size. Otherwise, the global degree is the upper bound.

This can be interpreted as the *local specificity* or *topical specificity* of the link evidence. Formally, the *local fraction* is calculated as:

$$\text{fraction}_{\text{loc}}(d) = \begin{cases} \deg_{\text{local}}(d) / \deg_{\text{global}}(d) & \text{if } \deg_{\text{global}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

This is radically different from local and global degrees by themselves. Although query-dependent, the local degree ranking is based on the amount of local evidence a page has. The page with the most links is ranked highest. The local fraction is based on the fraction of global links present in the local set. A page with a global in-degree of 1 and a local in-degree of 1 gets the maximum score. Pages with more local links cannot be ranked higher than this document.

The local degree and local fraction thus to reflect two aspects of topical relevance. The degree corresponds to the degree of relevance (from, say, marginally to highly relevant) and the fraction corresponds to how specifically about a topic a document is (from a small fragment to the whole article). This is related to the notions of exhaustivity and specificity used in early INEX evaluations (Pehcevski and Larsen, 2009).

The local fraction takes no account of the amount of link evidence, but this seems undesirable. We want to reduce the impact of the global link structure while retaining the impact of the amount of local links. We want to combine *local importance* and *local specificity*. Similar to the well-known term-weighting scheme TF·IDF, we can use the inverse of the log of the global degree as an inverted document frequency. The *weighted degree* is computed as:

$$\deg_{\text{weighted}}(d) = \begin{cases} \deg_{\text{local}}(d) / \log(1 + \deg_{\text{global}}(d)) & \text{if } \deg_{\text{global}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Pages with an extremely high global degree have a high a priori probability of a high local degree, which we tone down by dividing by the log of the global degree. Pages with high local degree are still promoted above pages with low local degree, but with equal amounts of local evidence, the page with less global evidence is preferred. We add one to the log of the global degree to ensure the denominator is more than zero. This has a minor impact on pages with a high global degree and tones down the impact of pages with a low global degree.

	Global					Local				
Degree	min.	max.	med.	mean	stdev.	min.	max.	med.	mean	stdev.
In-degree	0	74,937	4	20.63	282.94	0.00	48.83	1.14	3.17	6.65
Out-degree	0	5,098	12	20.63	36.70	0.04	21.01	2.34	3.17	3.37
Union	0	75,072	16	37.65	287.87	0.04	51.11	3.14	5.14	7.19
Intersection	0	1,488	2	3.62	9.10	0.00	14.68	0.44	1.20	2.15
	Fraction					Weighted				
In-degree	0.00	0.67	0.05	0.12	0.03	0.00	12.03	0.14	0.63	4.43
Out-degree	0.00	0.50	0.07	0.10	0.01	0.00	4.63	0.24	0.49	0.80
Union	0.00	0.49	0.07	0.10	0.01	0.00	9.74	0.32	0.73	2.91
Intersection	0.00	0.69	0.04	0.13	0.03	0.00	6.53	0.06	0.39	1.20

Table 20: Link statistics of the Wikipedia collections. Local statistics are macro averages over 221 topics.

5.2 RELATION BETWEEN DEGREES

In this section we analyse the extent to which degrees are correlated to each other. Local link evidence is derived from the global link structure, so how are local and global degrees related to each other? Incoming and outgoing links are derived from the same graph; how are they related to each other? We will first discuss statistics of the new types of link evidence discussed above. After that, we look at the relation between global and local link evidence and between incoming, outgoing and undirected link evidence.

5.2.1 *Degree statistics*

What do the new weighted degree distributions look like? Are they still power law distributions like the normal degrees?

Incoming and outgoing links are different types of evidence, but they have a similar impact on retrieval performance. Since they are different types of evidence, they might be complementary. We also look at undirected and bidirectional link evidence. The union of the in- and out-degree is the undirected degree, or the total number of pages that a page is connected to. The intersection of in- and out-degree is the set of bidirectional links, where pages A and B link to each other. The graph contains 12,401,667 undirected links and 1,182,558 bidirectional links (9.5%). Degree statistics are shown in Table 20.

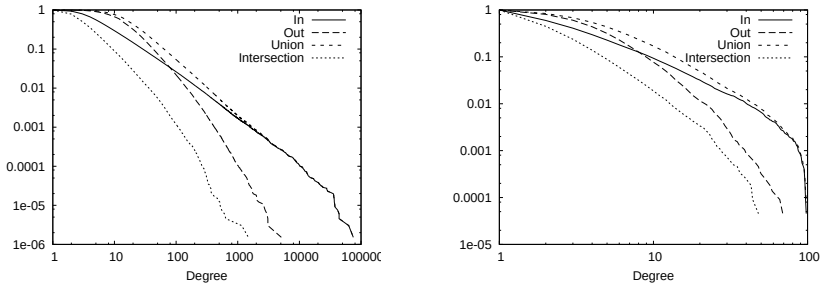

Figure 16: Complementary cumulative distribution function of retrieved documents over the global degrees (left) and over the local degrees (right).

Globally, the maximum and standard deviation of the in- and out-degrees differ by an order of magnitude (e.g. 74,937 versus 5,098), while among the local degrees this difference is smaller (48.83 versus 21.01).

The number of local links is of course smaller than the number of global links, but the link density is higher. An average of 37.65 undirected links (union) per document in the global set of 659,388 documents means that an average document is connected to 0.0057% percent of all documents. An average of 5.14 undirected links in a local set of 100 documents means that an average document is connected to 5.14% of the local documents. The density thus increases by almost three orders of magnitude.

The proportion of links that are bidirectional is the fraction of union that is also in intersection. This proportion is much higher in the local set ($1.20/5.14 = 0.23$) than in the global set ($3.62/37.65 = 0.10$). This can be partly explained by the higher link density in the local set. Given that page A links to page B, what is the probability that B links to A? In the global link graph, this is $20.63/659,388 = 0.00003$, while in the local link graph, this is $3.14/99 = 0.03$. However, the proportion of bidirectional links in the local set is much higher than $0.10 + 0.03 = 0.13$. The nature of Wikipedia links may also play a role: the Wikipedia guidelines on linking (Wikipedia, 2009) state that a link to another document should only be made when it is relevant to the context. Thus, in a set of documents related to the same query, many documents will be related to each other and therefore cross-linked. As we will see in the next chapter (page 140), the proportion of bidirectional links is indeed higher among linked pages that are semantically related to each other than among pages that are not.

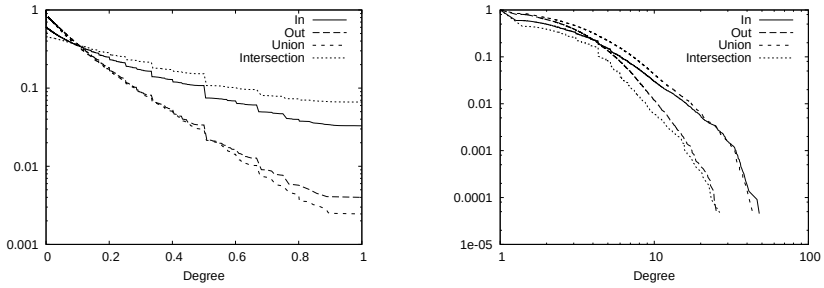

Figure 17: Complementary cumulative distribution function of retrieved documents over the local fractions (left) and the weighted degrees (right).

Figure 16 shows the CCDF of the global (left) and local degrees (right). We see that globally, the union degree distribution closely follows the in-degree distribution at the high degrees. The explanation is that the maximum out-degree is around 5,000 while the maximum in-degree is around 75,000. At the high end the union degrees are almost completely dominated by the contribution of the in-degree. We expect that in the top of the ranking, the in- and union degrees are strongly correlated. For the local degrees the pattern is very similar.

Figure 17 shows the CCDF of the local fractions (left), and weighted degrees (right). Note that the fractions are plotted on a normal scale, whereas the degrees are plotted on a logarithmic scale. The reason is that fractions lie between zero and one. Plotting them on a logarithmic scale between one and two (fraction plus one) results in straight lines, suggesting the local fractions follow a power law distribution.

With the local fractions we see that the local fraction of incoming links (in-fraction) distribution closely follows the intersection distribution and the out-fraction distribution closely follows the union distribution. With the weighted degrees it is vice versa.

5.2.2 *Correlation of degrees*

To what extent are local and global degrees related to each other? And are the local fractions more independent from the global degrees? We look at the Kendall's τ rank correlation between global, local and weighted degrees and local fractions. Many pages have the same degree, so ranking on degrees leads to many tied ranks. When computing the

	Incoming				Outgoing			
Degree	Global	Local	Fraction	Weighted	Global	Local	Fraction	Weighted
Global	–	0.46	0.01	0.29	–	0.32	-0.28	0.09
Local		–	0.57	0.87		–	0.44	0.84
Fraction			–	0.71			–	0.62
Weighted				–				–

Table 21: Correlation of global, local, fraction and weighted degrees over the top 100 results

rank correlation we should take this into account. Therefore, we ignore the pairs that are tied on both variables and compute Kendall's τ as:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_c + n_d + e_X)(n_c + n_d + e_Y)}}$$

where e_X is the number of pairs tied on y but not on x and e_Y is the number of pairs tied on x but not on y. Pairs tied on both x and y play no role in the computation of τ. We compute the rank correlation per topic and report the macro average over all 221 topics.

In Table 21 we see the rank correlations over the top 100 retrieved results for the global, local and weighted in-degrees (left) and out-degrees (right). The global and local in-degrees are moderately correlated (0.46). The global link structure has a significant impact on the local link degrees. Local degrees thus reflect both topical relatedness and document importance to some extent. As expected, the correlation with the global degrees decreases as we put more emphasis on the local context. The weighted in-degrees correlate somewhat less (0.29) with the global in-degrees than the local in-degrees. The local in-fractions appear to be uncorrelated with the global in-degrees (0.01). For the outgoing links, the correlations with the global degrees are lower, except for that of the local out-fractions, which are negatively correlated with the global out-degrees (-0.28). Recall from the previous chapter that the global out-degrees are strongly correlated with document length (see Table 12). This could mean that the out-fractions are also negatively correlated with document length and thus promote shorter documents that are very focused on the topic.

The local degrees are moderately correlated with the local fractions, with a stronger correlation for the in-degrees (0.57) than for the out-degrees (0.44). Down-weighting by the full global degrees really affects the ranking. The local degrees are strongly correlated with the weighted

degrees, both for incoming (0.87) and outgoing links (0.84), so down-weighting by the log global degrees apparently does not affect the ranking much. Since the weighted degrees are somewhere in between the local degrees and the local fractions, their correlations are strong with both. The log tones down the global degree to such an extent that the weighted degrees are closer to the local degrees (0.87 and 0.84) than to the fractions (0.71 and 0.62).

In summary then, to what extent are global, local and weighted degrees related to each other? As discussed above, global and local degrees are moderately correlated with each other. The ranking based on local link degrees is clearly influenced by the global link structure and possibly signals both topical relevance and document importance. The fraction of the global links that are present in the local graph is unrelated to the global degree and is only moderately related to the local degree. The TF·IDF-like weighted degree is very similar to the local degree but is less dependent upon the global degree, therefore less related to document importance and more with topical specificity.

5.2.3 *Directed and undirected link degrees*

In the previous chapter we saw that local in- and out-degrees have a similar impact on average precision. A simple explanation would be that they are strongly correlated, as they are derived from the same link graph. Incoming and outgoing link evidence are necessarily related in some way: a link between two documents is incoming link evidence for one document and not the other, and vice versa for outgoing link evidence.

By combining two different types of evidence, they might lessen each others impact. Given a link from document A to B, incoming link evidence will promote document B, outgoing link evidence will promote document A, whereas the undirected degree makes no distinction between the two. If documents with a high in-degree tend to have a low out-degree and vice versa, in union they would cancel each other out. If, on the other hand, the documents with high in-degree also tend have a high out-degree, in union they would increase the gap between the highly connected documents and those with only a few connections.

We will first look at the correlations of global degrees, then proceed with the local degrees, and the weighted and log weighted degrees.

If we focus on the global degrees and the correlation between the different link directions over the top 100 results (left side of Table 22), we see that the in- and out-degree rankings are moderately correlated (0.43).

	Top 100					Top 10				
Degree	In	Out	Un.	Int.	Cont.	In	Out	Un.	Int.	Cont.
In	–	0.43	0.63	0.74	0.02	–	0.31	0.83	0.43	0.04
Out		–	0.80	0.50	0.00	0.01	–	0.42	0.05	0.00
Un.			–	0.58	0.01	0.58	0.25	–	0.34	0.07
Int.				–	0.01	0.45	0.46	0.49	–	-0.02
Content					–	0.06	0.05	0.06	0.06	–

Table 22: Correlation of global degrees over the retrieved top 100 and top 10 of the 221 topics.

Recall from Table 12 that their linear dependence (Pearson correlation) is 0.19. In ranking they are more similar to each other. The in-degrees are more strongly correlated with the union and intersection degrees (0.63 and 0.74 respectively) than with the out-degrees, which is to be expected, given that the in-degree is part of the union and intersection degrees. The out-degrees are more strongly correlated with the union (0.80) than the in-degrees (0.63). We also show the rank correlations of the degrees with the content-based ranking (column 6). Obviously, the correlations are close to zero because the global degrees are independent of the content score.

The overall correlations give a broad idea of the relationship between degrees. Given that most documents have a low in- and out-degree, the correlation is dominated by these low degrees where the mass of the distribution is peaking, while we are mostly interested in the other end with the highest degrees. On the right of Table 22 we see the rank correlations between the top 10 results of the different degrees. Over the top 10, the correlation is not symmetrical: the top 10 documents by in-degree can be different from the top 10 documents by out-degree. That is, we take the top 10 results ranked by the column (say, in-degree) and compare their ordering with how the same 10 documents are ranked by the row (say, out-degree). Because of the asymmetry, we also look at the top 10 results ranked by out-degree and compare their ordering with how the in-degree orders them.

Among the top 10 documents according to in-degree, the correlation is low to moderate with out-degree (0.31) and intersection (0.43), but strong with union (0.83). In fact, the correlation between in-degree and union is higher in the top ranked documents than overall (0.63), which fits with what we saw in the degree distributions (Section 5.2.1). At the low end of the distribution, the union is more similar to the out-degree

	Degrees					Fractions				
Degree	In	Out	Un.	Int.	Cont.	In	Out	Un.	Int.	Cont.
In	–	0.27	0.78	0.40	0.12	–	0.06	0.12	0.28	-0.02
Out	0.10	–	0.46	0.20	0.12	0.10	–	0.52	0.09	0.00
Un.	0.61	0.28	–	0.36	0.12	0.24	0.51	–	0.11	0.06
Int.	0.45	0.46	0.43	–	0.12	0.34	0.11	0.10	–	-0.06
Content	0.17	0.19	0.19	0.19	–	0.12	0.13	0.12	0.15	–

Table 23: Correlation of local degrees (left) and local fractions (right) over the top 10.

while at the high end of the distribution, the union is more similar to the in-degree. The out-degree top 10 ranking is not correlated (0.01) with the in-degree ranking of the same documents. This shows where in- and out-degree are complementary. The out-degree pushes different documents to the top than the in-degree. The same holds for the out-degree top 10 and the intersection-based ranking (0.05). Only the union ranks the top 10 out-degree documents somewhat similarly (0.42). The documents with high union and intersection degrees get most of their evidence from the in-degrees. Even in the top of the ranking the degrees do not correlate with the content-based ranking (column 11).

The local degrees and fractions and the weighted degrees over the top 100 show roughly the same correlations between incoming and outgoing degrees and their union and intersection. Because we are mainly interested in how they compare in the top of the rankings, in Table 23 we only show the correlations over the top 10 results. We leave out the numbers of the weighted degrees, as their correlations are similar to the local degree correlations. The weights have little impact on the rankings.

We first discuss the correlations between the local degrees (left side of Table 23). Over the top 10 results, in- and out-degree correlate only weakly with each other (0.27 and 0.10). Local in- and out-degrees are not more strongly correlated than global in- and out-degrees. Local link evidence is thus not more symmetrical in terms of the degree structure than global link evidence. Local in- and out-degrees promote different documents. We looked at overlap between the top 10 documents of the local in- and out-degree and found that, averaged over 221 topics, the overlap is 4.7; thus each has 5.3 documents in the top 10 that are not in the top 10 of the other. The correlation with the content-based ranking is still low (column 6), but higher than for the global degrees.

The local degrees are related to the topic and thereby, to some extent, to the content of the documents.

From the local degrees to the local fractions (right side of Table 23), the main differences are that out-degree and union are more strongly related (0.52 and 0.51) as reflected by the degree distributions in Figures 16 and 17. The relation between incoming and outgoing link evidence has almost disappeared (0.06 and 0.10). The correlation between the fractions and the content-based rankings (column 11) are lower than those of the local degrees. Local or topical specificity seems unrelated to the content-based score.

For all the types of link evidence, the correlation with the content-only ranking is relatively low. Link evidence is thus quite different from textual evidence.

To what extent are the directed and undirected degrees related to each other? We found that there is a moderate correlation between document rankings based on in- and out-degree. However, this correlation is smaller in the tops of both rankings. There is also substantial difference between documents in the top 10, indicating that in-degree and out-degree provide information about different documents, which is reflected in the difference of their precision curves. Since both degrees give the same overall performance boost, their union might give a further boost. Because the out-degree distribution has a smaller spread, it has a smaller impact on the ranking. When combining it with the in-degree, that is, using the union of the degrees, it curbs the impact of the in-degree to some extent. As the union degree shows a very strong correlation with both degrees, and thus ranks documents similarly, it is also possible that it leads to only a small further performance boost.

The local degree structure is not more symmetric than the global degree structure. This is partly due to the moderate correlation between global and local degrees, but also partly to the fact that incoming and outgoing link evidence point in opposite directions and apparently promote different documents. The degree structure does not reveal to us any difference in symmetry between query-independent and query-dependent evidence. It also suggests that the relation between incoming and outgoing links is independent of whether the entire graph or a subgraph is used. Of course, this could be due to the fact that we average over a large number of topics, as there might be large differences between the correlations of individual topics.

We have found different levels of link evidence that range from completely query-independent to highly query-dependent, and seen that the similar impact of local incoming and outgoing link evidence

is not caused by symmetry in the degree structure. We now turn to investigate the ability of global and local incoming and outgoing link evidence to distinguish between relevant and non-relevant documents.

5.3 LINK EVIDENCE AND RELEVANCE RANKING

In the previous chapters we looked at the impact of combining link evidence with the document content evidence. Here, we look at how documents are ranked by link evidence alone. This allows us to directly compare query-independent and query-dependent link evidence for ad hoc retrieval in their ability to rank relevant before non-relevant documents. If we make link evidence more sensitive to the topical context, we expect it to become more effective for ad hoc retrieval. How does global link evidence compare to random ordering? How do local link evidence and weighted link evidence compare to content-based retrieval?

Because we only use the top 100 results for the local link degrees, we evaluate all runs only up to the first 100 results per topic. Because the full runs have many more results beyond the top 100, some of which are relevant, the MAP over the top 100 results is lower than the MAP over the full run. We compare the link-only ranking with the text-based ranking and a random ordering of the results. For the random ordering, we take the top 100 results of the text-based run and assign a random score to each document and rank the documents accordingly. Because two random orderings can have radically different performance scores, we average the scores of the randomly ordered run over 100 iterations. With 221 topics, this means the overall scores are based on the individual scores of 22,100 random orderings, giving very stable and reliable results.

The results are shown in Table 24. The text retrieval baseline (*Content*) leads to a much better ranking than the random ordering. We would expect that the ranking based on link evidence is better than random. If we look at the global evidence, we notice that in-degree performs better than random mainly at the early ranks (0.5025 versus 0.3962 for MRR, 0.2805 versus 0.2191 for P@10), while out-degree has worse MRR than random ordering (0.3638), perhaps because out-degree correlates with document length, and is made redundant by the length prior. The ability of global link evidence to distinguish relevant from non-relevant is not symmetric. The union of in- and out-degrees is less effective than in-degree alone.

Run	MAP	MRR	P@10	P@30	MAP	MRR	P@10	P@30
Random	0.1254	0.3962	0.2191	0.2181	0.1254	0.3962	0.2191	0.2181
Content	0.2679	0.8119	0.4937	0.3621	0.2679	0.8119	0.4937	0.3621
	Global				Local			
Indegree	0.1435	0.5024	0.2805	0.2490	0.2117	0.7259	0.4186	0.3243
Outdegree	0.1315	0.3638	0.2394	0.2377	0.2129	0.6374	0.4127	0.3321
Union	0.1382	0.4688	0.2611	0.2391	0.2206	0.7279	0.4235	0.3395
Intersection	0.1417	0.4811	0.2751	0.2446	0.2139	0.7132	0.4104	0.3205
	Fraction				Weighted			
In-degree	0.1925	0.5067	0.3457	0.3128	0.2294	0.7593	0.4348	0.3407
Out-degree	0.1921	0.5365	0.3543	0.3192	0.2268	0.7334	0.4258	0.3351
Union	0.1927	0.5434	0.3439	0.3103	0.2270	0.6714	0.4213	0.3436
Intersection	0.1999	0.5599	0.3552	0.3119	0.2382	0.7438	0.4394	0.3514

Table 24: Retrieval performance using link evidence alone on the INEX 2006–2007 Ad Hoc Track topics.

What is most striking is that all variants of local link evidence perform much better than the global link evidence. Global link evidence barely improves upon a random ordering while local link evidence is much closer in performance to the text-based ranking.

The local in- and out-degree have similar MAP scores (0.2117 and 0.2129), indicating the symmetry of the evidence. The union is even better (0.2206), showing that using the evidence in both directions is beneficial, again supporting the idea of symmetry.

The local fractions have somewhat lower scores than the local degrees. In other words, the amount of local evidence is more effective than the fraction of local evidence. Again, in- and out-fractions have similar MAP (0.1925 and 0.1921). Here the intersection is better than in- and out-degrees, perhaps because the smaller spread of the intersection degrees demotes high-degree pages less.

The weighted degrees are more effective than the local degrees, showing that making the evidence more sensitive to the topic is beneficial for ad hoc search. However, the fact that the weighted degrees are also better than the fractions indicates that the influence of the global structure does have a positive impact. Incoming and outgoing link evidence have similar MAP score, and their union is more effective than either by itself.

The fact that local link evidence is much more effective than global link evidence and random ordering is strong support for our conjecture that local link evidence is related to topical relevance. Putting more emphasis on the local context by using a TF·IDF weighting of local and global degrees improves the ranking further. Although we cannot directly observe the relation between local link evidence and topical relevance, the fact that performance increases as we make link evidence more sensitive to the topical context, and the similar performance of local incoming and outgoing link evidence, strongly suggest the relation with topical relevance. Global link evidence, although nowhere near as effective, is related to relevance, albeit a query-independent aspect of relevance.

This concludes our analysis of the ability of link evidence to distinguish between relevant and non-relevant documents. The next step is to zoom in on the relevant documents and investigate how link evidence affects the internal ordering of relevant documents.

5.4 LINK EVIDENCE AND AMOUNT OF RELEVANT TEXT

Query-dependent link evidence is more useful than query-independent link evidence for identifying relevant documents. But not all documents are equally relevant. Some documents might be mostly off-topic and only mention the topic in a few sentences, while others might be fully on-topic and cover the topic exhaustively.

This is where the INEX Ad Hoc relevance judgements on the Wikipedia collection allow a much deeper analysis than most other IR test collections. INEX studies the effectiveness of focused retrieval techniques, giving precise information about the location and amount of relevant text within documents. For the INEX Ad Hoc Track, assessors are asked to highlight all and only relevant text within each pooled document (Lalmas and Piwowarski, 2006, 2007). The relevance judgements thus contain not just binary judgements but the size (in number of characters) and location of the relevant text within relevant documents.

This allows us to study the relation between link evidence and the amount of relevant text. We make the assumption that documents that have more relevant text discuss the topic more exhaustively and are therefore more important to the topic. This assumption is not true in all cases, as some documents might be verbose and cover only a small aspect of the search topic in great detail, while others might be more concise but completely satisfy a user's information need in a few sentences or paragraphs. However, in general we expect the amount of

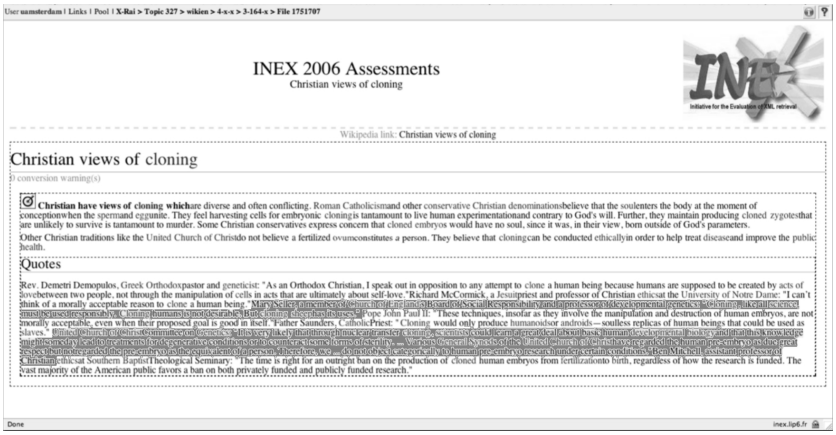

Figure 18: Example of relevant text highlighted by an assessor of the INEX Ad Hoc Track.

Degree	*In*	*Out*	*Union*	*Inter*
Global	0.12	0.17	0.14	0.14
Local	0.19	0.19	0.21	0.21
Fraction	0.11	0.03	0.05	0.14
Weighted	0.19	0.16	0.19	0.19

Table 25: Rank correlation coefficients between relevant text size and global degrees, local degrees, local fractions and weighted degrees.

relevant text to be a reasonable indicator of the utility of a document for ad hoc search.

To illustrate the detailed relevance information available through the INEX assessments, Figure 18 shows an example relevant document with some of its text highlighted. Assessors can highlight any of the text, and select multiple passages of relevant text. The resulting relevance judgements contain information about the character offset and length of each relevant passage. For our purposes we only consider the total amount of relevant text and the fraction of text relevant—the amount of relevant text divided by the total amount of text—in each article. The offset information is less useful for our current analysis because we look at the document level rather than the sub-document level.

In Table 25 we see the rank correlation between the amount of relevant text in articles and the link degrees. Note that these correlations are over

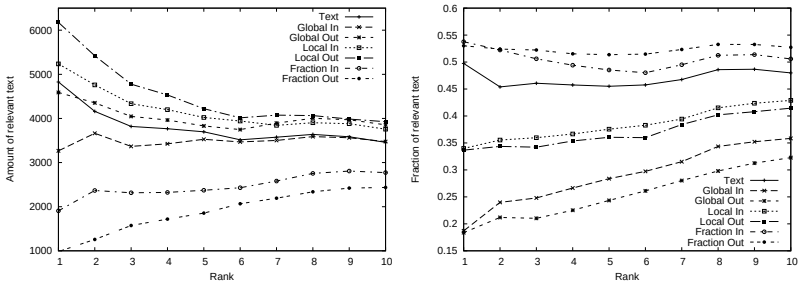

Figure 19: The average amount of relevant text at ranks 1 to 10 for the retrieved relevant documents ranked by content or link degree.

the relevant articles only. We are interested in the internal ranking of the *retrieved relevant* documents, therefore filter all the irrelevant documents out of the top 100 results. Of the global degrees, the out-degree has the highest correlation (0.17) with the amount of relevant text. This is probably due to the strong correlation between global out-degree and document length. Longer documents can have larger absolute amounts of relevant text. The local degrees correlate more with the amount of relevant text (between 0.19 and 0.21) than the global degrees, although all coefficients are still low. The fractions—especially the out-fractions (0.03) and union fractions (0.05)—are almost unrelated to the amount of relevant text in articles. The coefficients of the weighted degrees are very similar to those of the local degrees.

Again, the overall correlation does not tell us anything about the most interesting part of the data. What we want to know is whether the documents with the highest degrees are also the documents with the most relevant text. The average amount of relevant text over the first 10 retrieved relevant documents when ranked by degree is shown in Figure 19 (left). At each rank we computed the amount as the average over the n-th relevant documents over all topics that have at least n relevant documents in the top 100. We left out the curves for the weighted degree to keep the figure readable. They fall between the local degrees and local fraction curves, but closely follow the local degree curves.

The *Text* baseline has an average of almost 5,000 relevant characters in the highest ranked relevant document. This amount quickly drops to around 4,000 characters in the second and third relevant documents, and then slowly drops to around 3,500 at the tenth relevant document. The content-based evidence seems related with the amount of relevance.

