The meaning of structure: the value of link evidence for information retrieval
Koolen, M.H.A.

Citation for published version (APA):
Koolen, M. H. A. (2011). The meaning of structure: the value of link evidence for information retrieval
Amsterdam: IR Publications

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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
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
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\footnote{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).} 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
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.