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

Marijn Koolen
University of Amsterdam
m.h.a.koolen@uva.nl

15.04.2011

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. 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.

The value of link evidence for IR was investigated through the TREC Web Tracks of 1999–2004. First through the traditional ad hoc search methodology, with disappointing results. When ad hoc search was considered an inappropriate model for Web search, attention moved to more Web-centric search tasks such as navigational search, with great success. Home pages and other important Web pages tend to be pages with many incoming links, and link evidence such as indegree counts and PageRank proved highly effective.

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 available in its entirety, including all the links, and comes with a large, high-quality test collection in the form of the INEX Ad Hoc collection of 2006–2007. As an encyclopedia, Wikipedia is a natural collection to study topical relevance search tasks.

Link information can be derived from the link graph of the entire collection—giving global, query-independent information—or from a subset of the link graph derived from the top results for a given query produced by a text-based retrieval system, giving local, query-independent link information. PageRank is often computed on the global link graph, while algorithms like HITS are typically used on the local links of the top-ranked results.

We find that for ad hoc search on the INEX 2006 Wikipedia collection, local link evidence leads to significant performance improvements. Incoming and outgoing link information is equally effective; the direction of the links does not affect their value for ad hoc search. We compare this to Web-centric search tasks on the TREC 2004 Web track collection and to the ad hoc task on the TREC 2009 Web track collection. For Web-centric search tasks, such as home page and named page finding, incoming link evidence is more effective than outgoing

link evidence, and global link evidence is more effective than local link evidence. The value of link evidence for Web-centric search is to identify popular and authoritative pages, and the link direction determines the value. For ad hoc search on the more recent ClueWeb09 collection, global outgoing link evidence is very effective, but turns out to favour Wikipedia pages, which are densely interlinked. If we remove Wikipedia from the ClueWeb09 collection, local link evidence is more effective than global link evidence, and incoming and outgoing link evidence are equally effective. We also find that site-internal links are more effective for ad hoc search than site-external links. To determine the impact of link density, we conduct filtering experiments, which show that link density has a minimal impact on the effectiveness of global link evidence. For local link evidence, removing links gradually reduces the impact of link evidence, but even with a small number of links, link evidence is still effective. We also experiment with filtering links based on the semantic relatedness of the link documents and find that links between semantically related pages are more effective for identifying topically relevant documents than links between unrelated pages.

Further experiments on the INEX Wikipedia collection show that the amount of local link evidence is related to the amount of relevant text in documents, and the fraction of local link evidence (the percentage of the global links incident to a document that are present in the local link graph) is related to the fraction of text in a document that is relevant.

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.

Available at `http://staff.science.uva.nl/~mhakoole/2011/kool:mean11.pdf`