The amount of relevant text in documents gradually drops as the amount of textual evidence decreases.

The global in-degree shows no relation with the amount of relevant text. The relevant document with the highest in-degree has around 3,200 highlighted characters, while the tenth relevant document has around 3,400 highlighted characters. In contrast, the out-degree decreases with the amount of relevant text in documents. The global out-degree is strongly correlated to the length of documents. We already saw that document length is related to the *probability of relevance*, but document length also controls how much (relevant) text a document can possibly have. Global out-degree promotes longer documents, and within the set of relevant documents, the longer documents apparently have more relevant text. The highest ranked relevant document according to textual evidence has more relevant text than the highest ranked relevant document according to global out-degree. However, at ranks two to ten, the global out-degree finds more relevant text than the text-based baseline. Global in- and out-degrees are dissimilar in their relationship with the amount of relevant text.

For the local degrees, the amount of relevant text clearly decreases with increasing rank. The highest ranked relevant document according to local in-degree has 5,237 highlighted characters on average, while the document with highest out-degree has around 6,177 highlighted characters. At the tenth relevant document, in- and out-degree have 3,757 and 3,926 highlighted characters. The local in- and out-degrees are thus more related to the amount of relevant text in documents than the global degrees, and are more similar to each other in terms of the internal relevance ranking as well. This supports our conjecture that for the relation with topical relevance, the direction the local links plays no role; local link evidence is symmetric.

The local in-fractions show no relation with the amount of relevant text in documents. From rank 1 to 10, the average amount of highlighted text stays close to 2,700 characters. The local out-fractions have an inverse relation with the amount of relevant text in documents. The average amount of relevant text increases from 1,045 to 2,476 highlighted characters from rank 1 to 10. A possible explanation is that the fraction of outgoing links favours shorter documents with few outgoing links of which most are in the local graph over longer documents that have many outgoing links of which many are missing. These shorter documents have less relevant text in absolute terms, but might have a larger fraction of the text highlighted.

Therefore, we also look at the fraction of text highlighted over ranks. Within the same set of retrieved relevant documents, we look at the average percentage of text that is highlighted in the right side of Figure 19. For the *Text* ranking, the fraction of text highlighted fluctuates between 45% and 50%. Combined with the fact that the highest ranked relevant documents have more relevant text, this suggests that the content-based ranking is related with topical relevance.

Both the global in- and out-degrees show an inverse relation with the fraction of highlighted text in relevant documents. For the in-degree, the fraction goes up from 19% of the highest ranked relevant document to 36% of the tenth ranked relevant document. For the out-degree the percentages go up from 18% to 32%. The global out-degrees are positively related with the amount of relevant text, but negatively related with the fraction of relevant text. This shows that the global out-degree really promotes longer documents that have a larger scope than the search topic alone, and only as a consequence of this promote more relevant text. Again, global link evidence is not symmetric in its relation with the amount and fraction of relevant text in documents.

The local degrees are also negatively correlated to the fraction of highlighted text, but the percentages are substantially higher, going up from 34% at rank 1 to 43% at rank 10 for the in-degree and from 34% to 41% for the out-degree. These percentages are lower than for the text-based ranking. Textual evidence keeps more focus on the topic, but local link degrees find more relevant text early on. The difference between the local in- and out-degrees are small, supporting the conjecture that link evidence for topical relevance is symmetric in identifying and promoting relevant text.

The local fractions have the strongest relation with the fraction of highlighted text. At all ranks from 1 to 10, both the in- and out-fractions rank documents with a larger percentage of relevant text in the top 10 than the *Text* baseline does. Local specificity is a good indicator of topical specificity. The first ten relevant documents according to local out-fraction have more than 51% of their text highlighted. Thus, although the weighted degrees are less effective for finding documents with large amounts of relevant text, they keep strong focus on the search topic and instead first rank relevant documents that are mostly on-topic.

This shows the relation between the amount of local link evidence and the amount of relevant text, and between the fraction of link evidence present in the local graph and the fraction of relevant text.

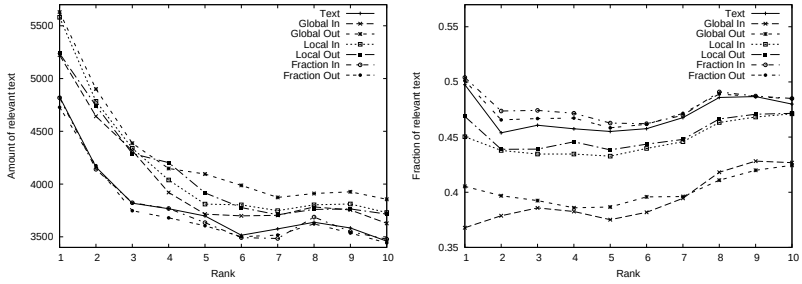

Figure 20: The average amount (left) and fraction (right) of relevant text at ranks 1 to 10 for the retrieved relevant documents ranked by the combination of content and link degree.

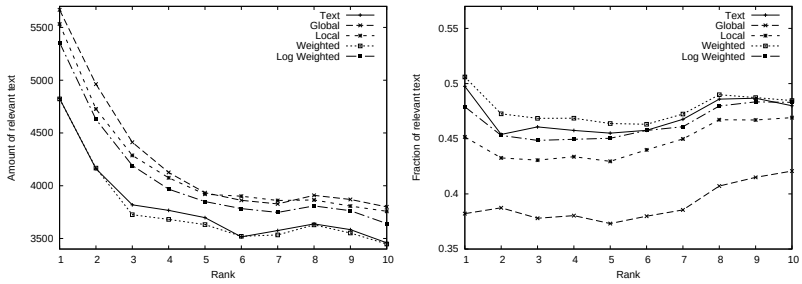

Figure 21: The average amount (left) and fraction (right) of text that is relevant at ranks 1 to 10 for the retrieved relevant documents ranked by the combination of content and union of in- and out-degree.

Local degrees reflect the exhaustivity dimension while local fractions reflect the specificity dimension.

In Figure 20 we see the amount (left) and fraction (right) of relevant text among the relevant documents using link evidence in combination with the text-based retrieval score. We see that the difference between incoming and outgoing link evidence is small, but that the level—global, local or weighted—determines the shape of the curve. Global and local degrees increase the amount of relevant text in the first 10 relevant documents, which is barely affected by the local fractions. The fraction of relevant text is slightly increased by the local fractions, somewhat decreased by the local degrees and strongly decreased by the global degrees. The local fractions seem to have little impact on the internal ranking of relevant documents, but the local degrees seem to improve the internal ranking.

In Figure 21 we see the amount (left) and fraction (right) of relevant text over the first 10 relevant documents for the baseline combined with the union of in- and out-degrees. The global degrees increase the amount of text in the first 10 relevant documents but at the price of a large drop in the fraction of relevant text. The highest ranked relevant documents are much longer and thus also have relatively more irrelevant text than the first 10 relevant articles of the baseline. The local degrees also increase the amount of relevant text but keep a much better focus on the topic because they introduce a smaller fraction of irrelevant text. The local fractions have little impact on the amount of relevant text at each rank but slightly increase the fraction of relevant text. They help increase the topical focus among the first 10 relevant documents by promoting documents that are more on-topic. Finally, the weighted degrees combine the increase in amount of relevant text from the local degrees with the better focus of the local fractions. The first 10 relevant documents have almost the same fraction of relevant text as those of the baseline, but a larger amount of relevant text.

Link evidence for document importance promotes more text and as a consequence also more relevant text, while link evidence for topical relevance focuses on relevant text.

To summarise, global link evidence is not symmetric in its relationship with the amount of relevant text and is unable to promote relevant text without losing focus. This makes sense if global link evidence is related to document importance but not to topical relevance. Important documents have a higher a priori probability of being relevant, but link evidence signalling importance cannot tell us how topically focused a document is. Local degrees have the strongest relation with the amount of relevant text. The local degree is a combination of topical specificity and global connectedness. As a consequence, it keeps more focus on the search topic than global link evidence and is better than the local fraction at ranking documents with respect to the amount of relevant text within them. Although the content-only score is a better indicator of the relevance of a document—it has higher precision scores in Table 24—the local link evidence seems a better indicator of the amount of relevant text in documents.

These findings offer further support of our conjectures. Global link evidence is unable to keep focus on the search topic but can only promote longer documents that have higher probabilities of being relevant. Therefore, global link evidence seems to signal document importance but not topical relevance, and is asymmetric in its relation to relevance. Local link evidence can keep focus on the search topic

and at the same time promote documents with more relevant text. If local link evidence signals topical relevance, it makes sense that the amount of evidence is related to the amount of relevant text, and that the fraction of links that is present locally is related to the fraction of text that is relevant. Local link evidence is also more symmetric in its relation to the exhaustivity and specificity dimensions of relevance, which further suggests its relation with topical relevance.

5.5 DISCUSSION AND CONCLUSIONS

In this chapter we investigated when and to what extent link evidence is related to document importance and topical relevance in Wikipedia. Our conjectures were:

1. Global link evidence is query-independent and therefore not related to topical relevance, but to document importance.
2. Local link evidence is query-dependent and is related to topical relevance.

First, we wanted to know how we can make link evidence less dependent on the global link structure. The local degree is in part determined by the global degree so the local in-degree might to some extent reflect the importance of pages. We can reduce the impact of the global degree by taking both local and global degrees into account. We introduced a type of evidence that is less related to the global degree and more related to the specific neighbourhood of the top retrieved results. If we take the fraction of the global links present in the local set, we focus on the local or topical specificity of a page, which expresses how strongly a page belongs to the local part of the global link structure. If we divide the local degree by the log of the global degree, we reduce the impact of the global link structure less heavily, and let the degree score reflect how much link evidence there is for a particular page.

The next part of our analysis addressed how different levels and directions of link evidence are related to each other. Our first question was how global, local and weighted degrees are related. The local degrees are moderately correlated to the global degrees, indicating that, if global link evidence signals document importance, then local link evidence might signal it as well, to some extent. The local fractions are unrelated to the global degrees and promote very different documents, suggesting they are unrelated to document importance. The weighted degrees are almost unrelated to the global degrees yet very similar to

the local degrees, but are slightly more sensitive to the local context. These four levels of link evidence reflect a gradual change from purely global to almost purely local.

Our next questions concerned the symmetry of the degree structure of the global and local links. Because they are derived from the same link graph, incoming and outgoing link degrees and their union and intersection are naturally related to each other. We saw that over the top 100 results, the document rankings based on incoming and outgoing link evidence are moderately correlated, but that this correlation is much smaller in the top of their rankings. Incoming and outgoing link evidence promote different documents. These correlations change little between global and local degrees. Local link evidence is not more symmetric in its degree structure than global link evidence.

In the third part of our analysis we looked at the ability of link evidence to distinguish relevant from non-relevant documents. The difference in performance between global and local link evidence is big. In isolation, global incoming link evidence provides a better-than-random ranking of documents based on binary relevance while global outgoing link evidence is almost ineffective. The direction of the links matters. Global link evidence seems to signal document importance but it is not symmetric. Local, query-dependent link evidence is better able to distinguish between relevant and non-relevant documents and shows similar impact of incoming and outgoing link evidence. Link evidence seems to signal topical relevance, as it is symmetric in its ability to identify relevant documents and is more useful than query-independent link evidence. However, if we reduce the impact of global degrees entirely, retrieval performance drops, showing that the global link structure does contribute useful information to the local evidence.

Finally, we looked at how link evidence is related to the amount and fraction of relevant text in documents. Global link evidence is asymmetric in its relation with the amount of relevant text. Out-degrees promote more relevant text but also more irrelevant text through their correlation with document length. Global in-degrees seem unrelated to the amount of relevant text in documents. Local link evidence is more symmetric in its relation to the amount of relevant text in pages and keeps more focus on the search topic than global link evidence. The fraction of links present in the local graph best reflects how specifically a page is about a topic, as it promotes documents that are mostly on-topic. However, there seems to be no relation between local fraction and the amount of relevant text in articles. The weighted degrees keep

better focus on the topic than the local degrees but are slightly less effective for promoting larger amounts of relevant text.

In Wikipedia, global outgoing link evidence is useful to rank documents with a lot of relevant text before documents with less relevant text, but as it is blind to the search topic, it also introduces more irrelevant text early in the ranking. Local link evidence is needed to find documents that are focused on the search topic. The amount of local evidence is related to the amount of relevant text, while the fraction is related to the topical focus of a document. There is no obvious difference between local incoming and outgoing link evidence—in terms of amount and fraction of relevance—suggesting their relation with relevance is the same and thus that it is independent of the link direction.

Our findings support our conjecture that global link evidence can be used as an indicator of document importance but not of topical relevance. The link direction determines which aspect of a document is indicated. Incoming links indicate popularity or authority. Outgoing links indicate length or 'hubness'. Local link evidence can be used as an indicator of topical relevance—regardless of the direction of the links—where the amount of evidence is related to the amount of relevance and the specificity of the evidence is related to the specificity of the relevant documents. The amount of local link evidence is to some extent dependent on the amount of global link evidence, but we have seen that query-dependent link evidence is more useful for ad hoc search than query-independent link evidence. The fact that the amount of local link evidence is more effective for ad hoc search than the specificity of the evidence shows that global link evidence still carries useful information.

The positive impact of combining content and query-dependent link evidence has three factors. First, the ranking of relevant documents with respect to the non-relevant documents is improved. Second, the internal ordering of the relevant documents is improved in terms of the amount of relevant text. Third, the curbing of infiltration ensures the relevant documents keep a strong focus on the search topic.

The transition from query-independent to query-dependent link evidence means we zoom in on the local link structure of the top retrieved results, thereby discarding most of the links in the global structure, which shows that not all links are equally useful. Only links between documents related to the search topic seem effective. In the next chapter we investigate how the semantic nature of links affect their value as evidence for retrieval.

6 LINK EVIDENCE AND SEMANTIC RELATEDNESS

Local link evidence is more related to topical relevance, whereas global link evidence is only related to either document importance or document length. Local link evidence is more effective for improving ad hoc search than global link evidence, indicating that local links are more useful than global links. Note that the query-dependent set of links is a proper subset of the global link structure. Evidently, some, but not all, links are useful as indicators of topical relevance. This confronts us with the question:

- Which links are useful as evidence for topical relevance?

Local links might differ from global links in quantity and quality. Because local link evidence is query-dependent and therefore derived from a set of documents that are plausibly semantically related to each other, local links might be more useful for topical relevance because they better signal the semantic relatedness of documents. *Our hypothesis is that links between semantically related documents are more effective.*

Wikipedia has a complex category structure, providing us with a hierarchical semantic classification of the articles. Thus, we can see whether a link connects two documents in the same category – in which case there is a clear semantic aspect to the link – or between two documents belonging to very different categories. For instance, the article on *Robert Hooke* and the article on *Christopher Wren* link to each other, and both these articles are assigned to the category Fellows of the Royal Society.[1] The article on *Robert Hooke* also has a link to the article on *1679*, the year in which he started a long correspondence with Isaac Newton. Years mentioned in a Wikipedia article are linked to specific pages on those years, listing important events of those years, whether they are related to the topic of the article mentioning the year or not. The article on the year 1679 lists major events that took place in 1679. The two articles *Robert Hooke* and *1679* do not share a category. From the category structure we can derive that *Robert Hooke* is semantically related to *Christopher Wren* but not to *1679*.

1 Robert Hooke and Christopher Wren were colleagues at the Royal Society and worked together rebuilding London after the great fire in 1666.

The category structure allows us to measure the semantic similarity of articles. Perhaps the effectiveness of link evidence for Wikipedia ad hoc retrieval comes from only those links that connect articles belonging to the same or similar categories. On the other hand, for link evidence to be effective, there must be enough links to be able to distinguish between articles. There is a trade-off between quality and quantity. How are the quantity and quality of links related to their effectiveness as evidence for document importance and topical relevance? Intuitively, we would associate quantity with document importance and quality with topical relevance. With very few links, link evidence would suggest all documents to be equally important. We would expect links between totally unrelated documents to be ineffective in improving document ranking in ad hoc retrieval.

Filtering on semantic similarity using the category structure makes the link structure more semantic but also more sparse. Using only the links between documents retrieved for a given query, is another way of filtering. This results in a set of links between documents related to the same topic and thus between documents that are to some extent similar to each other. This way of using feedback gives a similarly semantic but sparse graph. If there is a strong relation between the two methods, the query-independent approach of using the category structure to filter links allows us to reduce the size of the graph and compute the amount of link evidence at indexing time.

This leads to the following specific research questions:

- How can we measure the semantic relatedness between linked documents using the Wikipedia category structure?
- How is the link structure related to the categorical organisation in Wikipedia?
- Are links between semantically related documents more effective?

Fisher and Everson (2003) addressed a question similar to ours and found that links are useful for classification when the link density is sufficiently high and the links are of sufficiently high quality. However, sufficient density and sufficient quality are rather vague notions. Does it mean that the effectiveness of link evidence grows continuously with increasing density and quality? The datasets they used are small and have high link density. They argue that the TREC Web tracks found no benefit in using link information for ad hoc search because the link density in the WT2G and WT10G is too low. But the TREC Web collections are much bigger than the datasets used by Fisher and Everson (2003).

Because link density is quadratically related to collection size, large document collections tend to have low link density—Web pages are connected to a very small fraction of the entire Web. However, we found that link evidence is effective in the INEX Wikipedia collection where the link density is comparable to that of the WT2G collection. The more important measure is link degree. Documents need a minimum number of links for link evidence to have any impact. There is evidence that average link degree increases as collections grow (Leskovec et al., 2005).[2]

Several studies describe approaches to generate links using document similarity and lexical chaining. Blustein (1999) used LSI (Furnas et al., 1988) and a vector space model to match sentences and paragraphs within scientific articles and creates hyperlinks between the best matching document parts to study the value of semantic links for reading hypertexts. He used a minimum of 1.5 links per paragraph, regardless of there being any other part of the document that is semantically related, and found that semantic links are not more useful than structural links derived from the table of contents and the document layout. Green (1998) used WordNet to create lexical chains of related words within a document and created chain vectors per paragraph. Links were generated between paragraphs within a news article based on a similarity co-efficient of the lexical chain vectors. The same was done to generate links between news articles. A user study showed that links generated using lexical chains were considered as useful as links generated using simple term-based document similarity.

Kurland and Lee (2005, 2006) showed that generating links based on document similarity can help improve ad hoc retrieval effectiveness. Using language models to induce hyperlinks, they investigated multiple selection thresholds and weighted links, and found that weighted links lead to the best performance. Instead of generating links, we examine the semantic quality of existing links, and instead of measuring document similarity with language models, we use the explicit human judgements of the Wikipedia category structure. Davison (2000) showed that links on the Web tend to connect pages with topically related content. However, these studies do not prove our assumption that links are effective because they connect semantically related pages.

The rest of this chapter is organised as follows. We first describe the category structure and look at the semantic relatedness of documents in

2 Leskovec et al. (2005) use *density* to mean average degree and argue that link graphs become denser as they grow over time. However, density is usually defined as the proportion of possible links that are actually present. Their data show that the average degree increases while link density decreases.

Section 6.1. In Section 6.2 we address the issue of measuring semantic relatedness using the Wikipedia category structure We then analyse how linked documents in the global and local link graph are distributed over the semantic relatedness measure in Section 6.3. Then, in Section 6.4 we describe experiments with filtering links using the category structure and finish with conclusions in Section 6.5.

6.1 WIKIPEDIA CATEGORY STRUCTURE

The Wikipedia category structure is more or less hierarchical—categories are linked to each other via hypernym/hyponym relations but can have multiple parent categories—and allows us to determine how semantically related two documents are, even when they are not assigned to the same category, based on explicit human judgements of semantic relatedness. In fact, the only explicit human judgements are the assignment of a document to a category and the subsumption relation between two categories. That is, the article *Dog* is explicitly assigned to the category DOGS and the category DOGS is explicitly assigned as a subset of the category CANINES. We can say that the article *Dog* is semantically related to the category CANINES. This relation is not explicitly defined, but derived from the category structure.

Note that, because of the open nature of Wikipedia, anyone can edit the relations between categories, introduce new categories, remove existing categories and assign Wikipedia articles to categories. There is no standard way to create such taxonomies of categories: one person could introduce several intermediate levels between two categories where another would introduce none or only a few. Some of the relations are even cyclic in the sense that two categories can subsume each other. However, we assume that distances at the extreme ends of the distribution—the shortest and longest distances—can respectively be interpreted as semantically related and unrelated.

Some statistics on the category structure are given in Table 26. The category structure of the INEX 2006 Wikipedia collection contains 86,024 distinct categories. The top category in the hierarchy is called CATEGORIES, and almost all categories are connected to this top category via sub-category relations. There are 75,601 categories that contain articles and 10,423 categories that contain no articles but have only sub-categories. For instance, the category NOVELS BY AUTHOR has no articles assigned to it, but has many sub-categories such as CHARLES DICKENS NOVELS and SCIENCE FICTION NOVELS BY AUTHOR. The mean number of articles per category is 16.82, but the median is much lower

Description	min	max	mean	median	stdev
Category					
# articles	0	4,534	16.82	4	56.87
# children	0	1,581	1.69	0	8.55
# parents	0	55	1.69	2	1.17
distance	1	23	7.29	7	1.58
Article					
# categories	1	41	2.20	2	1.64

Table 26: Link degree and category size statistics of the Wikipedia collections.

(4), showing that the distribution is somewhat skewed. There are a few very large categories and many small ones. The mean number of parent and child categories is 1.69, but the median numbers of parent and child categories are 2 and 0 respectively. Thus, most categories are leaves in the category structure, connected to at least 2 broader categories. Some categories have many parents (the maximum is 55), while others have very many—up to 1,581—narrower categories. All articles in the collection are assigned to at least one category, with a mean (median) of 2.2 (2).

6.2 MEASURING SEMANTIC DISTANCE

How can we measure the semantic similarity of two documents in Wikipedia? There is a lot of literature on measuring semantic distances. A good overview can be found in Budanitsky and Hirst (2006). They discuss the difference between the terms *semantic relatedness*, *semantic similarity* and *semantic distance*. Two terms can be semantically related but not semantically similar. The words *coffee* and *mug* are semantically related (functional relationship) but not semantically similar. Antonyms (e.g. *long* and *short*) are also semantically related but dissimilar. Semantic distance can be measured using either semantic relatedness or semantic similarity. In the former case, the semantic distance between *coffee* and *mug* is small, in the latter case, it is not.

Milne and Witten (2008) derive the semantic relatedness of two Wikipedia articles from the link structure, compare their technique against manually defined relatedness measures and find it to be very competitive. For our analysis this method is inappropriate. We want a way to determine the semantic similarity of linked pages that does not

use the links themselves. Instead, we use the explicit relations between categories to measure the semantic nature of a link.

Given that Wikipedia is a collection of interrelated topics, we can view the category structure as a taxonomy of concepts and use methods from computational linguistics to measure semantic relatedness. Strube and Ponzetto (2006) used the Wikipedia category and link structures to measure word relatedness.

Perhaps the easiest way is to make a distinction between a pair of documents belonging to the same category and a pair of documents belonging to different categories, and say that the former pair is *semantically similar* whereas the latter pair is not. To give insight into how links in Wikipedia are related to the category structure, we adopt a path-based measure that simply counts the number of edges along the shortest path between two concept nodes (Rada et al., 1989, Resnik, 1995). The rationale behind this is that "the shorter the path from one node to another, the more similar they are" (Resnik, 1995) and "the relatedness of two words is equal to that of the most related pair of concepts they denote" (Budanitsky and Hirst, 2006).

The open nature of Wikipedia allows people to freely create new categories and hierarchical relations between categories, change existing relations and introduce intermediate levels in the hierarchy. This can cause the category structure to be imbalanced, which makes the path length hard to interpret in absolute terms. However, path-based measures using the category hierarchy have proven to perform well. Strube and Ponzetto (2006) compared the performance of path-based measures with more complex information-content-based measures using both Wikipedia and WordNet (WordNet, 2010) on a number of word relatedness datasets and found that path-based measures are more effective and that on large datasets, Wikipedia is more effective than WordNet. The effectiveness of simple path-based measures for semantic relatedness is supported by Zesch and Gurevych (2007), who performed a similar analysis using the German version of Wikipedia and found that path-based measures perform very well, suggesting the path lengths reflect semantic relatedness well.

An alternative is to ignore the hierarchical structure of the categories and use the co-occurrence of categories, that is, treat two categories as related to each other when an article is assigned to both. Holloway et al. (2007) used this approach to create a category map to visualise the space of topics in Wikipedia. They computed the cosine similarity between two categories based on their co-occurrence in articles. These similarities were then used as weighted edges, resulting in a category network.

However, we found that the category co-occurrence structure is very densely linked, such that the shortest path between two articles is very short even when the articles are entirely unrelated. The hierarchical structure is more fine-grained.

Another alternative is to use the textual content of the categories to measure distance. Kaptein and Kamps (2009) use KL-divergence to compute distance scores between categories based on the text of the documents assigned to each category. They model semantic relatedness by document similarity, which is computed using overlap in sets of tokens.

There are more elaborate algorithms to measure semantic similarity, several of which are reviewed in Budanitsky and Hirst (2006). We opt for the path-based measure over the category hierarchy because it is simple, uses the explicit semantic relation between categories based on human judgement, and has proven to be reasonably effective in semantic relatedness evaluations. Furthermore, in the next sections we will see that this approach is sufficient for our purpose of studying the impact of semantic relatedness on the effectiveness of link evidence for retrieval.

The simple approach of distinguishing between pairs of documents that share a category and pairs that do not, divides the link structure into two sets of links. This division follows the explicit human assignments. The more complex computation of categorical distance allows a more fine-grained division of links but derives this finer resolution of implicit relatedness from the category structure.

The category distance between two documents d_a and d_b is the minimum of the category distances between the categories of d_a and d_b:

$$\text{dist}_{\text{cat}}(d_a, d_b) = \min_{c_i \ni d_a, c_j \ni d_b} \text{dist}_{\text{cat}}(c_i, c_j)$$

where $c_i \ni d_a$ are the categories to which d_a is assigned. The distance $\text{dist}_{\text{cat}}(c_i, c_j)$ between the two categories c_i and c_j is defined as:

$$\text{dist}_{\text{cat}}(c_i, c_j) = \text{dist}_{\text{cat}}(c_i, \text{lso}(c_i, c_j)) + \text{dist}_{\text{cat}}(c_j, \text{lso}(c_i, c_j)) \tag{6.1}$$

where c_i and c_j are two categories, $\text{lso}(c_i, c_j)$ is the lowest super-ordinate (the lowest super category) of c_i and c_j and $\text{dist}_{\text{cat}}(c_i, \text{lso}(c_i, c_j))$ is the number of steps up the hierarchy from category c_i to $\text{lso}(c_i, c_j)$.

When we consider only categories connected to the top category CATEGORIES, the average shortest distance between two categories is 7.29 (median 7) and the maximum is 23.

What is the average category distance between two pages? We randomly sampled one million pairs of documents and computed the

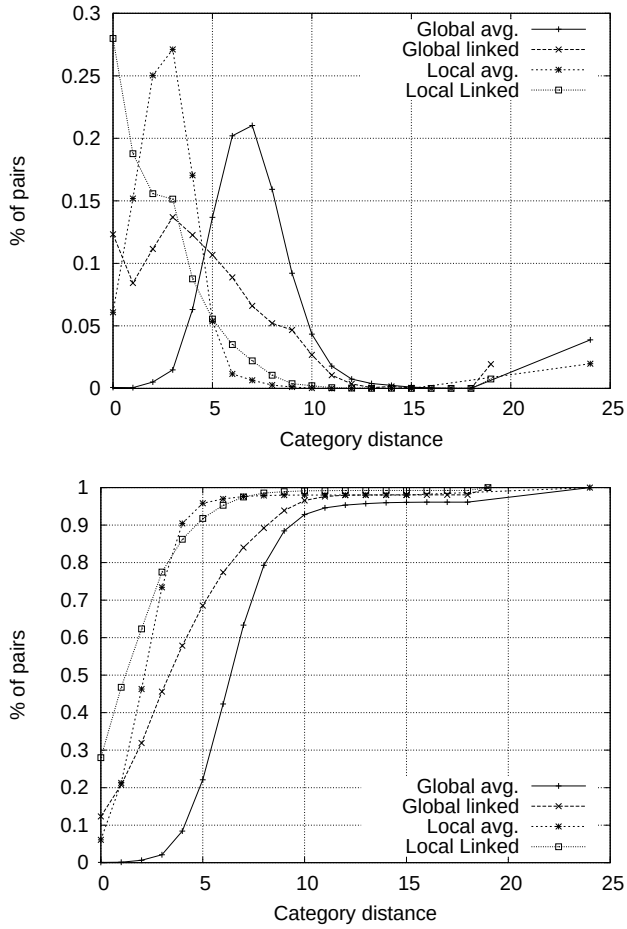

Figure 22: Distribution (top) and cumulative distribution (bottom) of category distances between documents.

shortest category distance between them. The distribution of category distances is shown in the top of Figure 22. The distribution of the global pairs is roughly normally distributed, with a peak at distance 7, with 21% of the documents pairs. The bulk of the document pairs are at a category distance of 4–10, and very few document pairs are semantically close to each other. The right-most data points represent document pairs belonging to unconnected parts of the category structure. Among the pairs that are connected via the category structure the average distance is 6.61, which is slightly below the average distance between two categories, which is 7.29. The cumulative distribution is shown at the bottom of Figure 22.

