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.