We also computed the category distance between all document pairs in the local top 100 documents for the 221 topics. This resulted in $221 \cdot \frac{1}{2} \cdot 100 \cdot 99 = 1,093,950$ document pairs. The distribution has roughly the same shape but is shifted towards the smaller distances. In the top 100 retrieved documents, 6% of the document pairs share at least 1 category (distance 0), the most frequent distance is 3 and almost all pairs have a distance less than 6. Among the pairs that are connected via the category structure, the average distance is 2.56. The documents in the top 100 results are more semantically related to each other than in the overall collection.

Now that we have chosen a method to measure semantic relatedness, we turn to the next question. How is the link structure related to semantic relatedness?

6.3 LINKS AND CATEGORIES

One of the main assumptions underlying algorithms like bibliographic coupling (Kessler, 1963b), spreading activation (Anderson and Pirolli, 1984) and relevance propagation (Shakery and Zhai, 2006) is that links are a signal that two documents are topically related to each other. But not *all* links connect documents that are topically related to each other (Qi et al., 2007). The performance of the algorithms just mentioned depends on the "semantic" quality of the links. The Wikipedia category structure provides a manually created semantic organisation of the Wikipedia articles. In this section we want to find out:

- How is the link structure related to the categorical organisation in Wikipedia?

Semantic relatedness is a symmetric relation and therefore independent of the direction of a link. If page A is semantically similar to page

B, then page B is also semantically similar to page A. We look at the shortest category distance between linked articles. The distribution of links over shortest category distance is given in Figure 22 and is shown both globally and locally over the top 100 retrieved results. In the global link structure, around 12% of the links connect two articles sharing at least one category—from here on referred to as *within-category links*, as opposed to *cross-category links*, which connect documents that share not a single category. The most frequent distance is 3 steps, above which the frequency gradually drops to almost 0 at 12 steps. There is a small peak again at the end, for the links between articles assigned to unconnected parts of the category structure (their distance is infinite).

Linked documents tend to be more semantically related to each other than randomly paired documents and share a category much more often. The median category distance of the linked documents is 4 while the median of the randomly paired documents is 7. Among the linked documents that are connected via the category structure, the average distance is 4.04, compared to 6.60 for the randomly sampled pairs. *There is a clear relation between global links and semantic relatedness.* However, compared to the documents in the top 100, the linked documents share a category more often but are also more frequently separated by greater semantic distances. Within the top retrieved results, the global link evidence has a weaker semantic signal than the text evidence.

If links are a signal of semantic relatedness, and semantic relatedness is symmetric, we would expect to find more bidirectional links between semantically related pages. If there is a higher probability that A links to B if A and B are more semantically related, then the probability that B links to A is also higher. In the entire INEX Wikipedia link graph, almost 10% of the links are bidirectional. Of the *cross-category* links, 7% are bidirectional, while among the *within-category* links, this is 33%. This difference is noteworthy for two reasons. First, it is further evidence that links in Wikipedia signal semantic relatedness. Second, as we saw above in Figure 22, the pages in the local set of top 100 results are more semantically related to each other than all the pages in the entire collection. We would thus expect more bidirectional links in the local link graph than in the global link graph, which, as mentioned in the previous chapter (see page 111), is indeed the case.

The category distance distribution over the local links is based on 63,435 links between the documents in the top 100 results of the 221 topics (5.8% of all possible pairs in the local sets). The local links show a very different distribution. Here, the 0 distance links are the most frequent and make up more than 25% of the link set, and the frequency

drops monotonously over category distance, with almost no pairs beyond 8 steps. There is a small set of links between articles assigned to unconnected categories. This means there is a clear relation between local link evidence and semantic similarity. In the query-dependent link set we more frequently find links between articles that are semantically similar. This is not surprising, because each article appears in the local set because it shows similarity with the search query and therefore also with the other documents in the local set. However, the average distance of the linked document pairs is 2.22 while over the entire local set the average is 2.56. In the top 100 results of a given query, the local links provide a stronger signal that two documents are semantically related than the text evidence.

How is the link structure related to the category structure? Compared to a random sample of document pairs, linked documents tend to be more semantically related and more often share a category. There is a clear relation between global links and semantic relatedness. However, this semantic signal is weaker than the text evidence in the top retrieved documents. In the local set, pages that are linked tend to be more semantically related than pages that are not linked. This further suggests that local link evidence signals topical relevance while global link evidence signals document importance but not topical relevance. Is the semantic nature of links also related to their effectiveness for information retrieval? This question is addressed in the next section.

6.4 SEMANTIC RELATEDNESS AND EFFECTIVENESS OF LINKS

By zooming in on the top ranked retrieval results, we filter the link graph on the search topic and as a consequence end up with links between semantically related pages. The global link graph contains the same links but also many more links between semantically unrelated pages. Local links are more effective for ad hoc search than global links and they are also a stronger signal that linked pages are semantically related.

- How is the impact of link evidence related to the semantic nature of links?

We can use the category structure to filter links and thereby control the semantic nature of link evidence. What happens to the impact of link evidence if we remove the within-category links? Does local link evidence become less effective? What happens when we remove only the longest distance links?

To show how the semantic nature of links affects their impact on effectiveness, we use two filtering methods: one where we remove the shortest semantic distance links (the SD filter), effectively degrading the semantic nature of the link graph, and one where we remove the longest semantic distance links (LD filter), effectively improving the semantic nature of the link graph. We filter links based on the path length distance measure described above. In the first filtering step the SD filter removes the links at distance 0, in the second step the links at distance 1, etc. The LD filter first removes the links between pages unconnected to each other via the category structure. In the second step the LD filter removes links at the largest distance (18 steps, see Figure 22), etc.

Note that by filtering we not only affect the semantic nature of the link graph, but also the link quantity. For comparison, we also look at the impact of randomly filtering links. We do this by assigning a random value between 0 and 1 to each page in the collection and sampling n% of the pages by selecting all pages with a value below $\frac{n}{100}$. The degree distribution of an n% sample is determined by the random assignment of the values, so repeating the experiment can result in different distributions. Therefore, the values reported are the averages over 20 iterations.

If we randomly remove links from the graph, we would expect that the degrees change uniformly. That is, all pages are affected in the same way. However, there is a difference between pages with high link degrees and pages with low link degrees. If we remove 90% of the links, a page with a thousand links in the full graph will have roughly one hundred links left in the filtered graph. A page with a single link in the full graph has a 90% chance of having no link in the filtered graph and a 10% chance of having one link in the filtered graph. For such a page, random filtering means all or nothing in terms of link degrees.

Do random filtering and semantic filtering lead to different degree distributions? We see statistics of the distribution in Table 27. Random filtering indeed leads to a similar distribution as the full graph, with all statistics being roughly 10% of those of the full graph. For instance, the maximum in-degree drops from 74,760 to 7,480 and the mean drops from 20.63 to 2.06. Semantic filtering has a very different impact. The shortest distance links (all links betweens documents that share a category, corresponding to just over 10% of the total number of links) are much more evenly distributed. The maximum in-degree is lower (1490) and similar to the maximum out- and union degrees (1488 and 1491 respectively). The standard deviations of the in- and out-degrees

Links	Degree	min.	max.	med.	mean	stdev.
All links	In-degree	0	74760	4	20.63	282.62
	Out-degree	0	5068	12	20.63	36.60
	Union	0	74895	16	37.62	287.55
10% random	In-degree	0	7480	0	2.06	28.31
	Out-degree	0	511	1	2.06	3.91
	Union	0	7523	2	4.09	29.22
$\text{dist}_{\text{cat}} = 0$	In-degree	0	1490	1	2.54	7.49
	Out-degree	0	1488	1	2.54	7.96
	Union	0	1491	2	3.81	10.40
$\text{dist}_{\text{cat}} \geq 9$	In-degree	0	53160	0	2.22	124.97
	Out-degree	0	937	0	2.22	7.06
	Union	0	53191	1	4.32	125.56

Table 27: Degree distribution statistics over all links, a 10% random sample, the shortest category links ($\text{dist}_{\text{cat}} = 0$) and the longest distance links with $\text{dist}_{\text{cat}} \geq 9$.

is also comparable (7.49 and 7.96). This again supports our hypothesis that the topical relevance aspect of link evidence is symmetric and therefore not determined by the direction of links.

The longest distance links (all links between documents separated by a category distance of at least 9 steps, corresponding to just over 10% of the total number of links) have an even more skewed distribution than a random 10% sample of the links. The maximum in-degree is close to that of the full graph (53,160 versus 74,760), while the average degree is much lower (2.22 versus 20.63). A relatively small number of pages is targeted by long distance links and most pages have no outgoing long distance links; the median in- and out-degree is 0. We would expect the longest distance links to have a small impact on effectiveness because most pages have no link evidence.

In the previous chapter we saw that ranking documents using only global link evidence results in a ranking that is better than random. What happens when we increase or decrease the semantic "quality" of global link evidence? We look again at the top 100 results retrieved by the text retrieval baseline, ranked by global link degrees alone.

The impact of filtering on the effectiveness of the global degrees is shown in Figure 23. For comparison, we added the effectiveness of not using link evidence, i.e., a random ordering of documents, as given in

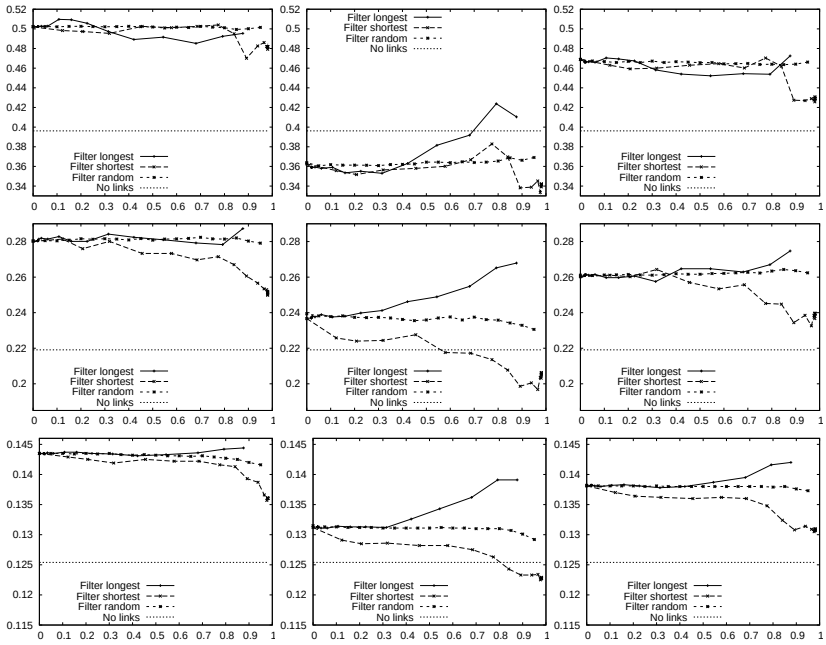

Figure 23: The impact of filtering links on the effectiveness of ranking the top 100 results of the baseline by global link degree in isolation. The x-axis shows the percentage of links removed. The top row shows MRR, the middle row shows P@10 and the bottom row shows MAP.

Table 24 in Section 5.3. The x-axis shows the percentage of links filtered. The top row shows the impact on MRR. Filtering has almost no impact on the effectiveness of global in-degrees, whether we remove links randomly or based on semantic similarity. Even removing the "most semantic" links has no impact. The effectiveness of global in-degree is unrelated to the semantic nature of links. The global out-degrees rank documents worse than random according to MRR and only slightly better than random according to P@10 and MAP, but if we remove links between semantically unrelated pages, the ranking gradually improves and becomes better than random for MRR with links between pages that are no more than 1 step away from each other in the category structure. Random filtering has no impact, while the SD filter starts to hurt performance when we remove all but the 10% longest distance links. For the union, which is dominated by in-degree in the top of the ranking, the impact of filtering is small.

The middle row shows the impact on P@10 and here we start to see larger differences. The Random and LD filters still have little impact on the in-degrees, but removing the shortest distance links hurts performance. Performance stays well above that of random ordering though. For the out-degrees, random filtering has no impact while performance improves as we remove long distance links and deteriorates as we remove short distance links. The union degrees show the same behaviour as the in-degrees, although performance stays below that of the in-degrees.

On MAP (bottom row) the impact of filtering is similar to the impact on P@10. The in-degree performance slightly improves with only the within-category links and drops with only the 10% longest distance links. Out-degree performance improves as the link evidence becomes more semantic and drops as it becomes less semantic and the union degrees behave more like the out-degrees on MAP.

From these observations we learn a few things about the nature of link evidence. First, random filtering has little impact on the global degrees, probably because the link graph is very rich and the high-degree pages are very robust against random filtering.

Second, filtering links between semantically unrelated pages has a positive impact on the effectiveness of the out-degrees but almost no impact on the in-degrees. Why is the impact of filtering bigger on out-degrees than on in-degrees? The in-degree distribution is more skewed, and the difference between the top ranked documents and the rest is big. The out-degree distribution is flatter and the difference between the top ranked documents and the rest is smaller. The high in-degree

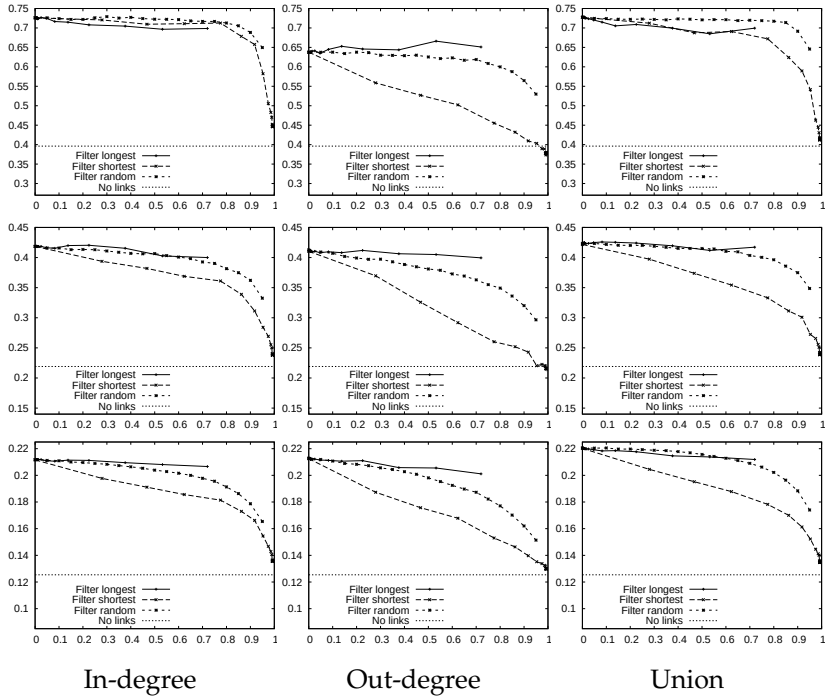

Figure 24: The impact of filtering links on the effectiveness of ranking on local link evidence in isolation. The x-axis shows the percentage of links removed. The top row shows MRR, the middle row shows P@10 and the bottom row shows MAP.

pages are more robust against filtering than the high out-degree pages. But random filtering does not hurt out-degree performance, so why is out-degree affected by the semantic nature of link evidence? A high out-degree signals that a page is long. A high out-degree of long semantic distance links signals that a page covers a lot of unrelated topics. A high out-degree of short semantic distance links signals that a page has a lot of text that is focused on a particular topic. In the set of documents retrieved in response to the query, a document discussing a topic at length has a higher chance of being relevant than a document that only briefly discusses many different topics. *The semantic nature of outgoing link evidence gives us information on the scope of a document's content.*

The impact of filtering on the local degrees is shown in Figure 24. Here, random filtering has a bigger impact. The local link graph is already filtered on the search topic and has far fewer links. Further filtering flattens the degree distribution even more. For the in-degree,

random filtering still has little impact on MRR up to 80%, but after that, performance starts to drop as the local link graph has almost no links left. Further down the ranking (P@10 and MAP) the impact is bigger because most pages will have no incoming links and are indistinguishable. For the out-degrees, the effect is even stronger. For union, the impact on MRR is similar to that of the in-degree, and on P@10 and MAP more similar to that of the out-degrees.

If we remove the shortest distance links first, performance drops faster than with random filtering, while if we remove the longest distance links first, performance remains stable. The shortest semantic distance links are the more effective links. If we want to improve ad hoc search by exploiting link evidence, we need links between semantically related pages. Another important thing to note is that filtering on the category structure does not make local link evidence more effective. The query-dependent filtering method of zooming in on the highest ranked retrieval results already gets rid of most links between unrelated pages.

What happens to the performance of link evidence in combination with the content-based score when we filter links (Figure 25)? Randomly removing links has a similar impact as removing the longest distance links first on MRR and P@10, except for the in-degree, where the LD filter curves falls faster than the random filter curve. However, for MAP, performance remains stable when we remove the longest distance links up to the within-category links. The other filters hurt MAP. Again, the links between the most semantically related documents are the most effective. When we filter the shortest distance links, out-degrees become ineffective and start hurting performance when more than 50% of the links are removed. This is not the case for the in-degrees. Overall, filtering links does not lead to improvements in the impact of link evidence.

A general conclusion is that the impact of incoming link evidence is less sensitive to the density of the link graph than the impact of outgoing link evidence. This is partly explained by the difference at the higher end of the distribution. The in-degree distribution is more skewed and the high in-degree pages are more robust against filtering. Incoming link evidence is also less sensitive to the semantic nature of link evidence. Making the link graph more semantic has almost no impact on the effectiveness of incoming link evidence. Global outgoing link evidence becomes more related with relevance if we remove links between semantically unrelated pages. A page with many links to

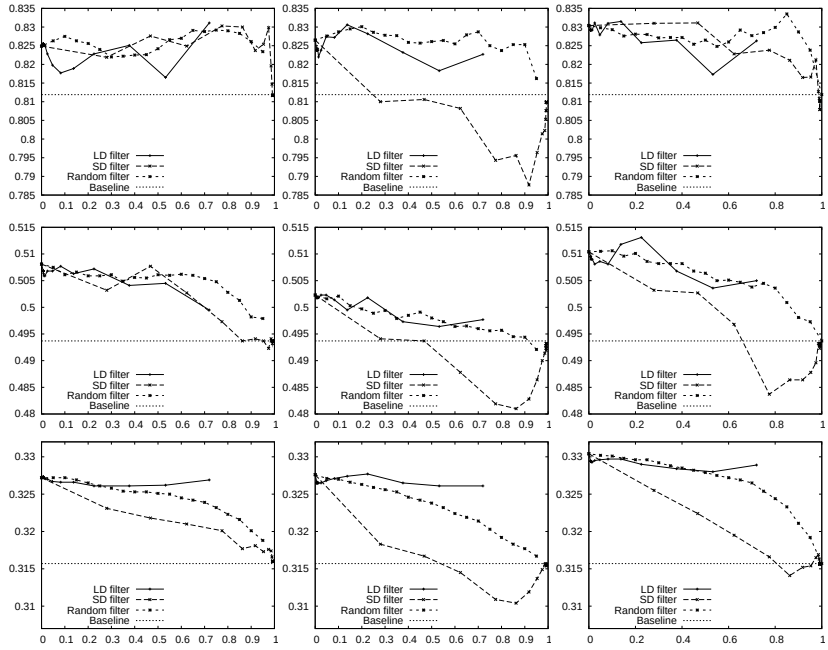

Figure 25: The impact of filtering links on the effectiveness of ranking on local link evidence and text evidence. The x-axis shows the percentage of links removed. The top row shows MRR, the middle row shows P@10 and the bottom row shows MAP.

		Within cat.			
		↑	=	↓	total
All	↑	130	0	20	150
	=	0	1	0	1
	↓	25	0	45	70
	total	155	1	65	221

Table 28: Per topic changes in AveP using all links set against per topic changes using within-category links based on the undirected degrees.

semantically related pages is an important page on its topic. This again shows the asymmetry of query-independent link evidence.

Local link evidence is not improved by removing long semantic distance links because it is already semantically filtered on the search topic, leaving very few long distance links to remove.

6.4.1 *Per topic analysis*

Averaged over the 221 topics, local link evidence using the within-category links seems to have a similar impact as local link evidence based on all links. But are there any differences for individual topics? We look at changes in average precision (AveP) with respect to the baseline, and compare the impact of using all links against the impact of using the within-category links. We would expect the within-category links to be less noisy and therefore more sensitive to the search topic. The changes in AveP in Table 28 are based on the undirected degrees. In the rows we see the number of topics for which AveP respectively deteriorates, stays the same or improves when using all links. In the columns we see the numbers when using the within-category links.

For the majority of the topics ($130 + 1 + 45 = 176$), the impact of link evidence on performance does not change direction by filtering out the cross-category links. For more than half of the topics (130 out of 221, or 59%), the impact of link evidence is positive (AvePgoes up), whether we use all links or only the within-category links. For 45 topics the impact of link evidence is negative with both sets of links and for 1 topic performance is the same with and without using link evidence.

When we filter the cross-category links, the impact of link evidence changes from positive to negative for 20 topics and from negative to positive for 25 topics. Overall, with filtering we increase the number

of topics for which performance improves and decrease the number of topics for which performance deteriorates. However, over all topics MAP

Next, we look at two topics where the filtered and unfiltered links have a very different impact. Topic 309 has title *"Ken Doherty" finals tournament* and topic 471 has title *Three greatest rivers +Japan* (and is also used for analysis in Section 3.4.1). The baseline system has an average precision of 0.6275 for topic 309 and 0.4167 for topic 471.

The unfiltered links have a positive impact on topic 309, with the union of the local in- and out-degrees improving the average precision from 0.6275 to 0.7249, while the filtered within-category degrees decrease average precision from 0.6275 to 0.5984. In Table 29 we see the 30 highest ranked articles according to the local union degree using the unfiltered links (left) and filtered within-category links (right). Relevant articles are shown in boldface. Without filtering, there are many relevant pages with a high local degree which are pushed up the ranking. In contrast, there are only 3 relevant pages with filtered local link evidence, and several pages with a higher degree. The number of local links drops drastically, from 160 to 28. The relevant pages have many links to related pages that are not in the same category, but which are useful for ranking. It seems the topic does not fit well within the category structure.

The article on *Ken Doherty* (a snooker player) does not share any categories with the other relevant pages. The article on *Snooker* does not share any categories with the articles on specific snooker tournaments such as *Masters (snooker)*, *Irish Masters (snooker)*, *Thailand Masters (snooker)* or *German Open (snooker)*, which all share the same category SNOOKER RANKING TOURNAMENTS. However, these articles on specific tournaments have no links to each other, apart from *World Snooker Championship* and *UK Championship (snooker)*. This shows that the search topic is less fine-grained than the category structure. The article on *Ken Doherty* and the articles on particular snooker tournaments are semantically related to each other, and the links between these articles are useful evidence for the topical relevance of these pages, but they are assigned to different categories. Although it explains why filtering out the cross-category links hurts performance, it also shows that the useful links are those connecting pages that are semantically related to each other. Using the top of the query-based ranking instead of the category structure focuses the local link graph at the right semantic level, so that the links between the snooker player and the tournaments he played in are retained while many irrelevant links are removed.

All links		Within category links	
Article	Degree	Article	Degree
Ken Doherty	23	Ken Jennings	3
Snooker	21	Jeopardy! Ultimate Tournament of Champions	3
Highest snooker break	14	Jerome Vered	3
Ireland	13	Jeopardy!	3
World Snooker Championship	10	Kazuma Kuwabara	2
Alan McManus	9	List of YuYu Hakusho episodes	2
Football World Cup	6	YuYu Hakusho	2
Sport in Ireland	6	Ken Shamrock	2
Masters (snooker)	6	Football World Cup	1
1998 in sports	6	Dan Severn	1
1997 in sports	5	Tenshinhan	1
Jimmy Connors	5	Leon White	1
2003 in sports	5	Highest snooker break	1
Jason's	4	European Under-21 Football Championship	1
Thailand Masters (snooker)	4	**World Snooker Championship**	1
Leon White	4	Double-elimination tournament	1
Grand Prix (snooker)	4	2005 NCAA Philippines Basketball Playoffs	1
Welsh Open (snooker)	4	Kuririn	1
Ken Shamrock	4	**Snooker**	1
Northern Ireland national football team	4	**UK Championship (snooker)**	1
Ding Junhui	4	Tournament	1
Steve James (snooker player)	4	2004 NCAA Philippines Basketball Playoffs	1
UK Championship (snooker)	4	Canadian Soccer Association	0
Jeopardy!	4	Paola Suárez	0
Ranelagh	4	June 2004 in sports	0
Ken Jennings	3	Royce Gracie	0
Irish Masters (snooker)	3	Playoff	0
Double-elimination tournament	3	World Universities Debating Championship	0
Strachan Open (snooker)	3	Tor Books	0
Players Championship (snooker)	3	Ranelagh	0

Table 29: Impact of link filtering on the local union degrees of topic 309 *"Ken Doherty" finals tournament*.

On topic 471, the unfiltered links have a negative impact, with the local union degrees decreasing the average precision from 0.4167 to 0.2875, while the filtered within-category degrees improve average precision from 0.4167 to 0.75. In Table 30 we see the 30 highest ranked articles according to the local union degree using the unfiltered links (left) and filtered within-category links (right). Without filtering, the articles *Japan*, *River* and *Honshū* (the largest Island of Japan) infiltrate in the top results because they are tangentially related to the topic and have many cross-category links with other pages in the local set. If we filter out those cross-category links, the union degree makes the topic more visible. The degrees of the three greatest infiltrating articles are greatly reduced, while the relevant pages only lose a few local links and move closer to the top. In combination with the text-based ranking, the articles *Rivers of Japan*, *Yoshino River* and *Tone River* are the 3 highest ranked results.

All 4 relevant pages (including which is not retrieved in the top 100) share the same category RIVERS OF JAPAN.

6.5 CONCLUSIONS

This chapter investigated the semantic nature of links, trying to answer which links are effective for retrieving relevant pages. Our first research question was:

- How can we measure the semantic relatedness between linked documents using the Wikipedia category structure?

The Wikipedia category structure is a mostly hierarchical semantic classification of the articles in Wikipedia. The open nature of Wikipedia has led to inconsistencies in the hierarchy, making complex information theoretical semantic relatedness algorithms inappropriate for our purposes. The simplest approach we used is distinguishing between links that connect documents that share at least one category and those that do not. We further computed the shortest path lengths between categories of linked documents using the hierarchical category structure.

Our second research question was:

- How is the link structure related to the categorical organisation in Wikipedia?

Compared to a random sample of document pairs, linked documents tend to be more semantically related to each other and more often share a category, showing a clear relation between global links and

All links		Within category links	
Article	Degree	Article	Degree
Japan	32	**Rivers of Japan**	5
River	17	River	5
Honshū	11	List of rivers of Asia	4
Rivers of Japan	8	List of rivers of Europe	3
Korea	7	List of rivers in China	2
Geography of Japan	6	Three Rivers Stadium	2
History of Japan	5	Sino-Japanese relations	2
List of waterways	5	**Yoshino River**	2
List of rivers of Asia	5	Heinz Field	2
Japan Self-Defense Forces	5	Pitt Stadium	2
National security of Japan	5	**Tone River**	2
List of rivers of Europe	4	Anti-Japanese sentiment	1
Chubu region	4	Greatest Hits (Queen)	1
Yoshino River	4	Honshū	1
Economic history of Japan	4	List of rivers of Iceland	1
Naval history of Japan	4	Economic relations of Japan	1
Japan-United States relations	4	Greatest Hits	1
Tone River	4	Korea	1
Occupied Japan	4	List of waterways	1
Flood	3	Geography of Japan	1
Japanese agriculture before WWII	3	Kuma River, Japan	1
Anti-Japanese sentiment	3	Japan	1
Weezer	3	Hvítárvatn	0
List of rivers in China	3	Districts of Bihar	0
Sino-Japanese relations	3	Three Rivers (district)	0
Kuma River, Japan	3	Jianzhen	0
Three Rivers Stadium	2	Lagarfljót	0
List of rivers of Iceland	2	Geography of Ecuador	0
Three Rivers (district)	2	Lake Maggiore	0
Pitt Stadium	2	Occupied Japan	0

Table 30: Impact of link filtering on the local union degrees of topic 471 *Three greatest rivers +Japan*.

semantic relatedness. However, within the top retrieved documents for a given query, the semantic signal of global link evidence is weaker than that of the textual evidence, providing an explanation why global link evidence is almost ineffective for topic relevance tasks. In the local set, pages that are linked tend to be more semantically related than pages that are not linked. Local link evidence is more clearly related to semantic relatedness and, even in the more topically focused set of top retrieved pages, links are a stronger signal that two pages are semantically related. This shows a difference in the semantic nature of global and local links. The query-independent link graph has many links between semantically unrelated pages. The semantic nature of link evidence changes as we zoom in on a subset of pages retrieved for a given query.

The finding that local links are more closely related to semantic relatedness and also more effective for retrieval brought us to our third research question, which was:

- Are links between semantically related pages more effective?

Removing the least semantic links (longest semantic distance links) has no negative impact on effectiveness and even improves performance of the global out-degrees. Removing the most semantic links (shortest semantic distance links) hurts performance of global and local link evidence, especially beyond the first few ranked documents. The effectiveness of global and local link evidence on the overall ranking is hurt more by random filtering than by removing the longest semantic distance links, and hurt most by removing the shortest distance links first. In other words, the effectiveness of link evidence is related to the semantic nature of links. Links between semantically related pages are more effective for ad hoc search than links between semantically unrelated pages.

The step from a global link graph to a local link graph works as a semantic link filter. Many of the links between semantically unrelated pages are removed. This is an essential step in making link evidence useful for ad hoc search. Our hypothesis that link evidence for topical relevance is symmetric hinges on the semantic relatedness of linked pages.

Finally, our main aim was to investigate:

- Which links are effective as evidence for topical relevance?

The effectiveness of link evidence is partly determined by the semantic nature of links. One prerequisite to make link evidence related

to topical relevance is that links connect documents that are semantically related. This explains why local link evidence is more effective than global link evidence. The documents in the local set are more semantically related to each other than documents in the global set. This operates as a semantic link filter, making the local link graph more "semantic" than the global link graph. On top of that, the local set is also filtered on the search topic and focuses on the right semantic level, making the local link graph also more semantically related to the search topic, which is another prerequisite to make link evidence related to topical relevance.

Part iv

Generalising to the Web

7

FROM WIKIPEDIA TO THE WEB

Now that we have analysed the nature of link evidence in Wikipedia, we increase our scope to the Web at large. In Chapter 4 we already looked at link evidence for Web-centric search tasks. For Web-centric tasks, global link evidence is very effective and needs no curbing, nor to be made sensitive to the topical context. In this chapter we look at ad hoc retrieval, and compare the impact of link evidence on the Web with our findings on Wikipedia for the same search task. In 2009, a new TREC Web Track was organised, using a large collection of Web pages called ClueWeb09 and a new set of topics and relevance judgements, which allow us to address the issue of Web ad hoc retrieval on a collection that is arguably more representative of the real Web than the WT2g and WT10g collections used for previous Web Ad Hoc tasks (see Table 2 on page 34). The links in Wikipedia are a special case of hyperlinks in the Web. Our findings about the value of link evidence might either be specific to the hyperlinks in Wikipedia or apply to hyperlinks in general. We used Wikipedia because its closed domain, dense link structure, categorical organisation and the availability of high-quality IR test collections allowed us a detailed analysis of the nature of link evidence. We found that local link evidence is more effective for ad hoc search than global link evidence and that for local link evidence, the direction of the links is not important. Do these findings also hold in general? In this chapter, we put our findings to the test on the larger and more general structure of links on the Web and address the question:

- To what extent does the value of link evidence in Wikipedia hold for link evidence in general?

In the course of writing this thesis, a new Web information retrieval test collection, called ClueWeb09, has been constructed based on a recent crawl of the Web with ad hoc search topics and relevance judgements. This new collection is meant to be a good representation of the first tier of the highest-quality pages in commercial search engine indexes and allow a better evaluation of Web search. There are two versions of the collection. ClueWeb09 Category A consists of 1 billion Web pages in various languages, the English pages making up roughly half of the collection. ClueWeb09 Category B contains a subset of 50

million pages in English, which were the first 50 million pages crawled. In this chapter, we use the Category B collection.

The new Web ad hoc retrieval test collection allows us to study the impact of link evidence on a more recent collection that should better resemble the collection of pages in the first tier of commercial search engine indexes. In the previous chapters we have established the value of link evidence for Wikipedia ad hoc search. We conduct experiments using link evidence for Web ad hoc search and compare the impact to the following findings in Wikipedia:

- *Global link evidence needs toning down, local does not*: The global degrees lead to heavy infiltration and can only improve mean average precision when we tone down their impact by using the log global degrees. The local degrees are more sensitive to the topical context and become less effective when toned down.
- *Local link evidence is more effective than global link evidence*: In Wikipedia, query-independent link evidence is less effective than query-dependent link evidence.
- *Local link evidence works in both directions, global link evidence does not*: In Chapter 4 we saw that log global in-degrees lead to a small but significant improvement in MAP while log global out-degrees have almost no impact on MAP. The local in- and out-degrees lead to very similar, significant improvements in MAP.

We then compare the new collection with older TREC Web collections in terms of the density and degree distribution of the link structure. At TREC 1999–2001, participants identified a number of aspects of hyperlinks that could affect the value of link evidence for retrieval, such as the number and density of links, and the relative importance of intra-server and inter-server links. Intra-server links are links between pages on the same server or Web site, while inter-server links are links from a page on one server to a page on a different server. To support meaningful experiments with hyperlinks and algorithms such as HITS and PageRank, test collections should have a large enough number of inter-server links (Bailey et al., 2003). As mentioned in Section 4.3.4.1, site-internal links are considered less useful as indicators of authority (Kleinberg, 1999) while Davison (2000) showed that linked pages on the same server are more similar to each other than linked pages on different servers. In this chapter, we address a number of more specific research questions:

- How has the Web changed in the decade between the TREC Web tracks of 1999–2001 and 2009?
- How is the effectiveness of link evidence affected by the density of the link graph?
- What is the relative impact of inter-server and intra-server links on the effectiveness of link evidence?

One important difference between the new collection and earlier TREC Web collections is that the new collection contains the full English Wikipedia, in which we now know that link evidence is effective. Differences observed between ClueWeb09 and earlier TREC Web collections might be due to the presence of Wikipedia. Therefore, another question we will address is:

- What is the impact of Wikipedia on the effectiveness of link evidence for Web ad hoc search?

This chapter is organised as follows. In Section 7.1 we describe the new Web collection and look at the relation between link degrees and the relevance of retrieval results. Then, in Section 7.2 we describe our experiments with using link evidence for the TREC 2009 Ad Hoc task, and in Section 7.3 we seek to understand the nature of link evidence in the new collection and address the research questions above in turn. We draw conclusions in Section 7.4.

7.1 THE CLUEWEB09 COLLECTION

The ClueWeb09 collection (CMU-LTI, 2009), consists of over a billion Web pages in several languages and contains some 25 terabytes of data. Because this amount of data might be too much to process for some research groups to participate in the TREC Web Track, a smaller, 10% subset of the collection is provided as an alternative. This subset, the ClueWeb09 category B collection consists of the first 45 million English Web pages of the crawl, and around 5 million pages representing the full English Wikipedia, which was crawled separately. In this chapter we use the category B collection, referred to as ClueWeb09 B.

The ClueWeb09 B collection was used for the TREC 2009 Web Track, which consisted of two tasks: the traditional Ad Hoc search task and the Diversity task. The ad hoc task is similar to the INEX Ad Hoc task on which the experiments in the previous chapters are based, and to the Ad Hoc task of the TREC 1999–2001 Web Tracks. The Diversity task

uses the same set of topics as the ad hoc task, but has a list of subtopics for each of those topics and challenges participants to develop retrieval techniques that return a list of diverse search results that cover relevant information for different aspects and interpretations of search queries.

7.1.1 *Relevance judgements*

Apart from the changes in the Web, any possible difference between the results in 1999–2001 and 2009 might be due to a difference in the setup. Although both have ad hoc search tasks, there might be differences in the numbers of pages judged, the specific judgement criteria (what counts towards the relevance of a page), and the types of queries and topics.

The 50 topics of the TREC 2009 Web Track were sampled from the query log of a real search engine (Clarke et al., 2009), with a preference for topics of medium popularity. Highly popular queries were assumed to be navigational and therefore less challenging, and rare queries were considered inappropriate because they may contain personally identifiable information.

For each of these topics, the relevant documents are identified by pooling the retrieved results of systems participating in the track and judging their relevance. Because of the large size of the ClueWeb09 collection, relevance judgements are necessarily incomplete. Two recently developed pooling strategies have been used to make evaluation over very large collections possible.

1. With the Minimal Test Collections (MTC) pooling strategy (Carterette et al., 2006), documents are selected for judgement that are most likely to determine the difference between two participating systems. The relevance judgements are meant to determine of any pair of participating systems which one is better than the other. Relevance judgements for documents ranked similarly by both systems are less useful, as they will not help identify which system is best. Therefore, documents are typically picked that are ranked very differently by the two systems. This pooling strategy is very sensitive to the particular systems contributing to the judgement phase and might be inappropriate for non-contributing systems that produce different rankings.

2. The second pooling strategy (Aslam et al., 2006) is based on the assumption that most relevant documents will be found in the top of the ranking, and samples documents from different parts (strata)

of the ranking using different sampling rates for each part. In the top 100 results the sampling rate is higher than in the results at ranks 100–1000. The sampling rate is then used to determine how many documents a pooled document represents. A document with a low sampling probability—meaning it was found far down the results list—represents more documents than a document with a high sampling probability.

Both pooling strategies use probabilities to determine how many documents a judged document represents, and both strategies skip many top ranked documents. Because we want to evaluate runs that did not contribute to the judgement pool, many of the top ranked results may not be judged. This might be problematic for our analysis of the impact of link evidence on reranking the top 100 results. The unknown relevance status of highly ranked documents might stop us from properly measuring qualitative differences between two results lists. To alleviate this possible problem, we can combine the Ad Hoc judgements of category B with the TREC 2009 Diversity relevance judgements. The Diversity judgements are based on the same topics and collection (Clarke et al., 2009), but on a different pooling strategy, namely, the traditional method of pooling the top n results of all officially submitted runs. In this case, the top 20 results of all runs were pooled and judged for relevance of a number of subtopics. Although the list of subtopics is not exhaustive, such that a document might be relevant for the overall topic but not for any of the subtopics, the extra judgements give us better information about the relevance of the top ranked results of our runs.

We evaluated all results with both the Ad Hoc relevance judgements and the combined relevance judgements, and found no qualitative differences. Systems doing better for one set of judgements tend to do better for the other set of judgements as well. This shows that the unjudged results introduce no bias in the evaluation. We use the category B documents and judgements in the experiments described below.

7.1.2 *Degree distribution*

We extracted over 1.5 billion links pointing to pages within the collection. Quite a large number of pages have multiple links to the same target URL (repeated links). If we collapse those repeated links and ignore self-referencing links (a link from a page to itself), we end up with 1.16 billion links between just over 50 million pages (see Table 31), which leads to a mean in-degree of 23.12. The median in-degree is 2,

Degree	#links	min.	max.	median	mean	stdev.
In-degree	1,161,178,732	0	5,958,953	2	23.12	2871.75
Out-degree	1,161,178,732	0	70,747	9	23.12	51.11

Table 31: Link degree statistics of the ClueWeb09 B collection.

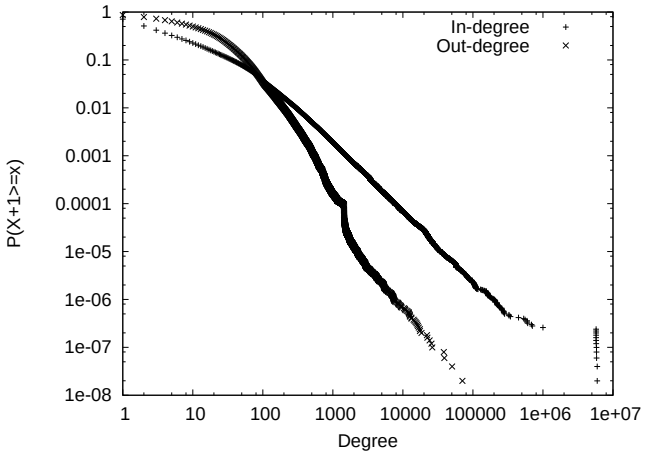

Figure 26: Complementary cumulative link degree distribution of the in- and out-degrees.

the median out-degree is 9. There are 35.7 million pages (71%) with at least one incoming link and 43.7 million pages (87%) with at least one outgoing link.

We only look at collection-internal links, that is, links from pages in the collection to other pages in the collection. Because the collection contains only a small part of the entire Web, many of the outgoing links of ClueWeb09 pages point to pages outside the collection. Therefore, the out-degree is lower than the actual out-degree of those pages. The same holds for the in-degree, as pages outside the collection may have links to pages in the collection as well. Wikipedia is a notable exception to this. Although it does have links to pages outside Wikipedia, the vast majority of its links are internal links to other Wikipedia pages. The English Wikipedia was crawled separately and is included in its entirety. The out-degrees of the Wikipedia pages will be much closer to their actual out-degrees.

The in- and out-degree distributions are shown in Figure 26. The straight vertical line at the high end of the in-degrees shows that a small

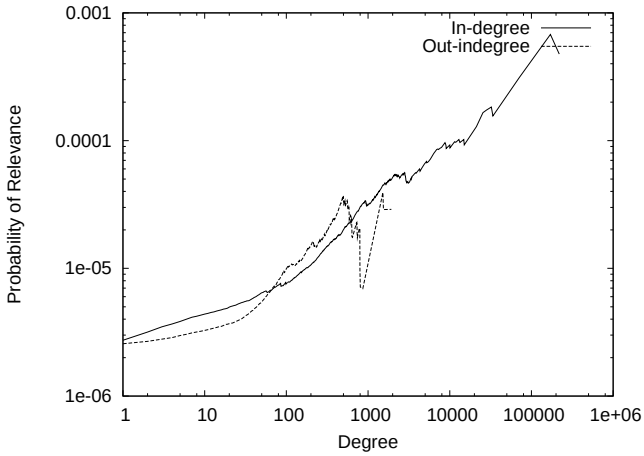

Figure 27: Prior complementary cumulative probability of relevance over in- and out-degrees.

number of pages have the same high in-degree. These are navigational and administrative pages in Wikipedia such as the Wikipedia main page, "Contact us", "Special:Random", "General_disclaimer" and portal pages. Every single Wikipedia page has a link to these pages, explaining why they have roughly the same in-degree.

Are the link degrees also related to the relevance of ad hoc retrieval results? The prior probability of relevance over in- and out-degrees is shown in Figure 27. Similar to the plots in Chapter 4, the in-degrees show a clear relation with relevance. The out-degrees also show a relation with relevance, although the probability does not increase all the way with out-degree. The maximum out-degree is 1,929, but the probability of relevance peaks at 500. Does this mean global link evidence can be used to improve ad hoc retrieval effectiveness in ClueWeb09 B? And is in-degree more effective than out-degree?

7.2 CLUEWEB EXPERIMENTS

We used Indri (Indri, 2009) for indexing. Stopwords are removed and all other terms are stemmed with the Krovetz stemmer. The main index is a standard full-text index.

For the full-text runs we again use a language model approach and linear smoothing. However, ad hoc search in large collections requires little smoothing (Kamps, 2006a). Therefore, we adjust the weight of the

smoothing parameter to $\lambda = 0.85$. That is, in the language modelling formula, the document model $P(q|d)$ has weight $\lambda = 0.85$ and the background model $P(q|D)$ has weight $1 - \lambda = 0.15$. We experimented with different smoothing values and found that on the ClueWeb09 B collection, the specific value for λ has little impact on the results. There are many documents matching all the words in the short Web queries used for the TREC 2009 Web Track, so the smoothing value has little effect in the top ranked documents. In Chapter 4 we saw that document length is not an effective document prior for Web-centric search tasks, but for ad hoc search, longer documents have a higher prior probability of relevance. The ClueWeb09 collection contains only Web data, so we first compare baseline runs with and without a document length prior. Recall that when document length is used as a prior, documents are scored using the document length as a prior probability $p_{\text{length}}(d) = \frac{|d|}{\sum_{d' \in D} |d'|}$, where d and d' are documents in collection D.

There are two sets of effectiveness measures used for the TREC 2009 Web Track: the expected precision measures based on the MTC pooling strategy and the statAP (Yilmaz and Aslam, 2006) measures based on the stratified sampling strategy (see Section 7.1.1). We use statAP, as it is more robust than the MTC-based measures when evaluating systems that did not officially participate in the Web Track and therefore did not contribute to the pool. We test for significant changes with respect to the full-text baseline using a one-tailed bootstrap test with 100,000 resamples.

The first two lines in Table 32 show the baseline results with ($\beta = 1$) and without ($\beta = 0$) length prior. As with ad hoc search on other collections, document length is related to relevance (Singhal et al., 1996). We conduct our link-based experiments on the $\beta = 1$ baseline. Recall that the union of in- and out-degree is the degree derived from an undirected version of the link graph. That is, bidirectional links count as one link. One notable difference with earlier Web tracks is that the full-text baseline on ClueWeb09 has a low early precision. Is this caused by spam or by the much larger size of the Web, with many low quality pages? Several Web Track participants found that spam filters and page quality scores improved performance (Cormack et al., 2010, Hauff and Hiemstra, 2010, Lin et al., 2010, Rajput et al., 2010), suggesting that standard text retrieval approaches indeed suffer from spam and low-quality pages in ClueWeb09.

We first look at the impact of the global degrees as true document priors over all retrieved results (rows 3–5). This leads to heavy infilt-

Reranked	ID	statMAP	MPC(30)	MAP	MRR	P@10
	baseline $\beta = 0$	0.0991	0.2208	0.0932	0.2982	0.2180
	baseline $\beta = 1$	0.1442	0.3079	0.1516	0.3061	0.2780
All results	Global in	0.0182	0.2632	0.0109	0.1688	0.0640
	Global out	0.0489	0.1293	0.0386	0.2270	0.1480
	Global union	0.0381	0.1382	0.0222	0.1836	0.0720
Top 100	Global in	0.1417	0.4488	0.1347	0.5008	0.2280
	Global out	0.1512	0.4771	0.1545	0.4347	0.3360
	Global union	0.1511	0.4865	0.1541	0.5010	0.3280
	Log Global in	0.1449	0.3344	0.1423	0.5179	0.2380
	Log Global out	0.1563	0.4689	0.1683	0.5040	0.3660
	Log Global union	0.1552	0.3786	0.1653	0.5217	0.3440
	Local in	0.1514	0.3704	0.1520	0.6616	0.2940
	Local out	0.1562	0.4114	0.1628	0.5550	0.3880
	Local union	0.1575	0.4210	0.1623	0.6395	0.3540
	Log Local in	0.1497	0.3836	0.1542	0.5329	0.2980
	Log Local out	0.1561	0.4049	0.1660	0.4724	0.3660
	Log Local union	0.1565	0.4202	0.1658	0.5183	0.3620

Table 32: Results for the 2009 Ad Hoc Task. Significance tests are with respect to the full text baseline, confidence levels are 0.95 (°), 0.99 (°) and 0.999 (•).

ration and is disastrous for early and overall precision, with statMAP going from 0.1442 to 0.0182, 0.489 and 0.0381 using the in-degrees, out-degrees and their union, respectively. If we limit the prior to the top 100 retrieved results (rows 6–8), we see improvements in early precision with in- and out-degrees and their union (MRR goes up from 0.3061 to 0.5008, 0.4347 and 0.5010 respectively), and improvements in overall precision with the out- and union degrees (statMAP goes up to 0.1512 and 0.1511 respectively). This is in direct contrast to the impact of global link evidence on Web-centric search tasks (see Chapter 4, Section 4.3.4), where global link evidence is best used as a true prior probability over all results. Ad hoc search requires more careful use of link evidence. The global in-degree prior over the top 100 results gives an improvement in MRR and MPC(30), but is not effective for average precision: MAP and statMAP go down. The low early precision of the baseline result and the improvements of the global degrees suggest that the text-based ranking has many low-quality documents at the top and the role of link evidence is to identify and push up the important pages. This fits with the findings of the TREC 2009 Web Track participants. The global out-degree priors (row 7) are more effective than the in-degree priors (row 6), and show improvement on all reported measures. The log global degree priors are even more effective (rows 9–11), and, apart from statMAP, the log global out-degrees are the most effective. Clearly, the number of outgoing links is a good signal for the relevance of a page. But why are pages with many outgoing links more often relevant? In Chapter 4 we saw that in Wikipedia the maximum out-degree is higher than in the .GOV collection (see Table 8 on page 75). Perhaps this is also the case in ClueWeb09 B, and the global out-degree favours Wikipedia pages, which might be higher quality documents, and therefore more often judged relevant. We discuss this in Section 7.3.4.

If we look at the local degrees (rows 12–14), we see that especially the in-degrees are effective for improving MRR. The local out-degrees are more effective for all other measures and thus for later and overall precision, consistent with our findings in Chapter 4. We note that this collection includes the full English Wikipedia, which might partly explain this consistency. The log local degrees (bottom 3 rows) only slightly improve traditional MAP but not the other measures. The local degrees need no toning down.

How do these results compare to the impact of link evidence in the INEX 2006 Wikipedia collection?

Global link evidence needs toning down, local does not: The normal, non-log global degree priors hurt performance only when used over all retrieved

results but lead to improvements when used over the top 100 results. In Wikipedia, the non-log global degree priors hurt performance both when applied to all results and when applied to only the top 100 results. Global link evidence is more effective on the Web than on Wikipedia. As in Wikipedia, using the log of the degrees improves the effectiveness of the global degrees but not of the local degrees. In Wikipedia, the normal local degrees were clearly better than the log local degrees. Here the difference is smaller.

Local link evidence is more effective than global link evidence: Local link evidence is more effective than global link evidence for statMAP, but not for MPC(30). The difference between the impact of local and global link evidence is also smaller. Link evidence for document importance seems more useful in the Web than in Wikipedia, which is possibly due to the large variation in document quality in the Web.

Local link evidence works in both directions, global link evidence does not: In Wikipedia, local in-degree has the same impact as local out-degree. Our hypothesis is that local links provide evidence for topical relevance, which is a symmetric relation, so should have the same impact in both link directions. In contrast to Wikipedia, there is a large difference between the impact of local incoming and outgoing link evidence in the Web. In contrast to the findings in Chapter 4, the local outgoing link degrees are more effective than the incoming link degrees in the ClueWeb09 B collection. This suggests that in the ClueWeb09 B collection, local link evidence signals more than just topical relevance. Perhaps the local out-degree is so effective because it is related to the global out-degree and also signals document importance. Again, this might be because the out-degree favours Wikipedia pages.

Why do the in- and out-degrees behave the same way in Wikipedia but differently in the ClueWeb09 collection? And why is link evidence effective for ad hoc search in ClueWeb09 where it is not in the WT10G collection? Perhaps link evidence is more effective than in previous TREC experiments because this collection contains the full Wikipedia, where we have seen that link evidence is effective for ad hoc search. Wikipedia pages are edited by many contributors, so the quality might be higher than that of many Web pages. Or perhaps the higher link density makes link evidence more effective. In the next section, we look at several factors of the ClueWeb09 link structure that play a role in the nature of link evidence.

7.3 WHY LINK EVIDENCE WORKS IN CLUEWEB09

Why is link evidence effective for ad hoc search in the ClueWeb09 collection, but not in older collections (see Table 2 in Section 2.3.3)? And why is global out-degree more effective than global in-degree and the local degrees, while our findings in the previous chapters would suggest otherwise? Several aspects have been mentioned as possible factors determining the effectiveness of link evidence.

Differences in the collection: The new collection is bigger, more recent, crawled using a different strategy. All these aspects could affect the value of link evidence in a collection.

The impact of link density: Gurrin and Smeaton (2004) have mentioned collection size and (inter-server) link density. Fisher and Everson (2003) also mention the link density of collections as a crucial factor to make links useful. In the previous chapter we have seen that link evidence can have a positive impact even with relatively small set of links.

The impact of inter-server links: Bailey et al. (2003) describe the construction of the WT10G test collection with the aim of having a larger inter-server link density. What is the relative impact of inter- and intra-server links?

The impact of Wikipedia: Another notable difference between the WT10G and ClueWeb09 collections is the presence of the full English Wikipedia in ClueWeb09. In the previous chapters we have seen that link evidence is effective in Wikipedia, so the positive impact of link evidence on the ClueWeb09 collection might be due to the presence of Wikipedia.

We will discuss each of these aspects in detail.

7.3.1 *Differences in the collection*

We compare the TREC 2009 Web Track with the earlier Web Tracks of 1999–2001 in terms of the collection, the relevance judgements and the link graph. The new ClueWeb09 collection is different from earlier TREC Web collections in several ways:

Collection size: The ClueWeb09 B collection (50 million pages) is much larger than the WT10G (1.7 million pages) and .GOV (1.2 million pages) collections and twice the size of the .GOV2 collection (25 million pages).

Size of Web: The Web has grown from an estimated 320 million pages in late 1997 (Lawrence and Giles, 1999) to tens or hundreds of billions

of pages. According to WorldWideWebSize (2010), in June 2010 the indexed Web contains over 23 billion pages and 120 million active Web sites.

Page quality: The ClueWeb09 collection was planned to reflect the first tier of highest quality Web pages in commercial search engine indexes (Callan et al., 2008), and was crawled in early 2009 (CMU-LTI, 2009) using a seed set of the highest PageRank pages from an earlier crawl (Fetterly et al., 2009b) and a crawling policy that schedules the most important pages to be crawled first (Abiteboul et al., 2003, Fetterly et al., 2009a). The value of link evidence might change with the quality of the pages. The INEX 2006 Wikipedia collection contains the full English Wikipedia, which arguably has little spam, and high-quality pages and links edited by many contributors.

Average link degree: The OPIC crawling policy (Abiteboul et al., 2003) determines page importance by counting incoming links. As a consequence, such a crawl leads to a densely interlinked collection of Web pages. As we can see in Table 1 on page 33, the average incoming and outgoing link degree of pages in the ClueWeb09 B collection is higher than in the earlier Web collections. With more links there is more evidence.

Age: The TREC Web Tracks of 1999-2000 used the WT2G, WT10G and VLC2 collections which were based on a truncated crawl of the Web of February 1997 (Hawking and Craswell, 2005), only four years after CERN declared the World Wide Web was freely available to anyone (W3C, 2010). The pages in the ClueWeb09 collection were crawled in early 2009, 12 years later. The World Wide Web has changed a lot since then. It has grown immensely in the meantime, as have commercial interests. The link graph is more complete and in some sense more stable.

Access: Users access many Web sites via search engines nowadays whereas in the early days access through hyperlinks was more common. Web site authors strive to get their Web pages as high in the search results ranking as possible to draw Web traffic. A lot of effort is put into search engine optimisation (SEO) so that site entry pages are placed high in search results rankings for particular queries, with particular attention to site-internal links structure and anchor text. Companies have analysed what makes sites and pages end up high in search results list. One particular strategy is to have pages in the site link to each other with high quality descriptive anchor text. Whereas in 1997 many authors used terms such as “here”, “next” and “home” as anchor text

to allow users to easily navigate within the site, focus has shifted to using anchor text terms that describe the content of a page so that a user typing the same terms as a query can find that particular page.

Spam: Because of the growing commercial interest and much larger number of users, Web spam has become an ever growing problem, with many different forms. The taxonomy of Web spam (Gyöngyi and Garcia-Molina, 2005) identifies two broad classes: text spam and link spam. Much like filtering out low-quality pages, text-based spam can be combatted with page importance measures such as PageRank, HITS and Online Page Importance Computation (OPIC, Abiteboul et al. (2003)).

Graph evolution: In Section 4.2.2 we looked at the phases of development of the INEX Wikipedia and .GOV collection, and observed that both collections are in the final phase, where almost all the pages are connected to the giant component. The numbers in Table 33 show that the .GOV, INEX Wikipedia and ClueWeb09 B collections have reached the final phase of the evolution of link graphs. The WT10G and .GOV2 collections—which were used for the earlier TREC Ad Hoc Tracks—have not reached this phase yet. In the previous chapter we saw that the links in Wikipedia are still effective when we randomly remove most of the links, which suggests that even in a relatively undeveloped link graph, link evidence is still effective. But perhaps randomly filtered fully evolved link graphs are different from link graphs in earlier evolutionary phases. Perhaps ad hoc search requires a fully developed link graph that is crawled to focus on the most important pages, for links to be effective.

Differences might also be found in the types of queries used and the relevance criteria given to relevance judges. Above we noted that the queries for the TREC 2009 Web track are mid-frequency queries from a search engine query log. If these queries are more general than those of earlier tracks, the number of relevant pages for these queries might also be higher.

We compare the relevance judgements of the TREC Web Tracks of 2000–2001 and 2009 in Table 34. The 2000–2001 Ad Hoc topics have around 70,000 judgements (1400 per topic) and around 3000 relevant pages (60 per topic). In contrast, the 2009 Ad Hoc topics have only 13,118 judgements (262 per topic) but 4002 relevant pages (80 per topic).[1] This could mean that the 2000–2001 topics are more specific. On the other hand, theory suggests that precision at a fixed cut-off (and therefore the number of relevant pages pooled from the participating

1 Because of the sampled pooling strategy used for the TREC 2009 Web Track, the 4002 judged relevant pages represent 25,036 estimated relevant pages in the entire collection.

	Collection				
	TREC 1999 WT10g	TREC 2002 .GOV	TREC 2004 .GOV2	INEX 2006 Wikipedia	TREC 2009 ClueWeb
# Docs	1,692,096	1,247,753	25,205,179	659,388	50,220,423
# Links	8,062,918	11,110,989	82,711,345	13,584,225	1,180,631,904
Thresh.					
1	831,848	612,286	12,516,624	322,118	24,974,092
2	846,048	623,877	12,602,590	329,694	25,110,212
3	860,248	635,467	12,688,555	337,270	25,246,331
4	6,066,790	4,378,632	107,390,194	2,208,796	222,626,286
5	12,133,579	8,757,264	214,780,388	4,417,592	445,252,571

Table 33: Threshold, levels and phase transitions for the TREC Web collections.

Edition	*Topics*	*Coll. size*	*# Judgements*	*# Relevant*	$\frac{\#\ Relevant}{\#\ Judged}$
TREC 9 (2000)	50	1.7M	70,070	2617	0.0373
TREC 2001	50	1.7M	70,400	3363	0.0478
TREC 2009	50	50M	13,118	4002	0.3051

Table 34: Comparison of the topics and relevance judgements of TRECs 2000–2001 and 2010.

systems) increases with collection size because the number of easy-to-find relevant documents increases (Hawking and Robertson, 2003). This might the reason that document importance is more effective here than in Wikipedia. As Kleinberg (1999) argued, the problem for broad search topics is not finding the relevant pages—which are abundant—but identifying the authoritative pages.

In very large collections with many relevant documents, the local graph might become sparse if too small a number of pages is chosen for the local set. With billions of Web pages, a larger local set might be required to derive meaningful evidence for topical relevance.

7.3.2 *The impact of link density*

What is the impact of link density? One of the hypotheses at TREC was that link evidence was not effective for ad hoc search because the

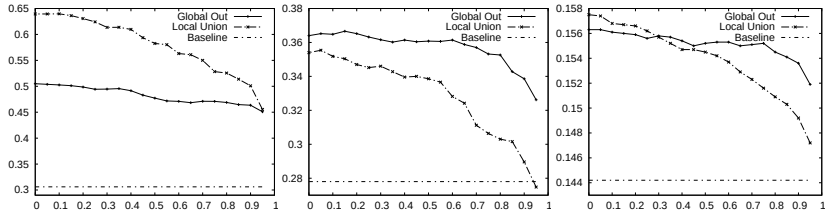

Figure 28: The impact of randomly filtering links on the effectiveness of link evidence for MRR (left), P@10 (middle) and statMAP (right).

number of (inter-server) links was too low. If there are no links at all, there is no impact of link evidence. With a complete graph (all pages link to all other pages), all pages have the same amount of evidence and link evidence has no impact. For link evidence to be effective, the link density must be somewhere in between. But is the optimal density closer to zero links or to the maximum number of links? As in the previous chapter, we randomly filter out links, this time to study the impact of link density.

Is the difference in effectiveness due to the density of the link graph? In fact, the link density of the ClueWeb09 B collection is lower than that of the WT10g collection. Link density is the fraction of all possible links that is actually present in the graph (Wasserman and Faust, 1994). With a collection of 1.7 million documents and 8 million links (see Table 1), the link density of the WT10g graph (treating all links as undirected) is $5.6 \cdot 10^{-6}$. The link density of the ClueWeb09 B collection with 50 million pages and 1.16 billion links, is $9.2 \cdot 10^{-7}$. However, the number of links per page, or average link degree, of the ClueWeb09 B collection is higher than that of the WT10g collection. Pages in the ClueWeb09 B collection have an average in-degree of 23.12, whereas pages in the WT10g collection have an average in-degree of 4.77.

The impact of filtering links on the effectiveness of link evidence is shown in Figure 28. Note that below 21% of the links, the average degree of the ClueWeb09 B collection is lower than that of the WT10g collection (The average degree in the WT10g collection, 4.77, is 21% of the average degree of the ClueWeb09 collection, 23.12). So, if we remove 80% of the links in the ClueWeb09 B collection, the average degree is roughly the same as that of the WT10g. Does link evidence become ineffective in ClueWeb09 if we randomly remove 80% of the links?

We show the impact of random filtering on the most effective global and local degrees, namely the log global out-degrees and the local undirected degrees. The impact on the other degrees is similar. The impact of global link evidence on MRR remains stable as we randomly remove links, up to the last 5% of the links. Link density has almost no impact on the effectiveness of global link evidence for MRR. The impact on the local degree is bigger, as the MRR slowly drops as we remove links and has the same impact as the global out-degrees after 95% of the links are removed. Link evidence is effective for MRR even at the lowest density.

For P@10, we see that the log global out-degrees remain effective when filtering links, and the impact is stable with up to 60% of the links removed. Beyond 60%, the improvement slowly drops but performance remains well above that of the baseline. The local union of in- and out-degrees is clearly affected by filtering. The improvement over the baseline steadily drops to zero as we filter out more links. The local graph is sparse to start with, so at 5% there are very few links left, changing very little in the ranking.

For statMAP we see a similar pattern as for P@10, although the local union of in- and out-degrees is more effective than the log global out-degrees without filtering. However, filtering has a bigger impact on local link evidence than on global link evidence, and as a consequence, local link evidence becomes less effective than global link evidence when more than 30% of the links are removed. The global link graph is very rich and many links can be removed before the structure falls apart and becomes meaningless. The local link graph is sparse and, with most links removed, carries almost no information to aid the ranking of retrieval results. However, even a very sparse local link graph can be enough to improve overall precision.

In sum, link density plays almost no part in the impact of global link evidence on the entire collection. Even a small set of links is enough to distinguish the important pages from the rest. In contrast, link density plays an important part in the impact of local link evidence.

Local link evidence becomes less effective on all three measures as we randomly remove links. The impact is bigger for measures that take into account a larger part of the ranking, like statMAP. If the global link graph is sparse, there is not enough local link information to have any impact beyond the first ranked documents. However, even the smallest samples contain enough links to improve retrieval performance with local link evidence.

Can link density explain why link information was not useful for Web ad hoc search in the WT10g collection? With 5% of the links the average degree is just above one, whereas in the WT10g collection where links were not effective, the average degree is almost 5. Link density alone cannot fully explain why links are effective in the ClueWeb09 B collection, but might be part of the explanation.

7.3.3 *The impact of inter-server links*

Another hypothesis brought forward at TREC 8 for the disappointing results of using hyperlinks for retrieval is that the WT2g collection has very few inter-server or site-external links. Links across sites are considered to be more meaningful than links between pages on the same site. The reasoning behind this is twofold. First, similar to the distinction between incoming and outgoing links, it is harder for a Web author to increase the number of links to her own page(s) from other sites than from pages under her control. Second, links within a site often serve a purely navigational purpose, such as links to the entry page and links in a navigation bar that allow users to quickly jump to the part of the site they are interested in. Perhaps the inter-server link density is high enough in the ClueWeb09 collection and is the determining factor in the effectiveness of link evidence. On the other hand, the internal links in Wikipedia are also intra-server or site-internal links, and have been proven meaningful in the previous chapters. Assuming that a single Web site is created and modified by a single author or group of cooperating authors, these authors have full control over the link structure on a global level, just as in Wikipedia. In this section, we want to find out:

- What is the relative impact of inter-server and intra-server links on the effectiveness of link evidence?

There are many more intra-server or site-internal links than inter-server or site-external links in ClueWeb09 B. There are 952 million intra-server links (88% of the total) and 132 million inter-server links (12%). If intra-server (site-internal) links are indeed less meaningful, the findings of the previous chapter suggest that the positive impact of link evidence mainly comes from inter-server (site-external) links. We test this by comparing in Table 35 the impact of using link evidence from only inter-server links versus using only intra-server links. We show only the log global degrees as they give the best improvements. The impact of inter-server versus intra-server links is largely the same

Links	ID	statMAP	MPC(30)	MAP	MRR	P@10
	Baseline	0.1442	0.3079	0.1516	0.3061	0.2780
Intra	Log Global Out	0.1566	0.4818	0.1659	0.5114	0.3680
	Log Global Union	0.1556	0.4701	0.1631	0.5129	0.3580
Inter	Log Global Out	0.1488	0.3336	0.1573	0.4333	0.3460
	Log Global Union	0.1469	0.3201	0.1547	0.3942	0.3280

Table 35: The impact of inter- and intra-server link evidence on retrieval effectiveness in ClueWeb09.

on global and local degrees. Although the inter-server links improve performance, the intra-server links lead to larger improvements. Of course, the intra-server links are larger in number, but these results show that the inter-server links are not the main contributors to the effectiveness of link evidence. In fact, the impact of the intra-server links is very similar to the impact of using both inter- and intra-server links.

The inter-server link structure is much more sparse and has less information to distinguish between pages. Site-entry pages are typically the pages with the most incoming inter-server links (Hawking et al., 2004), while most other pages within the same site have no incoming links from other sites. The inter-server links cover only a small part of the collection and mainly the entry pages. When searching information within a single (enterprise) Web site, Hawking et al. (2004) found that site-internal link evidence is effective for improving retrieval performance while site-external link evidence has almost no impact. Ad hoc search is closer to the enterprise search task of searching for information within a corporate Web site than to Web-centric tasks such as home page finding. In home page finding tasks, only the entry pages, which tend to have more incoming (inter-server) links than other pages, can be relevant. In ad hoc and enterprise search, any type of page can be relevant as long as it contains relevant information.

The bias towards intra-server links introduced by the fact that Web site authors can control both the incoming and outgoing links of a page has no negative impact on the value of intra-server links for ad hoc search. This bias is more troublesome for measuring authority and popularity.

7.3.4 *The impact of Wikipedia*

The English Wikipedia forms a substantial part of the ClueWeb09 B collection. With over 5.7 million pages, it takes up 11% of the collection, and could be the main reason for the effectiveness of link evidence.

To test this hypothesis, we indexed the ClueWeb09 B collection without Wikipedia, removed all links from and to Wikipedia from the link graph and re-ran our experiments. This version of Wikipedia is more recent than the version on which the INEX 2006 collection is based, and contains many more pages—although the crawl contains some duplicate pages and redirects. The average degree has also increased. The 5.7 million pages have 446 million links (ignoring repeated links and self-referencing links). All Wikipedia pages have the same navigational bar with links to the main page and a number of other pages, which are not present in the INEX 2006 Wikipedia collection. But even without those links, Wikipedia forms a very densely interlinked part of the Web. There are a further 21 million links from Wikipedia to other pages in Clueweb09 B and 1.5 million links to Wikipedia from the rest of ClueWeb09 B, which we also exclude. The non-Wikipedia part of the Web has 615 million links between 45 million pages. What is the impact of link evidence on ad hoc search in the non-Wikipedia part of the Web?

The results are shown in Table 36. Link degree priors are applied to the top 100 results of the $\beta = 1$ baseline. Performance drops when we remove Wikipedia from the collection. Wikipedia is a high-quality Web site with good informational pages that are often ranked high and contain little to no spam. If we remove them, statMAP drops considerably, from 0.1442 to 0.1044.

However, link evidence still improves the non-Wikipedia baseline. Without Wikipedia, the global degrees (rows 3–5) are effective for early precision—MRR goes up from 0.2814 to 0.3566, 0.3596 and 0.3537 for, respectively, the global in-degrees, out-degrees and their union—but not for statMAP and MAP. Although MPC(30) is improved, the traditional P@10 drops. The log global degrees (rows 6–8) are more effective than the normal global degrees. The local degrees (rows 9–11) improve on all reported measures. The log local degrees (rows 12–14) further improve normal MAP and P@10, but lead to a lower MRR than the normal local degrees. The presence of Wikipedia cannot explain the positive impact of link evidence on Web ad hoc search.

But these results reveal another interesting factor contributing to the impact of Wikipedia. With Wikipedia removed, local link evidence is

ID	statMAP	MPC(30)	MAP	MRR	P@10
Non–wiki baseline $\beta = 0$	0.0880	0.2181	0.0802	0.2784	0.2160
Non–wiki baseline $\beta = 1$	0.1044	0.2528	0.1015	0.2814	0.2260
Non–Wiki Global In	0.1008	0.3692	0.0883	0.3566	0.1720
Non–Wiki Global Out	0.1030	0.3552	0.0877	0.3596	0.1680
Non–Wiki Global Union	0.1012	0.3594	0.0882	0.3537	0.1760
Non–Wiki Log Global In	0.1043	0.2699	0.0951	0.4207	0.1880
Non–Wiki Log Global Out	0.1080	0.3812	0.1019	0.3961	0.2540
Non–Wiki Log Global Union	0.1072	0.2980	0.1014	0.4174	0.2260
Non–Wiki Local In	0.1114	0.3131	0.1035	0.4705	0.2440
Non–Wiki Local Out	0.1115	0.3121	0.1053	0.4667	0.2700
Non–Wiki Local Union	0.1132	0.3312	0.1070	0.4827	0.2700
Non–Wiki Log Local In	0.1102	0.3185	0.1059	0.4291	0.2580
Non–Wiki Log Local Out	0.1101	0.3250	0.1073	0.3961	0.2760
Non–Wiki Log Local Union	0.1108	0.3270	0.1080	0.4253	0.2720

Table 36: Impact of link evidence on the non-Wikipedia part of ClueWeb09 B.

more effective than global link evidence, and the difference between incoming and outgoing link evidence has almost disappeared. In-degree is better for very early precision while out-degree is better for later and overall precision, just as in the INEX 2006 Wikipedia collection. The union of the two degrees is even more effective. We compare this again with our findings in previous chapters.

Global link evidence needs toning down, local does not: The normal, non-log global degrees hurt MAP when used over the top 100 results, just as in Wikipedia, and using the log of the degrees improves the effectiveness of the global degrees. The local degrees become slightly more effective for MAP but not for statMAP. The log helps improve precision at a fixed rank cut-off but not MRR. In Wikipedia, the normal local degrees were clearly better than the log local degrees. In the Web, curbing the degrees is beneficial for performance.

Local link evidence is more effective than global link evidence: Without Wikipedia, local link evidence is more effective for Web ad hoc search than global link evidence, but global link evidence is still effective when applied with care. Document importance is useful for Web ad hoc search, but query-dependent evidence is more useful. Is local link evidence

more effective because it is a toned down version of the global degrees or because it is more related to topical relevance?

Local link evidence works in both directions, global link evidence does not: Local in-degrees are better for MRR while local out-degrees are better for P@10 and MPC(30). For MAP the differences are small. The union of the in- and out-degrees further improves performance. This is almost exactly the same as for the INEX Wikipedia collection. For overall precision, the impact of link evidence is symmetric, suggesting that local links signal topical relevance.

Without Wikipedia, link evidence in the Web behaves similar to link evidence in Wikipedia. Only when we combine Wikipedia with the rest of the Web do the global degrees become effective and the out-degrees more so than the in-degrees. What is the impact of Wikipedia on the global in- and out-degrees?

In the $\beta = 1$ baseline, there are 15 Wikipedia pages in the top 100 on average. The baseline run has 1.2 Wikipedia pages in the top 10. The local degrees and the global in-degrees push up Wikipedia pages, all with roughly 2.2 Wikipedia results in the top 10. The global out-degrees and union degrees have 4.2 Wikipedia pages in the top 10. The global out-degrees favour Wikipedia pages more than other degrees do. For the non-Wikipedia results in the top 100, the median global in-degree is 3 and the median global out-degree is 9. Among the Wikipedia results in the top 100, the median global in-degree is 0 and the median global out-degree is 147. The global out-degree works as a Wikipedia filter much more than the other degrees.

This can explain why the global out-degrees are effective. Wikipedia is densely linked and most pages have a large amount of outgoing links. The global out-degree seems to push up Wikipedia pages in the ranking. Wikipedia pages are often considered high-quality pages with informational text (a pre-requisite for relevance in the ad hoc methodology) and are on average longer than the non-Wikipedia pages—Wikipedia pages are on average 7944 characters long while non-Wikipedia pages are 4635 characters long. The impact of the length prior shows that document length is related to relevance. This would mean that Wikipedia pages have a higher probability of being relevant than non-Wikipedia pages, and performance is improved by favouring Wikipedia pages high in the ranking. The same observation was made by He et al. (2010), who reranked the search results by pushing all Wikipedia results to the top of the ranking.

We ran the same queries on a Wikipedia-only index of the ClueWeb09 B collection (see Table 37) and found that on the Wikipedia-only index,

ID	statMAP	MPC(30)	MAP	MRR	P@10
Wiki baseline $\beta = 0$	0.0483	0.1946	0.0584	0.2869	0.1920
Wiki baseline $\beta = 1$	0.0748	0.2433	0.0832	0.4295	0.3340
Wiki Global In	0.0608	0.2021	0.0485	0.3189	0.1520
Wiki Global Out	0.0627	0.2274	0.0590	0.2704	0.1860
Wiki Global Union	0.0623	0.2329	0.0571	0.3175	0.1900
Wiki Log Global In	0.0676	0.2382	0.0619	0.4454	0.2160
Wiki Log Global Out	0.0742	0.2610	0.0800	0.4111	0.3040
Wiki Log Global Union	0.0751	0.2496	0.0809	0.4400	0.3100
Wiki Local In	0.0694	0.2228	0.0645	0.4923	0.1860
Wiki Local Out	0.0765	0.2683	0.0847	0.4061	0.3420
Wiki Local Union	0.0769	0.2704	0.0787	0.4874	0.2520
Wiki Log Local In	0.0724	0.2520	0.0719	0.5091	0.2380
Wiki Log Local Out	0.0775	0.2905	0.0862	0.4512	0.3360
Wiki Log Local Union	0.0784	0.2674	0.0873	0.5015	0.3340

Table 37: Impact of link evidence on the Wikipedia part of ClueWeb09.

early precision is higher than on the whole ClueWeb09 B index, even though Wikipedia is a 10% subset. This is in direct contrast with the observation by Hawking and Robertson (2003) that precision at a fixed rank cut-off tends to increase with collection size. It is easier to find relevant pages in Wikipedia than in the rest of the ClueWeb09 B collection, suggesting that Wikipedia pages are are of higher quality than non-Wikipedia pages.

The Wikipedia pages form 11% of the ClueWeb09 B collection. The judged Wikipedia pages form 18% of the judged pages. This means Wikipedia pages have a higher probability of being retrieved in the top results. The relevant Wikipedia pages form 21% of the relevant pages, meaning Wikipedia pages have a higher probability of being relevant.

That the log global degrees still lead to improvements indicates that in the Web, part of the role of link evidence is to identify important, high-quality documents. The global out-degrees lead to a larger improvement than the global in-degrees, which might be a document length effect.

It seems that the special nature of Wikipedia creates a bias in the ClueWeb09 B collection which muddles the analysis of the impact of link evidence. Below, we look at the impact of link density and inter-server links in ClueWeb09 B with and without Wikipedia.

Links	ID	statMAP	MPC(30)	MAP	MRR	P@10
	Non–Wiki baseline	0.1022	0.2570	0.1011	0.2816	0.2260
Intra	Non–Wiki Local union	0.1119	0.3305	0.1058	0.4705	0.2033
	Non–Wiki Log Local union	0.1087	0.3201	0.1068	0.4163	0.2247
Inter	Non–Wiki Local union	0.1007	0.2593	0.0994	0.3361	0.2267
	Non–Wiki Log Local union	0.1017	0.2603	0.1005	0.3098	0.2287

Table 38: The impact of inter- and intra-server link evidence on retrieval effectiveness in the non-Wikipedia part of ClueWeb09 B.

7.3.4.1 *Inter-server and intra-server links*

We have seen that the Wikipedia part of ClueWeb09 B accounts for 446 million of the intra-server links and 22 million of the inter-server links (20.5 million links from Wikipedia to external pages and 1.5 million links from external pages to Wikipedia). With 5.7 million pages, it is by far the largest Web site in the collection. The next biggest Web site (in number of pages) has only 34,684 pages. The number of pages on the same site determines the maximum intra-server link degree. In a site with n pages, the maximum intra-server in- and out-degree is $n - 1$. Wikipedia pages can have much larger degrees than non-Wikipedia pages. Note that in ClueWeb09, Wikipedia has been crawled separately, which has an important impact on the composition of the collection. The crawling policy for the rest of the collection was to crawl new domains first and limit the number of pages per domain and the depth at which pages were crawled within each domain. Wikipedia is the only exception. In a normal crawl using the same policy, Wikipedia would form a much smaller part of the crawl and have a much sparser link graph. The change in the impact of link evidence on the ClueWeb09 B collection when we include Wikipedia might thus be an artefact of the way the collection is constructed. Because Wikipedia accounts for such a large part of the intra-server links and is a single giant Web site that accounts for 24% of the relevant documents, we look at the impact of inter- and intra-server links without Wikipedia in Table 38.

We show only the impact on the local and log local union degrees, which give the best performance. The impact is similar on all other degrees. Apart from the P@10 measure, intra-server link evidence outperforms inter-server link evidence on all measures. The presence of Wikipedia does not tip the balance in favour of the intra-server links. Even without Wikipedia, intra-server links are more effective than inter-

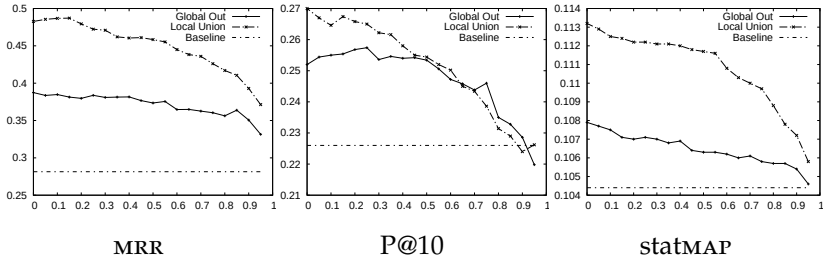

Figure 29: The impact of randomly filtering links on the effectiveness of link evidence in the non-Wikipedia part of ClueWeb09 B.

server links. The explanation given earlier is still valid: inter-server links cover a smaller number of pages than intra-server links and mainly point to site-entry pages. The bias of control over incoming links has no impact on the value of link evidence for topical relevance.

7.3.4.2 *Link density*

We repeat the link filtering method on the non-Wikipedia part of the collection and see the impact of filtering on effectiveness in Figure 29. The most striking difference from the impact on the whole collection is that, on the non-Wikipedia part, filtering has a larger impact on the effectiveness of the global degrees. Whereas on the whole collection the global degrees were barely affected by filtering, in the non-Wikipedia part the global degrees gradually lose their impact as we filter more links.

The small impact of filtering on the full collection can be explained by the fact that we used a random filter to reduce link density. With random filtering, all pages are affected in the same way. If page A has a higher link degree than page B in the full link graph, then A will also have a higher link degree on average in a randomly sampled link graph. In other words, the Wikipedia pages are promoted in the full collection regardless of the amount of links filtered. Highly connected pages are more robust against filtering links than other pages. The local graph is more fragile, so filtering has a larger impact on the local degrees.

7.4 CONCLUSIONS

In this chapter we looked at the impact of link evidence on ad hoc retrieval in the new ClueWeb09 B collection.

- What is the impact of link evidence on the ranking of Web ad hoc retrieval results?

Link evidence can significantly improve ad hoc retrieval effectiveness in the Web, when restricted to the top 100 results. This is in direct contrast to the findings of the TREC Web Tracks of 1999–2001. The main differences between the 1999–2001 evaluation on the WT10G collection and the evaluation in this chapter on the ClueWeb09 B collection is that TREC 2009 topics are more general, the collection is much larger and more densely linked and contains the full English Wikipedia.

- How is the effectiveness of link evidence affected by the density of the link graph?

We randomly filtered links from the graph to study the impact of link density. The effectiveness of the global degrees is hardly affected by randomly filtering links while the effectiveness of local link evidence gradually decreases. The global graph is richer and therefore more robust against filtering. The top of the degree-based ranking changes little by random filtering. The local graph is much sparser and the ranking is more sensitive to the presence of specific links. However, even with a small amount of links, both local and global link evidence can improve a standard full-text retrieval baseline.

- What is the relative impact of inter-server and intra-server links on the effectiveness of link evidence?

Intra-server links make up a large part of the link graph. The inter-server link graph is much more sparse and covers only a small number of pages. If inter-server links in ClueWeb09 B mainly point to entry pages, as they do in the dataset used by Hawking et al. (2004), it would make them less suitable for ad hoc search, where any type of page can be relevant. Intra-server links cover a much larger part of the collection, including many pages deep within sites. The impact of link evidence on ad hoc search mainly comes from intra-server links.

The main difference between inter-server and intra-server links is that Web authors have more control over the intra-server links. That is, assuming that a single Web site is authored by a single author or a cooperating group of authors, authors control both the incoming and outgoing links within their own Web site, just as in Wikipedia. Intra-server links are more similar to Wikipedia links and show the same impact on link evidence. There is little qualitative difference between incoming and outgoing site-internal links, making in-degrees

and out-degrees equally effective. The bias of control over the incoming intra-server links does not affect the relation between link evidence and topical relevance.

- What is the impact of Wikipedia on the effectiveness of link evidence for Web ad hoc search?

In Wikipedia the link structure is much denser than in the rest of the Web, and Wikipedia pages tend to have higher global out-degrees than other Web pages. Although Wikipedia might be different in nature from the rest of the Web, the higher density is also partly due to the crawling policy restricting the number of pages to be crawled from a single Web site—except for Wikipedia—and perhaps also to the fact that only a limited part of the Web has been crawled. There might be links to the pages in the collection from Web pages that have not been crawled. As a consequence, the global out-degree pushes Wikipedia up the ranking. Because Wikipedia pages have a higher probability of relevance, the global out-degree is effective for improving Web ad hoc search.

Without Wikipedia, link evidence in the Web behaves similar to link evidence in Wikipedia, lending support to the findings of previous chapters. Local links are more effective than global links, and incoming and outgoing link evidence have a similar impact on overall performance. This suggests that in the Web, local link degrees provide evidence for the topical relevance of search results. The fact that global degrees are still effective when toned down by taking the log of the degrees shows that, with the presence of many low-quality pages, document importance is useful for Web ad hoc search.

With Wikipedia, the impact of link evidence changes radically. Global out-degrees are very effective because they favour longer articles and especially Wikipedia articles, which have a higher prior probability of relevance. The Wikipedia link graph is very different from most other site-internal link graphs. Wikipedia is a large domain with millions of relatively long articles that are densely interlinked, whereas the next biggest Web site in ClueWeb09 B has only 25 thousand pages. The impact of Wikipedia might be smaller in a larger crawl of the Web, where there is no restriction on the number of pages per Web site.

- To what extent does the value of link evidence in Wikipedia hold for link evidence in general?

The presence of Wikipedia affects the value of link evidence. Without it, our hypotheses mostly hold. The fact that local link evidence is more

effective than global link evidence supports our hypothesis that local link evidence signals topical relevance and that topical relevance is useful for ad hoc search. That incoming and outgoing links have a similar impact on overall performance supports our hypothesis that the evidence for topical relevance is symmetric. The positive impact of global link evidence shows that document importance is also useful for ad hoc, at least on the Web.

However, Wikipedia is part of the World Wide Web, and general aspects of hyperlinks should hold in both the entire Web and in Wikipedia. Any aspect of hyperlinks where Wikipedia differs from the whole Web cannot be a general aspect. How should we interpret the impact of Wikipedia on the value of hyperlinks for retrieval? Perhaps the impact of Wikipedia is not so much that it changes the nature of links, but the nature of informational search on the Web. Wikipedia forms a special part of the Web that is important for informational search.

Part v

Conclusions

8

CONCLUSIONS

From a practical perspective, the value of link evidence has been established through the success of using PageRank, in-degree or anchor text and the findings in the Web-centric tasks of the TREC Web Tracks. The main goal of this thesis is clarifying and thereby providing a more thorough understanding of the relation between link evidence and the notion of relevance in information retrieval.

The history of scientific benchmarking for Web IR is plagued with the apparent contradiction between the experiences of Internet search engines, and the results of experiments at the TREC Web Tracks of 1999–2001 (Hawking and Craswell, 2005). This led to Google's Larry Page calling the entire formal evaluation process "irrelevant" during a heated panel debate at the 2000 Infonortics Search Engine Meeting (Sherman, 2000). After several years of disappointing results at TREC Web Tracks, it was surmised that Web structure is simply not effective for ad hoc search tasks. TREC moved on to Web-centric tasks, where link topology, anchor text, and URL structure were proven very effective for navigational search, such as site finding and home page finding.

For a very detailed analysis of the value of link evidence, we focused on the INEX Wikipedia Ad hoc test collection, which has a dense link structure that is not truncated by crawling limitations, a complex category structure and specific information on the location of relevant text in documents. The ClueWeb09 collection, which approximates the Web as indexed by Internet Search Engines closer than earlier collections, allowed us validate our findings in Wikipedia on a large Web collection. This prompted us to revisit the standing question of the value of link evidence for information retrieval.

In the first chapter we introduced a number of research questions and in the other chapters we stated a number of hypotheses. We will first address the research questions based on the findings of the previous chapters, then discuss the hypotheses and finally discuss avenues for future work.

8.1 RESEARCH QUESTIONS

In Chapter 1 we listed four sets of research questions that guided the work of this thesis. We will discuss each of them in turn.

1. Links for Wikipedia and Web retrieval: We decided to look at value of link evidence in the INEX 2006 Wikipedia collection, becomes it comes with a high-quality test collection with very detailed relevance judgements, a non-truncated link graph and a complex category structure, which together allow us a very detailed analysis of the impact of link evidence on retrieval performance. Because links in Wikipedia might differ from general Web hyperlinks in certain characteristics, their impact on retrieval might be different. Our first set of research questions prompted us establish the importance of link evidence in Wikipedia ad hoc retrieval and whether it differs from its role in Web search.

- Can link information in Wikipedia be used as evidence to improve the ranking of ad hoc retrieval results?

Finding: Link evidence can be used to improve ad hoc retrieval effectiveness if we derive it from the feedback of a text-based retrieval system. This finding is in direct contrast with the disappointing results of using link evidence for the Ad Hoc search task of the TREC Web Tracks of 1999–2001. This suggests that the value of link evidence Wikipedia is different than its value in the Web.

Evidence: In Chapter 3 we saw that incoming link degrees in Wikipedia are related to the relevance of retrieval results. Experiments on the INEX 2006 Wikipedia collection showed that global, query-independent link degrees can improve ad hoc retrieval performance if used carefully, although the improvements are small. Local, query-dependent link degrees are made sensitive to the search topic and are more effective for ad hoc search. The local graph is sparser and leads to a less extreme degree distribution which needs no toning down.

- Is the value of links in Wikipedia different from their value in the Web?

Finding: The value of links for retrieval in Wikipedia and the Web depends to a large extent on the nature of the search task. For navigational search, link evidence is best derived from the global incoming link structure while for informational search, link evidence is best derived from the local graph of incoming and outgoing links based on the feedback of a text retrieval system. However, the heterogeneity of the

Web puts more emphasis on the role of link evidence as an indicator of the authority or quality of documents, while the homogeneity of Wikipedia puts more emphasis on the relation between link evidence and topical relevance.

Wikipedia is a single homogeneous Web site containing articles written in a similar style and for a similar purpose, that is, to be informative. The Web is a vast collection of Web sites which are highly heterogeneous.

In the Web it is important to distinguish between Web-centric search tasks, such as home page finding, that are navigational in nature, and the informational search tasks as modelled by the ad hoc task. Locating home pages is a typical task in the Web, but makes no sense in the single domain of Wikipedia. Finding the best encyclopedic entry for a given search topic in Wikipedia resembles navigational search in the Web. Arguably, the difference between navigational and informational search is less relevant in Wikipedia.

Wikipedia content is edited by millions of contributors. Arguably, all Wikipedia pages have similar authority, such that the challenge for retrieval is to locate topically relevant information. Within the Web, pages and sites vary considerably in quality, reliability and popularity. For Web-centric search tasks, these aspects pose an additional challenge for retrieval.

Popularity and authority are derived mainly from site-external links, because Web page authors typically have no control over incoming links from other Web sites. The number of site-external incoming links is a measure of popularity and authority, which is derived from the global link graph.

Within a Web site, authors often have control over the site-internal links and thus over both the incoming and outgoing site-internal links of a page. In terms of conferring authority, site-internal links are less meaningful than site-external links. Wikipedia being a single domain within the larger Web, the links between the encyclopedic articles are also site-internal links.

Evidence: We saw in Chapter 4 that in Wikipedia, outgoing link degrees are structurally similar to incoming link degrees.

In Chapter 4 we looked at the impact of link evidence for Web-centric search tasks and observed a big difference in the value of link evidence for ad hoc search in Wikipedia and navigational search in the Web. For Web-centric search tasks, link evidence is effective when derived from the global link structure, which reflects the importance of individual pages, and identifies the right "type" of pages.

For ad hoc search in Wikipedia (Chapter 3) and the Web (Chapter 7), link evidence is effective when it is made sensitive to the topical context. Within the local set, incoming and outgoing links are equally effective. In the Web, global link evidence is more useful than in Wikipedia. This is probably related to the fact that Wikipedia is single homogeneous domain with high-quality pages, while the quality of pages in the Web varies strongly, which makes evidence of page importance more useful.

We should also make a distinction between the Web as represented by the collections used for the TREC Web Tracks of 1999–2001 and the Web represented by the collection used for the 2009 Web Track. Whether the more recent collection is a better representation of the Web, or whether it reflects a Web structure that has changed with respect to that of the earlier collections, the impact of link evidence is different. In the more recent collection, link evidence helps improve ad hoc retrieval performance, whereas in the 1999-2001 experiments on the older collections it did not. In Section 7.3.1 we discussed a number of differences such as the more general topics used in 2009, the higher density of the link graph and structural changes due to insights in search engine optimisation.

2. Global and local link evidence: Link information can be derived from the entire link graph of the collection, or from a subset of query-dependent retrieval results.

- How is global, query-independent link evidence related to relevance?

Finding: Global link degrees are related to query-independent aspects of documents. Insofar as global link evidence signals document importance, the direction of the link evidence determines what the evidence is an indicator of. The in-degree is an indicator of how well-known or popular a Web page is and how general or common the topic of a Wikipedia article is. The out-degree is an indicator of document length in Wikipedia. These aspects are related to query-independent aspects of relevance. By definition, global link evidence is not related to topical relevance. It cannot be used to determine whether a document is on the requested search topic. Within a set of relevant pages, it can promote pages with more relevant information, but it has no way to keep focus. Promoted pages often have more relevant information because they have more information in total, including more irrelevant information. Therefore, we conjectured that global link evidence is related to document importance but not to topical relevance.

For search tasks that demand high-quality, popular Web pages, global link evidence is best used unrestricted to clearly distinguish the import-

ant pages from the rest. For search tasks focussing on topical relevance, global link evidence needs to be curbed so as not to disrupt the content-based relevance ranking too much.

Evidence: In Chapter 3 we saw that global link evidence, if used in a logarithmic scale, can improve ad hoc retrieval performance in Wikipedia, although the improvement is very small. In Chapter 4 we saw that global incoming link evidence is more effective than global outgoing link evidence.

For Web-centric search tasks on the .GOV collection, it is best to use global link evidence on a non-logarithmic scale. Both incoming and outgoing link evidence are effective, but incoming link evidence is more effective than outgoing link evidence. In the Web, global incoming link evidence identifies home pages and other important Web pages. Global outgoing link evidence identifies long pages in Wikipedia but is also useful for identifying entry pages in the Web, as shown in Chapter 4.

In Chapter 5 we saw that global link evidence provides a better-than-random ordering if we use incoming link evidence.

In Chapter 7 we saw that for Web ad hoc search, global link evidence is effective, especially when used on a logarithmic scale.

- How is local, query-dependent link evidence related to relevance?

Finding: Local, query-dependent link evidence is made sensitive to the search topic and is more subtle in changing the ranking of search results. Local link degrees are necessarily related to the global link structure and therefore reflect the importance of documents to a certain extent. The local degree distribution also has a similar shape as the global degree distribution. However, local link evidence is related to topical specificity as well and keeps reasonable focus on the search topic. Insofar as local link degrees are related to document importance, the direction of the links determines the value of the evidence. Incoming links signal authority or popularity, which is one of the key motivations behind the HITS and SALSA algorithms. Outgoing links weakly signal document length as well as perhaps accessibility; a page with many links to other, relevant search results is a good entry point for accessing information on the search topic. Insofar as local link degrees are related to topical relevance, the direction of the links is irrelevant. The link provides the same semantic relatedness information about both the page from which it originates and the page it points to. Semantic relatedness is a symmetric relation. Link evidence for topical relevance—which is by necessity query-dependent—works in both directions. In Wikipedia, the amount and fraction of local link evidence are related to the notions

of exhaustivity (the extent to which the search topic is discussed) and specificity (the extent to which the document is focused on the topic) of relevance.

Local link evidence can improve the relevance ranking in two ways. First, it can be used to distinguish relevant from non-relevant documents. Second, within the set of retrieved relevant documents, it can be used to rank documents more favourably with respect to the amount of relevant text they contain, while retaining topical focus.

Evidence: In Chapter 5, we found that, in Wikipedia, local link degrees are moderately correlated with global degrees, showing that local link evidence can still reflect query-independent aspects of relevance. We can make local link evidence less dependent on the global degrees by looking at the fraction of global links present in the local graph (the local fraction) or down-weighting the local degree by the log of the global degree (the weighted degree).

In Section 5.3 we saw that local link evidence in Wikipedia is symmetric in the sense that incoming and outgoing link evidence have the same impact on average precision and in Section 5.4 that they have a similar relation with the amount and fraction of relevant text in documents. In the same section, we saw that the amount of local link evidence is related to the amount of relevant text and the fraction of local link evidence is related to the fraction of relevant text, reflecting the notions of exhaustivity and specificity. As we increasingly focus the link evidence on the local context, pages with less irrelevant text are promoted.

Restricting the link graph to links between the top retrieved documents for a given query acts as a filter in two ways. First, the local links connect documents that are more semantically related to each other than documents in the local set that are not linked (Chapter 6). Second, the documents are filtered on the search topic, making link evidence more focused on the search topic.

In the Web, local link evidence is effective for both ad hoc search tasks and Web-centric search tasks. For ad hoc search local link evidence is more effective than global link evidence when we leave out Wikipedia (Section 7.3.4), and local incoming and outgoing link evidence are equally effective. With Wikipedia included, local link evidence is less effective than global outgoing link evidence because global outgoing link evidence works as a Wikipedia prior. For Web-centric search, local link evidence is less effective than global link evidence (Section 4.3.4) and incoming link evidence is more effective than outgoing link evidence.

3. Importance and topical relevance: Links can be used as indicators of popularity or importance of documents, or as indicators of the topical relevance of linked documents.

- Is link evidence of document importance useful as evidence for ranking ad hoc retrieval results?

Finding: Document importance is a broad term covering different aspects, which we here associate with query-independent aspects of relevance. Importance can be related to the amount of information in a document, or the popularity or authority of a document or its author. All these aspects can be useful for relevance. Long documents have a higher probability of being relevant for ad hoc search because they have a lot of text, any part of which can be relevant to some information need. Popular documents have been found worthwhile to link to by many different people. This is a signal that this document is a desirable document to return as search result when it matches the query. Authoritative documents receive many votes of confidence from different sources, indicating that is has reliable and useful content.

Evidence: Document length priors are very effective for ad hoc search, as we have seen for Wikipedia in Chapter 4 and for the Web in Chapter 7. In Wikipedia, pages vary little in quality and authority, making global link evidence only marginally useful, and only when used with care. In the Web, where quality and authority of pages and sites varies much more, indicators of popularity and authority are more effective.

- Is link evidence of topical relevance useful as evidence for ranking ad hoc retrieval results?

Finding: Our conjecture is that links can provide evidence of topical relevance when derived from the feedback of a text-based retrieval model. Because it is based on the ranking of a text-based retrieval model, it is not obvious that it is complementary to textual evidence. Although we cannot measure the relation between link evidence and topical relevance directly, there are certain factors of its impact that we expect to observe.

If we have content-based evidence that a certain document is relevant for a given search topic, than we would expect another document with similar content to be relevant as well. Insofar as links are evidence that linked documents are similar in content, they are a signal that the content-based evidence of one document is also evidence for the documents it is linked to. The content similarity relation is a symmetric relation, which should mean that this signal is independent of the

direction of the link. If we have content-based evidence for the relevance of two linked documents, the link between them reinforces the content-based evidence of both documents. Therefore, we would expect link evidence for topical relevance to work in both the incoming and outgoing direction.

This symmetry should have consequences for non-local links as well. Links between a retrieved document and a non-retrieved document provide evidence that the non-retrieved document is relevant, but also that the retrieved document is not relevant.

Another factor which we would expect to observe is that the amount of link evidence for topical relevance is related to the extent to which a document is relevant. A small amount of link evidence provides little support for the relevance of a document and suggests a document is only marginally relevant. More link evidence for topical relevance signals that a document is more topically relevant.

Evidence: To find evidence for these expected observations, we used the detailed relevance assessments of the INEX Wikipedia Ad Hoc test collection, which allow a detailed analysis of the relation between link evidence and relevance. We saw symmetry in the impact of link evidence in Wikipedia in Chapter 5 where incoming and outgoing local link evidence led to similar average precision (page 119) and showed a similar relation with the amount and fraction of relevant text in relevant pages (page 122). Although their impact on performance is symmetric, local incoming and outgoing link evidence do promote different documents. The shape of their distributions is somewhat different, which is reflected in differences in their precision curve (see Table 24 on page 119). Combining incoming and outgoing links leads to more evidence and better performance. With more evidence, links are better able to distinguish between documents. The local fractions of incoming and outgoing links showed an even stronger symmetry. The symmetry increases as we make link evidence more sensitive to the topical context.

Links between local documents and non-local documents provide contrasting evidence for topical relevance, which might explain why expanding the set of retrieved documents with documents that are linked to them is not effective for ad hoc search in Wikipedia (see Section 4.3.7 on page 98). The lack of content-based evidence for the non-retrieved document signals that the link lacks evidence for topical relevance.

In Section 5.4 we observed that the local degrees are related to the amount of relevant text in pages. If we look at the internal ranking

of relevant documents according to the local degrees, the amount of relevant text decreases with increasing rank, while the fraction of relevant text remains stable. We also saw that the fraction of global links present in the local set is related to the fraction of relevant text in documents. Relevant documents with a large fraction of global links in the local set have a large fraction of relevant text. This supports our conjecture that link evidence based on feedback from a content-based retrieval model is related to topical relevance.

4. Quantity and semantics: the information conveyed by links is affected by the quantity of links and the semantic relatedness of linked documents.

- What is the impact of link density or link quantity on the value of link evidence?

Finding: Link density has little impact on global link evidence. Even a small amount of links can make link evidence effective for ad hoc search. In networks where preferential attachment is one of the guiding principles, the most important (popular, authoritative) pages build up connections quicker than less important pages. This pattern will quickly distinguish the important pages even in sparse graphs. However, for local link evidence, a denser global graph will lead to denser local graphs. With very few links in the global graph, a small subset of nodes will have an almost completely unconnected subgraph. Increasing link density makes local link evidence more effective as long as the links connect semantically related pages and the density is not so high that all local pages have the same amount of link evidence.

Evidence: In Chapter 6 we saw that the impact of global link evidence in Wikipedia remains stable as we randomly filter out links while the impact of local link evidence gradually decreases as we remove more links from the graphs. The global link structure is very rich and random filtering does not affect the order of the high degree pages much. Pages with many links are robust against random filtering. The local link graph is much sparser, and removing links reduces the ability of local link evidence to distinguish between relevant and non-relevant pages. The impact of filtering is more visible for evaluation measures that use a larger part of the ranking.

In Section 7.3.2 we saw a similar impact of filtering links in the Web. With Wikipedia included, randomly filtering links affects the impact of local link evidence more than the impact of global link evidence. However, even when we remove 90% of the links in the ClueWeb09 B collection, performance is improved by link evidence. When we leave

Wikipedia out of the collection, in Section 7.3.4.2, we saw that the impact of both global and local link evidence gradually drops to zero as more links are removed. The impact of local link evidence is still affected more than the impact of global link evidence.

- How does the semantic relatedness of linked documents affect the value of link evidence?

Finding: The semantic relatedness of linked documents determines the effectiveness of link evidence for topical relevance in Wikipedia. The more semantically related the linked documents are, the more effective the link. This supports ideas behind relevance propagation. If a document matches a search query well, this is evidence that the document is relevant for the information need of the user. In turn, this is evidence that semantically similar documents are also relevant. Links that connect semantically similar documents are useful links that are supported by the content of connected pages.

Global incoming link evidence is unaffected by the semantic relatedness of linked documents. The semantic nature of global outgoing link evidence to some extent reflects the topical scope of a document. In Wikipedia, documents with many outgoing links to topically unrelated pages are long documents discussing many unrelated topics, while documents with many outgoing links to topically related pages are long documents focused on a particular topic, which have a higher probability of being relevant.

Evidence: In Section 6.4 we saw that the positive impact of link evidence in Wikipedia ad hoc retrieval drops quickly as we remove the links with the shortest category distance first. When we remove the longest category distance links first, performance remains stable.

In Section 6.4 on page 141, we saw that the performance of ranking the top 100 documents by global in-degree remains stable whether we filter links randomly or based on category distance. Global outgoing link evidence becomes more effective when we remove links between semantically unrelated documents.

8.1.1 *Main research question*

We have analysed how the direction and semantic nature of links and the nature of the link graph from which link evidence is derived affect the meaning signalled by links. Our aim was to get Our main research question was:

- What is the value of link evidence for information retrieval?

The value of link information for information retrieval depends on a number of things: the nature of the search task, the semantic nature and the direction of the links, and the nature of the document set from which the link information is derived.

In a large collection of documents, the links between those documents provide us with evidence about the popularity, authority or length of documents, which are query-independent aspects related to the relevance of documents. At the collection level, the direction of the link determines for which aspect it provides evidence. Hyperlinks are anchored in pieces of text in documents and thus take up space. Outgoing links therefore provide evidence about document length. as well as connectedness and "hubness". Those hyperlinks are pointed at other pages in the collection, but require the author of the source document to know about the documents targeted by hyperlinks and to put in effort to create those links, therefore the incoming links say something about the popularity and/or authority of those targeted pages. Because link evidence derived at this level is blind to the topic of request, it provides no information about the topical relevance of documents.

When we zoom in on a subset of documents retrieved in response to a given query and retain only the links between these retrieved documents, the nature of link evidence changes. The signals about document length and popularity diminish, and are joined by a signal that the linked documents are related to other documents retrieved for the same query, providing evidence for the topical relevance of the linked documents. This signal forms evidence for both linked documents in equal measure, therefore is independent of the direction of the link. The strength of this evidence is determined by the semantic relation of the linked documents. The evidence is strong when the linked documents have semantically similar content and weak when the linked documents have semantically unrelated content. The evidence for document length and popularity are not affected by the semantic relatedness of the linked documents.

For ad hoc search, inter-server links are not more valuable than intra-server links. The main distinction between inter-server and intra-server links is the control over a page's incoming links and affects the flow of authority and popularity. It has no notable impact on the relation between link evidence and topical relevance. Although document importance indicators such as authority and popularity are useful for ad hoc search, link evidence as an indicator of topical relevance is more useful. The relation between link evidence and topical

relevance is established by filtering links on the search topic, which greatly reduces the number of links. The greater quantity of intra-server links makes them more useful than inter-server links for ad hoc search.

Perhaps the main contribution of this thesis is that it solves the apparent contradiction between the experiences of Internet search engines, and the results of experiments at TREC. Negative results for ad hoc informational search using Web structure have tainted the reputation of reproducible IR evaluation. The positive results in this thesis may help to set the record straight. This turns the earlier negative results into something positive in a sense: they aid to our understanding of when and why link evidence works, and when not.

8.2 HYPOTHESES

Throughout this thesis, we introduced a number of hypotheses and conjectures. We will discuss the status of each with respect to the conclusions above.

- **In Wikipedia, incoming and outgoing links are similar to each other (Chapter 4).** In Chapter 4 we argued that in contrast to the Web in general, authors contributing to Wikipedia have control over both the incoming and outgoing links of a page. The conferral of authority from source to targets disappears, making incoming and outgoing link evidence equally important. In Chapter 7 we saw that on the Web, site-internal links behave very much like the links in Wikipedia. This is not surprising, given Wikipedia is itself a single Web site and that the links between its articles are site-internal links. The control over incoming and outgoing site-internal links is more general than Wikipedia alone. However, the encyclopedic organisation of Wikipedia, where each topic has its dedicated article, makes it clear which related articles to link to. On top of that, the content and links of Wikipedia are edited by millions of contributors, resulting in a high-quality link graph where relevant links are preserved, missing links are added, and irrelevant and redundant links are removed. As a consequence, the Wikipedia link graph might be more complete in terms of connections between related pages.

- **Links in Wikipedia are similar to links in the World Wide Web (Chapter 4).** Both in Wikipedia and the Web, the link structures show power law distributions, indicating that both are scale-free networks which grow according to the principle of preferential attachment—possibly in combination with other principles such as (dis)assortative

mixing. For both Wikipedia and the Web, we saw that the shape of the degree distributions is the same on global and local levels. That is, within a set of documents retrieved for a given query, the degree distribution has the same shape as over the whole collection, which was also observed by Chakrabarti et al. (2002). In Chapter 7 we discovered that the site-internal links in the non-Wikipedia part of the Web have a similar impact on retrieval as the site-internal links in Wikipedia.

- **Global link evidence is query-independent and is related to document importance, but not to topical relevance (Chapter 5).** Global evidence is by definition blind to the topic of the query, therefore cannot be related to topical relevance. Because relevance is a broad notion with many different aspects, global evidence can be related to other aspect of relevance such as document length, authority or popularity.

- **Local link evidence is query-dependent and related to topical relevance (Chapter 5).** Local link evidence is a form of relevance feedback. The link graph is filtered on the search topic, and thereby made more semantic. The amount of local link evidence is related to the exhaustivity dimension of relevance while the fraction of global link evidence that is present in the local set is related to the specificity dimension of relevance. Links in the local set have an impact in both the incoming and outgoing directions, supporting the notion that local link evidence is related to topical relevance.

- **Link evidence for topical relevance is more useful for ad hoc search than link evidence for document importance (Chapter 5).** We have seen that in Wikipedia, global link evidence results in a ranking that is only slightly better than random, although removing the global component from local link evidence hurts performance. Document importance plays a small role in Wikipedia ad hoc search, but a role nonetheless. In the TREC 2009 Web Ad Hoc task, the role of document importance is larger. Removing the global component in local link evidence seriously hurts performance. The topics for this task are more general and relevant documents are more abundant. The challenge of identifying relevant documents is smaller, and as a consequence, the challenge of identifying the best ones becomes more important. The role of document importance for ad hoc search seems related to the generality of the topic.

- **Link evidence for document importance is asymmetric. The direction of the link determines what the signal means (Chapter 5).** In the Web, where authors are typically not in control of incoming links from other sites, a site-external link from page A to page B confers authority from A to B, making B important but not A. In the Web, a link from page A to page B is a signal that page B contains useful information. If A and B are pages in different sites, the link confers authority from A to B. In both cases, the link signals that page B is important but not page A. At the same time, the link confers information about the length of page A, but not of page B. The direction of the link determines for what aspect of a document the link provides evidence.

- **Link evidence for topical relevance is symmetric. The meaning of the signal is independent of the link direction (Chapter 5).** In a graph of links between semantically related pages, incoming and outgoing links behave in a similar way and lead to similar degree distributions. Insofar as link evidence is related to topical relevance, it is independent of the link direction. Topical relevance is related to semantic relatedness and therefore symmetric.

- **Links between semantically related pages are more effective than links between unrelated ones (Chapter 6).** The effectiveness of link evidence depends on the semantic nature of the links. Evidence from links between semantically unrelated pages has almost no impact on retrieval, whereas evidence from links between semantically related pages can improve retrieval effectiveness.

- **Global link evidence is more useful for ad hoc retrieval in the Web than in Wikipedia (Chapter 7)**. The variation in document quality and authority, and the amount of spam are much bigger in the Web at large than in Wikipedia. This puts more emphasis on retrieving high-quality results in Web ad hoc search, and makes global link evidence more effective in the Web as a whole than in Wikipedia.

8.2.1 *Future Research*

Our findings that link evidence is effective for ad hoc retrieval in the INEX Wikipedia and ClueWeb09 B collections is in contrast with the findings of the 1999–2001 TREC Web Tracks, which used earlier crawls of the Web. We have discussed several factors that might contribute to this gap. Further analysis could provide a more complete answer

and might lead to an even better understanding of the value of link evidence for information retrieval.

First, in Chapters 4 and 7 we discussed the evolutionary phases of link graphs and how they might affect the value of link evidence for retrieval. Random filtering of links to control link density showed that, even in sparsely linked collections, link evidence can improve ad hoc retrieval. However, randomly removing links from a fully developed link graph is different from a link graph that is in an earlier evolutionary stage. Through preferential attachment (Barabási and Albert, 1999), authority and popularity become visible at an early stage, making global link evidence useful to find important pages. Only at a later stage do other pages acquire enough connections to derive topical relevance information from the graph. An interesting line of future research would be to look at the impact of graph evolution on the effectiveness of link evidence for retrieval.

Second, an aspect that is related to this is the nature of the growth process of link graphs. We have seen that the Web and Wikipedia both fit the model of scale-free networks, where the degree distribution adheres to a power law. Other hyperlinked document collections might grow in a different fashion, without preferential attachment and disassortative mixing. What is the relation between global link evidence and document importance in such networks? How does the growth process of document networks affect the value of link structure for information retrieval?

Third, we also briefly discussed the impact of the crawling policy on the composition of the resulting crawl. Policies that favour highly connected pages to be crawled first result ensure that the first part of the crawl is densely linked and that it contains many important pages. Other crawling policies result in a different composition of the crawl. Of course, this depends on the length of the crawl. In theory, if the crawl is exhaustive, making sure every reachable Web pages is crawled, the crawling order does not affect the final crawl. When the crawl is stopped at an earlier stage, the crawling order is important. Similar to the work by Fetterly et al. (2009a,b), we could measure the impact of link evidence on retrieval effectiveness using various stages of a crawl based on different crawling strategies and different sets of seed documents.

Fourth, the topic generality might play a role in the value of link evidence. The HITS algorithm was designed for broad topics, for which the search engine retrieves very many relevant pages with high precision. For these topics, the challenge is to identify the most authoritative

pages. For more specific (non-navigational) topics, the number of relevant pages is smaller and therefore, possibly the number of useful links as well. Here, it seems we need to make link evidence sensitive to the topical context. This suggests that link evidence has to be tailored to the specificity of the search topic and prompts us to look at how the value of link evidence is affected by the generality of the search topic. Another question is whether the link structure of the Web is rich enough to be effective for very specific topics found in the long tail of infrequent queries.

Another aspect for future research is the way link evidence is incorporated into the retrieval model. In Chapter 3 we have only used degrees as link evidence priors in combination with a standard language model retrieval system. But link evidence and text evidence could be combined in different ways. The difficulty with obtaining local, query-dependent link evidence is that it is dependent on the specific ranking function used to obtain the top retrieved results. The number of top retrieved results is also important and the optimal number is dependent on the query and the retrieval model. An alternative would be to look for ways in which local link evidence could be used directly in the ranking function, and where the retrieval model itself determines the size of the local set.

Finally, the positive results in this thesis cast a more optimistic light on the value of link information for information retrieval and might bring renewed interest in finding new ways of using link information for retrieval. In this thesis, we mainly looked at degrees, because we wanted to understand the nature of links, and only briefly looked at the impact of the more complex algorithms PageRank and HITS. However, there are many other ways of deriving meaning from the link structure. Outgoing links have proven to be a valuable source of link evidence, but is often ignored. The results in this thesis merit further investigation of when and how to use outgoing link evidence. Also, our findings prompt the question whether anchor text is also effective for ad hoc retrieval, and initial experiments have confirmed that in the ClueWeb09, anchor text can improve ad hoc retrieval performance (Koolen and Kamps, 2010), which might stimulate further research into the use of anchor text.

This thesis sheds more light on the meaning of link information, broadens the scope of its use for IR tasks and gives a more optimistic outlook for future research.

BIBLIOGRAPHY

S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *WWW*, pages 280–290, 2003. (Cited on pages 33, 171, and 172.)

L. A. Adamic. Zipf, Power-laws, and Pareto - a ranking tutorial, 09 2007. URL `http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html`. Last checked: 12-03-2010. (Cited on page 26.)

M. Agosti. Hypertext and information retrieval. *Inf. Process. Manage.*, 29(3):283–286, 1993. (Cited on page 26.)

D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. Using Wikipedia at the TREC QA Track. In *Proceedings of* TREC *2004*, 2004. (Cited on page 40.)

J. R. Anderson and P. L. Pirolli. Spread of activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(4):791–798, October 1984. (Cited on pages 24 and 139.)

J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In SIGIR *'06: Proceedings of the 29th annual international ACM* SIGIR *conference on Research and development in information retrieval*, pages 541–548, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: http://doi.acm.org/10.1145/1148170.1148263. (Cited on page 162.)

R. Baeza-yates and C. Castillo. Crawling the infinite web. *Journal of Web Engineering*, 1, 2005. (Cited on page 72.)

R. A. Baeza-Yates, C. Castillo, M. Marín, and A. Rodríguez. Crawling a country: better strategies than breadth-first for web page ordering. In A. Ellis and T. Hagino, editors, *WWW (Special interest tracks and posters)*, pages 864–872. ACM, 2005. ISBN 1-59593-051-5. (Cited on page 33.)

P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, 39 (6):853–871, 2003. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(02)00084-5. (Cited on pages 29, 160, and 170.)

A.-L. Barabási and R. Albert. The emergence of scaling in random networks. *Science*, 286:509–512, 1999. (Cited on pages 27, 52, and 203.)

M. J. Bates. Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution. In H. Bruce, R. Fidel, P. Ingwersen, and P. Vakkari, editors, *Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*, pages 137–150. Libraries Unlimited, 2002. (Cited on page 25.)

F. Bellomi and R. Bonato. Network Analysis for Wikipedia. *Proceedings of Wikimania*, 2005. (Cited on pages 39 and 52.)

T. Berners-Lee. Information Management: A Proposal, 1990. URL `http://www.w3.org/History/1989/proposal.html`. (Cited on page 26.)

K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: http://doi.acm.org/10.1145/290941.290972. (Cited on pages 28 and 35.)

W. J. Blustein. *Hypertext Versions of Journal Articles: Computer-aided linking and realistic human-based evaluation*. PhD thesis, University of Western Ontaria, London, Ontario, Canada, April 1999. (Cited on page 133.)

A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Internet Techn.*, 5(1):231–297, 2005. (Cited on page 38.)

R. A. Botafogo and B. Shneiderman. Identifying aggregates in hypertext structures. In *Hypertext*, pages 63–74. ACM, 1991. ISBN 0-89791-547-X. (Cited on page 24.)

R. A. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Trans. Inf. Syst.*, 10(2):142–180, 1992. (Cited on page 24.)

A. Broder. A taxonomy of web search. SIGIR *Forum*, 36(2):3–10, 2002. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/792550.792552. (Cited on pages 20, 29, and 73.)

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th International World-Wide Web conference WWW9*, pages 309–320. Elsevier Science, Amsterdam, 2000. (Cited on pages 26, 27, and 77.)

C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In SIGIR *'04: Proceedings of the 27th annual international ACM* SIGIR *conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-881-4. doi: http://doi.acm.org/10.1145/1008992.1009000. (Cited on page 47.)

A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006. (Cited on pages 135, 136, and 137.)

L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2747-7. doi: http://dx.doi.org/10.1109/WI.2006.164. (Cited on pages 40, 53, and 77.)

V. Bush. As we may think. *The Atlantic Monthly*, July 1945. URL `http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush-all.shtml`. (Cited on page 15.)

J. Callan, C. Yoo, and L. Zhao. Web08-PR Dataset, 2008. Project planning document. (Cited on pages 35 and 171.)

A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of Wikipedia. Physical Review E, Feb 2006. URL `http://arxiv.org/abs/physics/0602026`. (Cited on pages 39 and 52.)

A. Capocci, F. Rao, and G. Caldarelli. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia wikipedia. *EPL (Europhysics Letters)*, 81(2):28006+, 2008. doi: 10.1209/0295-5075/81/28006. URL `http://dx.doi.org/10.1209/0295-5075/81/28006`. (Cited on page 41.)

S. J. Carrière and R. Kazman. WebQuery: Searching and Visualizing the Web Through Connectivity. *Computer Networks*, 29(8-13):1257–1267, 1997. (Cited on pages 35 and 107.)

B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In SIGIR *'06: Proceedings of the 29th annual international ACM* SIGIR *conference on Research and development in information retrieval*, pages 268–275, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: http://doi.acm.org/10.1145/1148170.1148219. (Cited on page 162.)

S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock. The structure of broad topics on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 251–262, New York, NY, USA, 2002. ACM. ISBN 1-58113-449-5. doi: http://doi.acm.org/10.1145/511446.511480. (Cited on pages 28, 35, 54, 84, and 201.)

S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantics relationships between wikipedia categories. In M. Völkel and S. Schaffert, editors, *SemWiki*, volume 206 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006. (Cited on page 41.)

C. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In TREC *2009 Notebook*, pages 16–23, 2009. (Cited on pages 31, 162, and 163.)

C. Cleverdon. *The Cranfield tests on index language devices*, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-454-5. URL `http://portal.acm.org/citation.cfm?id=275537.275544`. (Cited on pages 20 and 21.)

P. R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manage.*, 23 (4):255–268, 1987. (Cited on page 24.)

W. S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971. (Cited on page 16.)

G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010. (Cited on page 166.)

E. Cosijn and P. Ingwersen. Dimensions of relevance. *Inf. Process. Manage.*, 36(4):533–550, 2000. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(99)00072-2. (Cited on page 15.)

N. Craswell, P. Bailey, and D. Hawking. Is it fair to evaluate Web systems using TREC ad hoc methods? In *ACM SIGIR '99 Workshop*

on Evaluation of Web Document Retrieval, 1999. (Cited on pages 19 and 29.)

N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 250–257. ACM Press, New York NY, USA, 2001. (Cited on pages 30 and 93.)

N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076106. (Cited on page 30.)

W. B. Croft and H. R. Turtle. Retrieval strategies for hypertext. *Inf. Process. Manage.*, 29(3):313–324, 1993. (Cited on page 25.)

B. D. Davison. Topical locality in the web. In *Research and Development in Information Retrieval (SIGIR)*, pages 272–279, 2000. URL `citeseer.ist.psu.edu/article/davison00topical.html`. (Cited on pages 27, 69, 133, and 160.)

A. P. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 entity ranking track. In Fuhr et al. (2008b), pages 245–251. ISBN 978-3-540-85901-7. (Cited on page 40.)

L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, June 2006. (Cited on pages 40 and 46.)

S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Internet Technol.*, 2(3):205–223, 2002. ISSN 1533-5399. doi: http://doi.acm.org/10.1145/572326.572328. (Cited on page 54.)

DMOZ. Open Directory Project, 2010. URL `http://www.dmoz.org`. (Cited on pages 35 and 54.)

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979. (Cited on page 63.)

N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages

459–460, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: http://doi.acm.org/10.1145/860435.860550. (Cited on pages 31 and 39.)

D. Ellis, J. Furner, and P. W. 0002. On the creation of hypertext links in full-text documents: Measurement of retrieval effectiveness. *JASIS*, 47(4):287–300, 1996. (Cited on page 25.)

R. A. Fairthorne. Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. *Journal of Documentation*, 25(4):319–343, 1969. (Cited on page 27.)

M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262. ACM Press, New York NY, USA, 1999. (Cited on page 52.)

D. Fetterly, N. Craswell, and V. Vinay. The impact of crawl policy on web search effectiveness. In *Proceedings of the 32nd international ACM* SIGIR *conference on Research and development in information retrieval*, pages 580–587, New York, NY, USA, 2009a. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572041. URL `http://portal.acm.org/citation.cfm?id=1571941.1572041`. (Cited on pages 33, 171, and 203.)

D. Fetterly, N. Craswell, and V. Vinay. Measuring the search effectiveness of a breadth-first crawl. In M. Boughanem, C. Berrut, J. Mothe, and C. Soulé-Dupuy, editors, *ECIR*, volume 5478 of *Lecture Notes in Computer Science*, pages 388–399. Springer, 2009b. ISBN 978-3-642-00957-0. (Cited on pages 33, 171, and 203.)

M. Fisher and R. Everson. When are links useful? experiments in text classification. In *In Advances in IR, 25th European Conference on IR research, ECIR*, pages 41–56. Springer-Verlag, 2003. (Cited on pages 132 and 170.)

W. Foundation. Wikimedia Downloads, 2009. URL `http://download.wikimedia.org/`. (Cited on page 46.)

H.-P. Frei and D. Stieger. The use of semantic links in hypertext information retrieval. *Inf. Process. Manage.*, 31(1):1–13, 1995. (Cited on page 25.)

N. Fuhr, J. Kamps, M. Lalmas, S. Malik, and A. Trotman. Overview of the INEX 2007 ad hoc track. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Pre-Proceedings of INEX 2007*, pages 1–22, 2007. (Cited on page 46.)

N. Fuhr, J. Kamps, M. Lalmas, S. Malik, and A. Trotman. Overview of the INEX 2007 ad hoc track. In N. Fuhr, M. Lalmas, A. Trotman, and J. Kamps, editors, *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, volume 4862 of *Lecture Notes in Computer Science*, pages 1–23. Springer Verlag, Heidelberg, 2008a. (Cited on page 40.)

N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, editors. *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval,* INEX *2007, Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers*, volume 4862 of *Lecture Notes in Computer Science*, 2008b. Springer. ISBN 978-3-540-85901-7. (Cited on pages 209 and 221.)

G. W. Furnas, S. C. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In Y. Chiaramella, editor, *SIGIR*, pages 465–480. ACM, 1988. (Cited on page 133.)

E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. (Cited on page 40.)

D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *HYPERTEXT '98: Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, pages 225–234, New York, NY, USA, 1998. ACM. ISBN 0-89791-972-6. doi: http://doi.acm.org/10.1145/276627.276652. (Cited on page 27.)

S. J. Green. Automatically generating hypertext in newspaper articles by computing semantic relatedness. In *NeMLaP3/CoNLL '98: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 101–110, Morristown, NJ, USA, 1998. Association for Computational Linguistics. ISBN 0-7258-0634-6. (Cited on page 133.)

C. Gurrin and A. F. Smeaton. Replicating web structure in small-scale test collections. *Inf. Retr.*, 7(3-4):239–263, 2004. ISSN 1386-4564. doi: http://dx.doi.org/10.1023/B:INRT.0000011206.23588.ab. (Cited on pages 7, 30, and 170.)

Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *AIRWeb*, pages 39–47, 2005. (Cited on page 172.)

D. Harman. Overview of the second text retrieval conference (trec-2). In *TREC*, pages 1–20, 1993. (Cited on page 14.)

S. P. Harter. Psychological relevance and information science. *JASIS*, 43 (9):602–615, 1992. (Cited on page 17.)

C. Hauff and D. Hiemstra. University of twente @ TREC 2009: Indexing half a billion web pages, 2010. (Cited on page 166.)

T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. on Knowl. and Data Eng.*, 15 (4):784–796, 2003. ISSN 1041-4347. doi: http://dx.doi.org/10.1109/TKDE.2003.1208999. (Cited on page 35.)

D. Hawking. Overview of the TREC-9 Web Track. In *TREC*, 2000. (Cited on page 29.)

D. Hawking. Overview of the TREC-9 Web Track. In *The Ninth Text REtrieval Conference (TREC-9)*, pages 87–102. National Institute for Standards and Technology. NIST Special Publication 500-249, 2001. (Cited on page 6.)

D. Hawking and N. Craswell. Overview of the TREC-2001 Web Track. In *Proceedings of TREC-2001*, November 2001. `http://david-hawking.net/pubs/hawking_trec01wt.pdf`. (Cited on page 30.)

D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005. (Cited on pages 5, 6, 29, 30, 171, and 189.)

D. Hawking and S. Robertson. On collection size and retrieval effectiveness. *Inf. Retr.*, 6(1):99–105, 2003. ISSN 1386-4564. doi: http://dx.doi.org/10.1023/A:1022904715765. (Cited on pages 173 and 181.)

D. Hawking, N. Craswell, P. B. Thistlewaite, and D. Harman. Results and challenges in web search evaluation. *Computer Networks*, 31 (11-16):1321–1330, 1999a. (Cited on page 19.)

D. Hawking, E. M. Voorhees, N. Craswell, and P. Bailey. Overview of the trec-8 web track. In *TREC*, 1999b. (Cited on page 29.)

D. Hawking, F. Crimmins, N. Craswell, and T. Upstill. How valuable is external link evidence when searching enterprise webs? In K.-D. Schewe and H. E. Williams, editors, *ADC*, volume 27 of *CRPIT*, pages 77–84. Australian Computer Society, 2004. ISBN 1-920682-06-6. (Cited on pages 177 and 184.)

J. He, K. Balog, K. Hofmann, E. J. Meij, M. de Rijke, E. Tsagkias, and W. Weerkamp. Heuristic ranking and diversification of web documents. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, February 2010. (Cited on page 180.)

W. Hersh. Relevance and retrieval evaluation: perspectives from medicine. *J. Am. Soc. Inf. Sci.*, 45(3):201–206, 1994. ISSN 0002-8231. doi: http://dx.doi.org/10.1002/(SICI)1097-4571(199404)45:3<201::AID-ASI9>3.0.CO;2-W. (Cited on page 17.)

D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001. (Cited on page 48.)

T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles. *Complex.*, 12(3):30–40, 2007. ISSN 1076-2787. doi: http://dx.doi.org/10.1002/cplx.v12:3. (Cited on page 136.)

D. W. Huang, Y. Xu, A. Trotman, and S. Geva. Overview of INEX 2007 link the wiki track. In *Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers*, pages 373–387, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85901-7. doi: http://dx.doi.org/10.1007/978-3-540-85902-4_32. (Cited on page 40.)

W. J. Hutchins. On the Problem of "Aboutness" in Document Analysis. *Journal of Informatics*, 1:17–35, 1977. (Cited on page 16.)

ILPS. The *ilps* extension of the *lucene* search engine, 2005. `http://ilps.science.uva.nl/Resources/`. (Cited on page 48.)

Indri. Language modeling meets inference networks, 2009. `http://www.lemurproject.org/indri/`. (Cited on page 165.)

B. J. Jansen and A. Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Inf. Process. Manage.*, 42(1):248–263, 2006. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/j.ipm.2004.10.007. (Cited on page 29.)

B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. SIGIR *Forum*, 32(1):5–17, 1998. (Cited on page 29.)

S. Janson, T. Luczak, and A. Rucinski. *Random Graphs*. Wiley Interscience, 2000. (Cited on page 76.)

J. Kamps. Web-centric language models. In A. Chowdhury, N. Fuhr, M. Ronthaler, and H.-J. Schek, editors, *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 307–308. ACM Press, New York NY, USA, 2005. (Cited on pages 30, 86, 88, and 93.)

J. Kamps. Effective smoothing for a terabyte of text. In E. M. Voorhees and L. P. Buckland, editors, *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*. National Institute of Standards and Technology. NIST Special Publication 500-266, 2006a. (Cited on page 165.)

J. Kamps. Experiments with document and query representations for a terabyte of text. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006b. (Cited on page 31.)

J. Kamps and M. Koolen. The importance of link evidence in Wikipedia. In C. Macdonald, I. Ounis, V. Plachouras, I. Rutven, and R. W. White, editors, *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282. Springer Verlag, Heidelberg, 2008. (Cited on page 11.)

J. Kamps and M. Koolen. Is Wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, pages 232–241. ACM Press, New York NY, USA, 2009. (Cited on page 12.)

J. Kamps, S. Fissaha Adafre, and M. de Rijke. Effective translation, tokenization and combination for cross-lingual retrieval. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images*, volume

3491 of *Lecture Notes in Computer Science*, pages 123–134. Springer Verlag, Heidelberg, 2005. (Cited on page 31.)

J. Kamps, S. Geva, A. Trotman, A. Woodley, and M. Koolen. Overview of the INEX 2008 ad hoc track. In S. Geva, J. Kamps, and A. Trotman, editors, *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, volume 5631 of *LNCS*, pages 1–28. Springer Verlag, Berlin, Heidelberg, 2009. (Cited on page 40.)

R. Kaptein and J. Kamps. Using links to classify Wikipedia pages. In S. Geva, J. Kamps, and A. Trotman, editors, *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, volume 5631 of *LNCS*. Springer Verlag, Berlin, Heidelberg, 2009. (Cited on page 137.)

R. Kaptein, M. Koolen, and J. Kamps. Using Wikipedia categories for ad hoc search. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 824–825. ACM Press, New York NY, USA, 2009. (Cited on page 40.)

L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953. doi: 10.1007/BF02289026. URL `http://dx.doi.org/10.1007/BF02289026`. (Cited on page 27.)

M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963a. (Cited on page 24.)

M. Kessler. An experimental study of bibliographic coupling between technical papers (corresp.). *Information Theory, IEEE Transactions on*, 9(1):49 – 51, jan 1963b. ISSN 0018-9448. (Cited on pages 24 and 139.)

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. ISSN 0004-5411. doi: http://doi.acm.org/10.1145/324133.324140. (Cited on pages 4, 27, 50, 91, 160, and 173.)

J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. In *COCOON*, pages 1–17, 1999. (Cited on pages 26, 27, and 72.)

M. Kochen. *Principles of information retrieval*. Melville Pub. Co. Los Angeles,, 1974. ISBN 0471496979. (Cited on page 15.)

M. Koolen and J. Kamps. What's in a link? from document importance to topical relevance. In L. Azzopardi, G. Kazai, S. Robertson, S. Rüger, M. Shokouhi, D. Song, and E. Yilmas, editors, *Proceedings of the 2nd International Conferences on the Theory of Information Retrieval (ICTIR 2009)*, volume 5766 of *LNCS*, pages 313–321. Springer Verlag, Berlin, Heidelberg, 2009. (Cited on page 12.)

M. Koolen and J. Kamps. The importance of anchor-text for ad hoc search revisited. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York NY, USA, 2010. (Cited on pages 14 and 204.)

M. Koolen and J. Kamps. Are semantically related links effective for retrieval? In P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, and V. Murdoch, editors, *Advances in Information Retrieval: 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *LNCS*, pages 92–103. Springer, 2011. (Cited on page 13.)

W. Kraaij and T. Westerveld. How Different are Web Documents? In *Proceedings of the ninth Text Retrieval Conference, TREC-9*. NIST Special Publication, May 2001. (Cited on pages 6 and 69.)

W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, New York NY, USA, 2002. (Cited on pages 30 and 93.)

R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Computer Networks*, volume 31, pages 1481–1493, May 1999. (Cited on page 27.)

O. Kurland and L. Lee. Pagerank without hyperlinks: structural re-ranking using links induced by language models. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, editors, *SIGIR*, pages 306–313. ACM, 2005. ISBN 1-59593-034-5. (Cited on page 133.)

O. Kurland and L. Lee. Respect my authority!: Hits without hyperlinks, utilizing cluster-based language models. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors, *SIGIR*, pages 83–90. ACM, 2006. ISBN 1-59593-369-7. (Cited on page 133.)

M. Lalmas and B. Piwowarski. INEX 2006 Relevance Assessment Guide. In *Pre-Proceedings of* INEX *2006*, pages 389–395, 2006. (Cited on pages 46 and 120.)

M. Lalmas and B. Piwowarski. INEX 2007 Relevance Assessment Guide. In *Pre-Proceedings of* INEX *2007*, pages 454–463, 2007. (Cited on pages 46 and 120.)

S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999. (Cited on page 170.)

R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001. (Cited on page 38.)

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM Press, New York, NY, USA, 2005. (Cited on pages 75 and 133.)

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1):2, 2007. (Cited on page 75.)

J. Lin, D. Metzler, T. Elsayed, and L. Wang. Of Ivory and Smurfs: Loxodontan MapReduce Experiments for Web Search. In TREC, 2010. (Cited on page 166.)

D. Lizorkin, O. Medelyan, and M. Grineva. Analysis of community structure in wikipedia. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 1221–1222. ACM, 2009. ISBN 978-1-60558-487-4. (Cited on page 40.)

Lucene. Welcome to Lucene, 2010. URL `http://lucene.apache.org/`. (Cited on page 47.)

S. Malik, A. Trotman, M. Lalmas, and N. Fuhr. Overview of INEX 2006. In N. Fuhr, M. Lalmas, and A. Trotman, editors, INEX, volume 4518 of *Lecture Notes in Computer Science*, pages 1–11. Springer, 2006. ISBN 978-3-540-73887-9. (Cited on pages 40 and 46.)

M. Marchiori. The quest for correct information on the web: Hyper search engines. *Computer Networks*, 29(8-13):1225–1236, 1997. (Cited on page 35.)

O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67 (9):716–754, 2009. ISSN 1071-5819. doi: http://dx.doi.org/10.1016/j.ijhcs.2009.05.004. (Cited on page 41.)

D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In SIGIR *'09: Proceedings of the 32nd international ACM* SIGIR *conference on Research and development in information retrieval*, pages 219–226, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1571981. (Cited on pages 31 and 39.)

D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*, pages 25–30. AAAI, 2008. (Cited on pages 40 and 135.)

S. Mizzaro. How many relevances in information retrieval? *Interacting with Computers*, 10(3):303–320, 1998. (Cited on page 16.)

M. A. Najork, H. Zaragoza, and M. J. Taylor. HITS on the Web: How does it compare? In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 471–478. ACM, New York, NY, USA, 2007. (Cited on pages 35 and 75.)

L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 91–98, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: http://doi.acm.org/10.1145/1148170.1148189. (Cited on page 30.)

ODP. Open Directory Project, 2010. URL `http://www.dmoz.org/`. last checked on 7 July 2010. (Cited on page 30.)

P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 143–150. ACM Press, New York NY, USA, 2003. (Cited on page 30.)

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. URL `citeseer.ist.psu.edu/page98pagerank.html`. (Cited on pages 4, 27, 33, and 36.)

S. Pal, M. Mitra, and A. Chakraborty. Stability of INEX 2007 evaluation measures. In *Proceedings of the Second International Workshop on EValuating Information Access (EVIA 2008), NTCIR 7*, pages 23–29, 2008. (Cited on page 47.)

S. Pal, M. Mitra, and J. Kamps. Evaluation effort, reliability and reusability in XML retrieval. *Journal of the American Society for Information Science and Technology*, 61, 2010. (Cited on page 47.)

J. Pehcevski and B. Larsen. Relevance. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 2377–2378. Springer US, 2009. ISBN 978-0-387-35544-3, 978-0-387-39940-9. (Cited on page 109.)

J. Pehcevski, A.-M. Vercoustre, and J. A. Thom. Exploiting locality of wikipedia links in entity ranking. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2008. ISBN 978-3-540-78645-0. (Cited on page 40.)

J. Picard. Modeling and combining evidence provided by document relationships using probabilistic argumentation systems. In *SIGIR*, pages 182–189. ACM, 1998. (Cited on page 25.)

J. Picard and J. Savoy. Enhancing retrieval with hyperlinks: A general model based on propositional argumentation systems. *JASIST*, 54(4): 347–355, 2003. (Cited on page 25.)

P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: extracting usable structures from the web. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 118–125, New York, NY, USA, 1996. ACM. ISBN 0-89791-777-4. doi: http://doi.acm.org/10.1145/238386.238450. (Cited on page 26.)

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: http://doi.acm.org/10.1145/290941.291008. (Cited on page 48.)

D. D. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976. doi: 10.1002/asi.4630270505. (Cited on page 27.)

X. Qi, L. Nie, and B. D. Davison. Measuring similarity to detect qualified links. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 49–56, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-732-2. doi: http://doi.acm.org/10.1145/1244408.1244418. (Cited on page 139.)

R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989. (Cited on page 136.)

S. Rajput, V. Pavlu, J. As, and E. Kanoulas. Northeastern University in the TREC 2009 Web Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, February 2010. (Cited on page 166.)

W. B. Rayward. Visions of xanadu: Paul otlet (1868–1944) and hypertext. *Journal of the American Society for Information Science*, 45(4):235–250, 1994. ISSN 0002-8231. (Cited on pages 24 and 41.)

P. Resnik. Using information content to evaluate semantic similarity in a taxanomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, 1995. (Cited on page 136.)

S. Robertson. On the history of evaluation in IR. *J. Information Science*, 34(4):439–456, 2008. (Cited on pages 15 and 21.)

S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC*, pages 0–, 1994. (Cited on page 50.)

D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM. ISBN 1-58113-844-X. doi: http://doi.acm.org/10.1145/988672.988675. URL `http://doi.acm.org/10.1145/988672.988675`. (Cited on page 20.)

T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975. (Cited on pages 15, 16, and 22.)

T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):1915–1933, 2007. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/asi.v58:13. (Cited on page 16.)

J. Savoy. Bayesian inference networks and spreading activation in hypertext systems. *Inf. Process. Manage.*, 28(3):389–406, 1992. (Cited on page 26.)

J. Savoy. A learning scheme for information retrieval in hypertext. *Inf. Process. Manage.*, 30(4):515–534, 1994. (Cited on page 25.)

J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33:495–512, 1997. (Cited on page 63.)

J. R. Seeley. The net of reciprocal influence. *Canadian Journal of Psychology*, 3:234–240, 1949. (Cited on page 27.)

A. Shakery and C. Zhai. A probabilistic relevance propagation model for hypertext retrieval. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, editors, *CIKM*, pages 550–558. ACM, 2006. ISBN 1-59593-433-2. (Cited on pages 28, 38, and 139.)

C. Sherman. 'Old Economy' Info Retrieval Clashes with 'New Economy' Web Upstarts at the Fifth Annual Search Engine Conference. *Information Today Newsbreaks*, 2000. `http://web.archive.org/web/20001217211000/www.infotoday.com/newsbreaks/nb000424-2.htm`. (Cited on page 189.)

C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. SIGIR *Forum*, 33(1):6–12, 1999. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/331403.331405. (Cited on page 29.)

A. Singhal and M. Kaszkiel. At&t at trec-9. In *TREC*, 2000. (Cited on page 29.)

A. Singhal and M. Kaszkiel. A case study in web search using trec algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 708–716, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: http://doi.acm.org/10.1145/371920.372186. (Cited on page 29.)

A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In SIGIR *'96: Proceedings of the 19th annual international ACM* SIGIR *conference on Research and development in information retrieval*, pages 21–29, New York, NY, USA, 1996. ACM. ISBN 0-89791-792-8. doi: http://doi.acm.org/10.1145/243199.243206. (Cited on pages 48, 88, and 166.)

I. Soboroff. Do trec web collections look like the web? *SIGIR Forum*, 36 (2):23–31, 2002. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/792550.792554. (Cited on pages 70 and 77.)

D. Soergel. Indexing and retrieval performance: The logical evidence. *JASIS*, 45(8):589–599, 1994. (Cited on page 17.)

K. Sparck-Jones and C. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. Technical report, Computer Laboratory, University of Cambridge, 1975. (Cited on page 21.)

M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, July 2006. (Cited on pages 40 and 136.)

CMU-LTI. The ClueWeb09 Dataset, 2009. URL `http://boston.lti.cs.cmu.edu/Data/clueweb09/`. (Cited on pages 31, 161, and 171.)

ILPS. The ILPS extension of the Lucene search engine, 2005. URL `http://ilps.science.uva.nl/resources/lm-lucene`. (Cited on page 47.)

INEX. INitiative for the Evaluation of XML retrieval, 2009. `http://www.inex.otago.ac.nz/`. (Cited on page 7.)

TREC. Text-REtrieval Conference, 2009. `http://trec.nist.gov/`. (Cited on pages 5 and 29.)

T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. P. de Vries. Structured document retrieval, multimedia retrieval, and entity ranking using pf/tijah. In Fuhr et al. (2008b), pages 306–320. ISBN 978-3-540-85901-7. (Cited on page 38.)

P. Vakkari. Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval. *Inf. Process. Manage.*, 35(6):819–837, 1999. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(99)00028-X. (Cited on page 18.)

E. M. Voorhees. The philosophy of information retrieval evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, CLEF 2001, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2002. (Cited on page 21.)

J. Voss. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005. URL `http://www.citebase.org/cgi-bin/citations?id=oai:eprints.rclis.org:3610`. (Cited on pages 39 and 52.)

W3C. A Little History of the World Wide Web, 2010. URL `http://www.w3.org/History.html`. Last checked 17 June 2010. (Cited on page 171.)

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge MA, 1994. (Cited on pages 4, 27, and 174.)

Wikipedia. Linking, 2009. URL `http://en.wikipedia.org/wiki/Wikipedia:Linking`. (Cited on page 111.)

Wikipedia. the free encyclopedia, 2010. URL `http://en.wikipedia.org/wiki/Wikipedia:Linking#Overlinking_and_underlinking`. (Cited on page 8.)

P. Willett. Recent trends in hierarchic document clustering: a critical review. *Inf. Process. Manage.*, 24(5):577–597, 1988. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/0306-4573(88)90027-1. (Cited on page 25.)

P. Wilson. Situational relevance. *Information Storage and Retrieval*, 9(8): 457–471, 1973. (Cited on page 16.)

WordNet. A lexical database for English, 2010. `http://wordnet.princeton.edu/`. (Cited on page 136.)

WorldWideWebSize. Daily estimated size of the World Wide Web, 2010. URL `http://www.worldwidewebsize.com`. Last checked 22 June 2010. (Cited on page 171.)

J. Xu and W. B. Croft. Query expansion using local and global document analysis. In SIGIR *'96: Proceedings of the 19th annual international ACM* SIGIR *conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM. ISBN 0-89791-792-8. doi: http://doi.acm.org/10.1145/243199.243202. (Cited on page 37.)

J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/333135.333138. (Cited on page 37.)

E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. doi: http://doi.acm.org/10.1145/1183614.1183633. (Cited on page 166.)

H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1015–1018, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: http://doi.acm.org/10.1145/1321440.1321599. (Cited on page 40.)

T. Zesch and I. Gurevych. Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pages 1–8, Apr 2007. (Cited on page 136.)

V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks, Jul 2006. URL `http://arxiv.org/abs/physics/0602149`. (Cited on pages 39 and 53.)

Titles in the SIKS *Dissertation Series:*

1998-01: **Johan van den Akker (CWI)**
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-02: **Floris Wiesman (UM)**
Information Retrieval by Graphically Browsing Meta-Information

1998-03: **Ans Steuten (TUD)**
Conversations within the Language/Action Perspective

1998-04: **Dennis Breuker (UM)**
Memory versus Search in Games

1998-05: **E.W.Oskamp (RUL)**
Computerondersteuning bij Straftoemeting

1999-01: **Mark Sloof (VU)**
Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-02: **Rob Potharst (EUR)**
Classification using decision trees and neural nets

1999-03: **Don Beal (UM)**
The Nature of Minimax Search

1999-04: **Jacques Penders (UM)**
The practical Art of Moving Physical Objects

1999-05: **Aldo de Moor (KUB)**
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-06: **Niek J.E. Wijngaards (VU)**
Re-design of compositional systems

1999-07: **David Spelt (UT)**
Verification support for object database design

1999-08: **Jacques H.J. Lenting (UM)**
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.

2000-01: **Frank Niessink (VU)**
Perspectives on Improving Software Maintenance

2000-02: **Koen Holtman (TUE)**
Prototyping of CMS Storage Management

2000-03: **Carolien M.T. Metselaar (UVA)**
Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.

2000-04: **Geert de Haan (VU)**
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-05: **Ruud van der Pol (UM)**
Knowledge-based Query Formulation in Information Retrieval.

2000-06: **Rogier van Eijk (UU)**
Programming Languages for Agent Communication

2000-07: **Niels Peek (UU)**
Decision-theoretic Planning of Clinical Patient Management

2000-08: **Veerle Coup (EUR)**
Sensitivity Analyis of Decision-Theoretic Networks

2000-09: **Florian Waas (CWI)**
Principles of Probabilistic Query Optimization

2000-10: **Niels Nes (CWI)**
Image Database Management System Design Considerations, Algorithms and Architecture

2000-11: **Jonas Karlsson (CWI)**
Scalable Distributed Data Structures for Database Management

2001-01: **Silja Renooij (UU)**
Qualitative Approaches to Quantifying Probabilistic Networks

2001-02: **Koen Hindriks (UU)**
Agent Programming Languages: Programming with Mental Models

2001-03: **Maarten van Someren (UvA)**
Learning as problem solving

2001-04: **Evgueni Smirnov (UM)**
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-05: **Jacco van Ossenbruggen (VU)**
Processing Structured Hypermedia: A Matter of Style

2001-06: **Martijn van Welie (VU)**
Task-based User Interface Design

2001-07: **Bastiaan Schonhage (VU)**
Diva: Architectural Perspectives on Information Visualization

2001-08: **Pascal van Eck (VU)**
A Compositional Semantic Structure for Multi-Agent Systems Dynamics.

2001-09: **Pieter Jan 't Hoen (RUL)**
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10: **Maarten Sierhuis (UvA)**
Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design

2001-11: **Tom M. van Engers (VUA)**
Knowledge Management: The Role of Mental Models in Business Systems Design

2002-01: **Nico Lassing (VU)**
Architecture-Level Modifiability Analysis

2002-02: **Roelof van Zwol (UT)**
Modelling and searching web-based document collections

2002-03: **Henk Ernst Blok (UT)**
Database Optimization Aspects for Information Retrieval

2002-04: **Juan Roberto Castelo Valdueza (UU)**
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05: **Radu Serban (VU)**
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents

2002-06: **Laurens Mommers (UL)**
Applied legal epistemology; Building a knowledge-based ontology of the legal domain

2002-07: **Peter Boncz (CWI)**
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08: **Jaap Gordijn (VU)**
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09: **Willem-Jan van den Heuvel(KUB)**
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10: **Brian Sheppard (UM)**
Towards Perfect Play of Scrabble

2002-11: **Wouter C.A. Wijngaards (VU)**
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12: **Albrecht Schmidt (Uva)**
Processing XML in Database Systems

2002-13: **Hongjing Wu (TUE)**
A Reference Architecture for Adaptive Hypermedia Applications

2002-14: **Wieke de Vries (UU)**
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15: **Rik Eshuis (UT)**
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16: **Pieter van Langen (VU)**
The Anatomy of Design: Foundations, Models and Applications

2002-17: **Stefan Manegold (UVA)**
Understanding, Modeling, and Improving Main-Memory Database Performance

2003-01: **Heiner Stuckenschmidt (VU)**
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02: **Jan Broersen (VU)**
Modal Action Logics for Reasoning About Reactive Systems

2003-03: **Martijn Schuemie (TUD)**
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04: **Milan Petkovic (UT)**
Content-Based Video Retrieval Supported by Database Technology

2003-05: **Jos Lehmann (UVA)**
Causation in Artificial Intelligence and Law - A modelling approach

2003-06: **Boris van Schooten (UT)**
Development and specification of virtual environments

2003-07: **Machiel Jansen (UvA)**
Formal Explorations of Knowledge Intensive Tasks

2003-08: **Yongping Ran (UM)**
Repair Based Scheduling

2003-09: **Rens Kortmann (UM)**
The resolution of visually guided behaviour

2003-10: **Andreas Lincke (UvT)**
interaction between medium, innovation context and culture

2003-11: **Simon Keizer (UT)**
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12: **Roeland Ordelman (UT)**
Dutch speech recognition in multimedia information retrieval

2003-13: **Jeroen Donkers (UM)**
Nosce Hostem - Searching with Opponent Models

2003-14: **Stijn Hoppenbrouwers (KUN)**
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15: **Mathijs de Weerdt (TUD)**
Plan Merging in Multi-Agent Systems

2003-16: **Menzo Windhouwer (CWI)**
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17: **David Jansen (UT)**
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18: **Levente Kocsis (UM)**
Learning Search Decisions

2004-01: **Virginia Dignum (UU)**
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02: **Lai Xu (UvT)**
Monitoring Multi-party Contracts for E-business

2004-03: **Perry Groot (VU)**
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04: **Chris van Aart (UVA)**
Organizational Principles for Multi-Agent Architectures

2004-05: **Viara Popova (EUR)**
Knowledge discovery and monotonicity

2004-06: **Bart-Jan Hommes (TUD)**
The Evaluation of Business Process Modeling Techniques

2004-07: **Elise Boltjes (UM)**
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08: **Joop Verbeek (UM)**
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise

2004-09: **Martin Caminada (VU)**
For the Sake of the Argument; explorations into argument-based reasoning

2004-10: **Suzanne Kabel (UVA)**
Knowledge-rich indexing of learning-objects

2004-11: **Michel Klein (VU)**
Change Management for Distributed Ontologies

2004-12: **The Duy Bui (UT)**
Creating emotions and facial expressions for embodied agents

2004-13: **Wojciech Jamroga (UT)**
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14: **Paul Harrenstein (UU)**
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15: **Arno Knobbe (UU)**
Multi-Relational Data Mining

2004-16: **Federico Divina (VU)**
Hybrid Genetic Relational Search for Inductive Learning

2004-17: **Mark Winands (UM)**
Informed Search in Complex Games

2004-18: **Vania Bessa Machado (UvA)**
Supporting the Construction of Qualitative Knowledge Models

2004-19: **Thijs Westerveld (UT)**
Using generative probabilistic models for multimedia retrieval

2004-20: **Madelon Evers (Nyenrode)**
Learning from Design: facilitating multidisciplinary design teams

2005-01: **Floor Verdenius (UVA)**
Methodological Aspects of Designing Induction-Based Applications

2005-02: **Erik van der Werf (UM)**
AI techniques for the game of Go

2005-03: **Franc Grootjen (RUN)**
A Pragmatic Approach to the Conceptualisation of Language

2005-04: **Nirvana Meratnia (UT)**
Towards Database Support for Moving Object data

2005-05: **Gabriel Infante-Lopez (UVA)**
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06: **Pieter Spronck (UM)**
Adaptive Game AI

2005-07: **Flavius Frasincar (TUE)**
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08: **Richard Vdovjak (TUE)**
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09: **Jeen Broekstra (VU)**
Storage, Querying and Inferencing for Semantic Web Languages

2005-10: **Anders Bouwer (UVA)**
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11: **Elth Ogston (VU)**
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12: **Csaba Boer (EUR)**
Distributed Simulation in Industry

2005-13: **Fred Hamburg (UL)**
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14: **Borys Omelayenko (VU)**
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15: **Tibor Bosse (VU)**
Analysis of the Dynamics of Cognitive Processes

2005-16: **Joris Graaumans (UU)**
Usability of XML Query Languages

2005-17: **Boris Shishkov (TUD)**
Software Specification Based on Re-usable Business Components

2005-18: **Danielle Sent (UU)**
Test-selection strategies for probabilistic networks

2005-19: **Michel van Dartel (UM)**
Situated Representation

2005-20: **Cristina Coteanu (UL)**
Cyber Consumer Law, State of the Art and Perspectives

2005-21: **Wijnand Derks (UT)**
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006-01: **Samuil Angelov (TUE)**
Foundations of B2B Electronic Contracting

2006-02: **Cristina Chisalita (VU)**
Contextual issues in the design and use of information technology in organizations

2006-03: **Noor Christoph (UVA)**
The role of metacognitive skills in learning to solve problems

2006-04: **Marta Sabou (VU)**
Building Web Service Ontologies

2006-05: **Cees Pierik (UU)**
Validation Techniques for Object-Oriented Proof Outlines

2006-06: **Ziv Baida (VU)**
Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling

2006-07: **Marko Smiljanic (UT)**
XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08: **Eelco Herder (UT)**
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09: **Mohamed Wahdan (UM)**
Automatic Formulation of the Auditor's Opinion

2006-10: **Ronny Siebes (VU)**
Semantic Routing in Peer-to-Peer Systems

2006-11: **Joeri van Ruth (UT)**
Flattening Queries over Nested Data Types

2006-12: **Bert Bongers (VU)**
Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13: **Henk-Jan Lebbink (UU)**
Dialogue and Decision Games for Information Exchanging Agents

2006-14: **Johan Hoorn (VU)**
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15: **Rainer Malik (UU)**
CONAN: Text Mining in the Biomedical Domain

2006-16: **Carsten Riggelsen (UU)**
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17: **Stacey Nagata (UU)**
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18: **Valentin Zhizhkun (UVA)**
Graph transformation for Natural Language Processing

2006-19: **Birna van Riemsdijk (UU)**
Cognitive Agent Programming: A Semantic Approach

2006-20: **Marina Velikova (UvT)**
Monotone models for prediction in data mining

2006-21: **Bas van Gils (RUN)**
Aptness on the Web

2006-22: **Paul de Vrieze (RUN)**
Fundaments of Adaptive Personalisation

2006-23: **Ion Juvina (UU)**
Development of Cognitive Model for Navigating on the Web

2006-24: **Laura Hollink (VU)**
Semantic Annotation for Retrieval of Visual Resources

2006-25: **Madalina Drugan (UU)**
Conditional log-likelihood MDL and Evolutionary MCMC

2006-26: **Vojkan Mihajlović (UT)**
Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27: **Stefano Bocconi (CWI)**
Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28: **Borkur Sigurbjornsson (UVA)**
Focused Information Access using XML Element Retrieval

2007-01: **Kees Leune (UvT)**
Access Control and Service-Oriented Architectures

2007-02: **Wouter Teepe (RUG)**
Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03: **Peter Mika (VU)**
Social Networks and the Semantic Web

2007-04: **Jurriaan van Diggelen (UU)**
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-05: **Bart Schermer (UL)**
Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-06: **Gilad Mishne (UVA)**
Applied Text Analytics for Blogs

2007-07: **Natasa Jovanovic' (UT)**
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

2007-08: **Mark Hoogendoorn (VU)**
Modeling of Change in Multi-Agent Organizations

2007-09: **David Mobach (VU)**
Agent-Based Mediated Service Negotiation

2007-10: **Huib Aldewereld (UU)**
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11: **Natalia Stash (TUE)**
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12: **Marcel van Gerven (RUN)**
Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

2007-13: **Rutger Rienks (UT)**
Meetings in Smart Environments; Implications of Progressing Technology

2007-14: **Niek Bergboer (UM)**
Context-Based Image Analysis

2007-15: **Joyca Lacroix (UM)**
NIM: a Situated Computational Memory Model

2007-16: **Davide Grossi (UU)**
Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems

2007-17: **Theodore Charitos (UU)**
Reasoning with Dynamic Networks in Practice

2007-18: **Bart Orriens (UvT)**
On the development an management of adaptive business collaborations

2007-19: **David Levy (UM)**
Intimate relationships with artificial partners

2007-20: **Slinger Jansen (UU)**
Customer Configuration Updating in a Software Supply Network

2007-21: **Karianne Vermaas (UU)**
Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005

2007-22: **Zlatko Zlatev (UT)**
Goal-oriented design of value and process models from patterns

2007-23: **Peter Barna (TUE)**
Specification of Application Logic in Web Information Systems

2007-24: **Georgina Ramírez Camps (CWI)**
Structural Features in XML Retrieval

2007-25: **Joost Schalken (VU)**
Empirical Investigations in Software Process Improvement

2008-01: **Katalin Boer-Sorbán (EUR)**
Agent-Based Simulation of Financial Markets: A modular,continuous-time approach

2008-02: **Alexei Sharpanskykh (VU)**
On Computer-Aided Methods for Modeling and Analysis of Organizations

2008-03: **Vera Hollink (UVA)**
Optimizing hierarchical menus: a usage-based approach

2008-04: **Ander de Keijzer (UT)**
Management of Uncertain Data - towards unattended integration

2008-05: **Bela Mutschler (UT)**
Modeling and simulating causal dependencies on process-aware information systems from a cost perspective

2008-06: **Arjen Hommersom (RUN)**
On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective

2008-07: **Peter van Rosmalen (OU)**
Supporting the tutor in the design and support of adaptive e-learning

2008-08: **Janneke Bolt (UU)**
Bayesian Networks: Aspects of Approximate Inference

2008-09: **Christof van Nimwegen (UU)**
The paradox of the guided user: assistance can be counter-effective

2008-10: **Wauter Bosma (UT)**
Discourse oriented summarization

2008-11: **Vera Kartseva (VU)**
Designing Controls for Network Organizations: A Value-Based Approach

2008-12: **Jozsef Farkas (RUN)**
A Semiotically Oriented Cognitive Model of Knowledge Representation

2008-13: **Caterina Carraciolo (UVA)**
Topic Driven Access to Scientific Handbooks

2008-14: **Arthur van Bunningen (UT)**
Context-Aware Querying; Better Answers with Less Effort

2008-15: **Martijn van Otterlo (UT)**
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.

2008-16: **Henriette van Vugt (VU)**
Embodied agents from a user's perspective

2008-17: **Martin Op 't Land (TUD)**
Applying Architecture and Ontology to the Splitting and Allying of Enterprises

2008-18: **Guido de Croon (UM)**
Adaptive Active Vision

2008-19: **Henning Rode (UT)**
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search

2008-20: **Rex Arendsen (UVA)**
Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven

2008-21: **Krisztian Balog (UVA)**
People Search in the Enterprise

2008-22: **Henk Koning (UU)**
Communication of IT-Architecture

2008-23: **Stefan Visscher (UU)**
Bayesian network models for the management of ventilator-associated pneumonia

2008-24: **Zharko Aleksovski (VU)**
Using background knowledge in ontology matching

2008-25: **Geert Jonker (UU)**
Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency

2008-26: **Marijn Huijbregts (UT)**
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled

2008-27: **Hubert Vogten (OU)**
Design and Implementation Strategies for IMS Learning Design

2008-28: **Ildiko Flesch (RUN)**
On the Use of Independence Relations in Bayesian Networks

2008-29: **Dennis Reidsma (UT)**
Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans

2008-30: **Wouter van Atteveldt (VU)**
Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content

2008-31: **Loes Braun (UM)**
Pro-Active Medical Information Retrieval

2008-32: **Trung H. Bui (UT)**
Toward Affective Dialogue Management using Partially Observable Markov Decision Processes

2008-33: **Frank Terpstra (UVA)**
Scientific Workflow Design; theoretical and practical issues

2008-34: **Jeroen de Knijf (UU)**
Studies in Frequent Tree Mining

2008-35: **Ben Torben Nielsen (UvT)**
Dendritic morphologies: function shapes structure

2009-01: **Rasa Jurgelenaite (RUN)**
Symmetric Causal Independence Models

2009-02: **Willem Robert van Hage (VU)**
Evaluating Ontology-Alignment Techniques

2009-03: **Hans Stol (UvT)**
A Framework for Evidence-based Policy Making Using IT

2009-04: **Josephine Nabukenya (RUN)**
Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-05: **Sietse Overbeek (RUN)**
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-06: **Muhammad Subianto (UU)**
Understanding Classification

2009-07: **Ronald Poppe (UT)**
Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-08: **Volker Nannen (VU)**
Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-09: **Benjamin Kanagwa (RUN)**
Design, Discovery and Construction of Service-oriented Systems

2009-10: **Jan Wielemaker (UVA)**
Logic programming for knowledge-intensive interactive applications

2009-11: **Alexander Boer (UVA)**
Legal Theory, Sources of Law & the Semantic Web

2009-12: **Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)**
Operating Guidelines for Services

2009-13: **Steven de Jong (UM)**
Fairness in Multi-Agent Systems

2009-14: **Maksym Korotkiy (VU)**
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)

2009-15: **Rinke Hoekstra (UVA)**
Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16: **Fritz Reul (UvT)**
New Architectures in Computer Chess

2009-17: **Laurens van der Maaten (UvT)**
Feature Extraction from Visual Data

2009-18: **Fabian Groffen (CWI)**
Armada, An Evolving Database System

2009-19: **Valentin Robu (CWI)**
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets

2009-20: **Bob van der Vecht (UU)**
Adjustable Autonomy: Controling Influences on Decision Making

2009-21: **Stijn Vanderlooy (UM)**
Ranking and Reliable Classification

2009-22: **Pavel Serdyukov (UT)**
Search For Expertise: Going beyond direct evidence

2009-23: **Peter Hofgesang (VU)**
Modelling Web Usage in a Changing Environment

2009-24: **Annerieke Heuvelink (VUA)**
Cognitive Models for Training Simulations

2009-25: **Alex van Ballegooij (CWI)**
RAM: Array Database Management through Relational Mapping"

2009-26: **Fernando Koch (UU)**
An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27: **Christian Glahn (OU)**
Contextual Support of social Engagement and Reflection on the Web

2009-28: **Sander Evers (UT)**
Sensor Data Management with Probabilistic Models

2009-29: **Stanislav Pokraev (UT)**
Model-Driven Semantic Integration of Service-Oriented Applications

2009-30: **Marcin Zukowski (CWI)**
Balancing vectorized query execution with bandwidth-optimized storage

2009-31: **Sofiya Katrenko (UVA)**
A Closer Look at Learning Relations from Text

2009-32: **Rik Farenhorst (VU) and Remco de Boer (VU)**
Architectural Knowledge Management: Supporting Architects and Auditors

2009-33: **Khiet Truong (UT)**
How Does Real Affect Affect Affect Recognition In Speech?

2009-34: **Inge van de Weerd (UU)**
Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35: **Wouter Koelewijn (UL)**
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36: **Marco Kalz (OUN)**
Placement Support for Learners in Learning Networks

2009-37: **Hendrik Drachsler (OUN)**
Navigation Support for Learners in Informal Learning Networks

2009-38: **Riina Vuorikari (OU)**
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context

2009-39: **Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)**
Service Substitution – A Behavioral Approach Based on Petri Nets

2009-40: **Stephan Raaijmakers (UvT)**
Multinomial Language Learning: Investigations into the Geometry of Language

2009-41: **Igor Berezhnyy (UvT)**
Digital Analysis of Paintings

2009-42: **Toine Bogers (UvT)**
Recommender Systems for Social Bookmarking

2009-43: **Virginia Nunes Leal Franqueira (UT)**
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients

2009-44: **Roberto Santana Tapia (UT)**
Assessing Business-IT Alignment in Networked Organizations

2009-45: **Jilles Vreeken (UU)**
Making Pattern Mining Useful

2009-46: **Loredana Afanasiev (UvA)**
Querying XML: Benchmarks and Recursion

2010-01: **Matthijs van Leeuwen (UU)**
Patterns that Matter

2010-02: **Ingo Wassink (UT)**
Work flows in Life Science

2010-04: **Joost Geurts (CWI)**
A Document Engineering Model and Processing Framework for Multimedia documents Olga Kulyk (UT) Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05: **Claudia Hauff (UT)**
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06: **Sander Bakkes (UvT)**
Rapid Adaptation of Video Game AI

2010-07: **Wim Fikkert (UT)**
Gesture interaction at a Distance

2010-08: **Krzysztof Siewicz (UL)**
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-09: **Hugo Kielman (UL)**
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-11: **Rebecca Ong (UL)**
Adriaan Ter Mors (TUD) The world according to MARP: Multi-Agent Route Planning

2010-12: **Susan van den Braak (UU)**
Sensemaking software for crime analysis

2010-13: **Gianluigi Folino (RUN)**
High Performance Data Mining using Bio-inspired techniques

2010-14: **Sander van Splunter (VU)**
Automated Web Service Reconfiguration

2010-15: **Lianne Bodenstaff (UT)**
Managing Dependency Relations in Inter-Organizational Models

2010-16: **Sicco Verwer (TUD)**
Efficient Identification of Timed Automata, theory and practice

2010-17: **Spyros Kotoulas (VU)**
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18: **Charlotte Gerritsen (VU)**
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19: **Henriette Cramer (UvA)**
People's Responses to Autonomous and Adaptive Systems

2010-20: **Ivo Swartjes (UT)**
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21: **Harold van Heerde (UT)**
Privacy-aware data management by means of data degradation

2010-22: **Michiel Hildebrand (CWI)**
End-user Support for Access to Heterogeneous Linked Data

2010-24: **Bas Steunebrink (UU)**
The Logical Structure of Emotions Dmytro Tykhonov Designing Generic and Efficient Negotiation Strategies

2010-25: **Zulfiqar Ali Memon (VU)**
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-27: **Ying Zhang (CWI)**
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines Marten Voulon (UL) Automatisch contracteren

2010-28: **Arne Koopman (UU)**
Characteristic Relational Patterns

2010-29: **Stratos Idreos(CWI)**
Database Cracking: Towards Auto-tuning Database Kernels

2010-30: **Marieke van Erp (UvT)**
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31: **Victor de Boer (UVA)**
Ontology Enrichment from Heterogeneous Sources on the Web

2010-32: **Marcel Hiel (UvT)**
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33: **Robin Aly (UT)**
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34: **Teduh Dirgahayu (UT)**
Interaction Design in Service Compositions

2010-35: **Dolf Trieschnigg (UT)**
Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36: **Jose Janssen (OU)**
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-37: **Niels Lohmann (TUE)**
Correctness of services and their composition

2010-38: **Dirk Fahland (TUE)**
From Scenarios to components

2010-39: **Ghazanfar Farooq Siddiqui (VU)**
Integrative modeling of emotions in virtual agents

2010-40: **Mark van Assem (VU)**
Converting and Integrating Vocabularies for the Semantic Web

2010-41: **Guillaume Chaslot (UM)**
Monte-Carlo Tree Search

2010-42: **Sybren de Kinderen (VU)**
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43: **Peter van Kranenburg (UU)**
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44: **Pieter Bellekens (TUE)**
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45: **Vasilios Andrikopoulos (UvT)**
A theory and model for the evolution of software services

2010-46: **Vincent Pijpers (VU)**
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47: **Chen Li (UT)**
Mining Process Model Variants: Challenges, Techniques, Examples

2010-48: **Milan Lovric (EUR)**
Behavioral Finance and Agent-Based Artificial Markets

2010-49: **Jahn-Takeshi Saito (UM)**
Solving difficult game positions

2010-51: **Bouke Huurnink (UVA)**
Search in Audiovisual Broadcast Archives Alia Khairia Amin (CWI) Understanding and supporting information seeking tasks in multiple sources

2010-52: **Peter-Paul van Maanen (VU)**
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention

2010-53: **Edgar Meij (UVA)**
Combining Concepts and Language Models for Information Access

2011-01: **Botond Cseke (RUN)**
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02: **Nick Tinnemeier(UU)**
Work flows in Life Science

2011-03: **Jan Martijn van der Werf (TUE)**
Compositional Design and Verification of Component-Based Information Systems

2011-04: **Hado van Hasselt (UU)**
Insights in Reinforcement Learning: Formal analysis and empirical evaluation of temporal-difference learning algorithms

2011-05: **Base van der Raadt (VU)**
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

2011-06: **Yiwen Wang (TUE)**
Semantically-Enhanced Recommendations in Cultural Heritage

2011-07: **Yujia Cao (UT)**
Multimodal Information Presentation for High Load Human Computer Interaction

2011-08: **Nieske Vergunst (UU)**
BDI-based Generation of Robust Task-Oriented Dialogues

2011-09: **Tim de Jong (OU)**
Contextualised Mobile Media for Learning

2011-10: **Bart Bogaert (UvT)**
Cloud Content Contention

2011-11: **Dhaval Vyas (UT)**
Designing for Awareness: An Experience-focused HCI Perspective

2011-12: **Carmen Bratosin (TUE)**
Grid Architecture for Distributed Process Mining

2011-13: **Xiaoyu Mao (UvT)**
Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14: **Milan Lovric (EUR)**
Behavioral Finance and Agent-Based Artificial Markets

2011-15: **Marijn Koolen (UvA)**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

Titles in the ILLC Dissertation Series:

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ISBN 978-90-814485-5-0