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.
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 be 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 between 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 \textsc{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:

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

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

$$ P_{Q}@n = \frac{1}{Q} \sum_{i=1}^{Q} P_{i}@n $$

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:

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

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

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, 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 result 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 be 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 showing 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 WTI0G collection was still very low, and extracted a subset of the collection, WTI-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. 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
### 2.3 Web Retrieval

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Domain</th>
<th># Pages</th>
<th># Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>vlc2 (wt100g)</td>
<td>1997</td>
<td>.com</td>
<td>18,571,671</td>
<td></td>
</tr>
<tr>
<td>wt2g</td>
<td>1997</td>
<td>.com</td>
<td>247,491</td>
<td>1,166,702</td>
</tr>
<tr>
<td>wt10g</td>
<td>1997</td>
<td>.com</td>
<td>1,692,096</td>
<td>8,062,918</td>
</tr>
<tr>
<td>.gov</td>
<td>2002</td>
<td>.gov</td>
<td>1,247,753</td>
<td>11,110,985</td>
</tr>
<tr>
<td>.gov2</td>
<td>2004</td>
<td>.gov</td>
<td>25,205,179</td>
<td>82,711,345</td>
</tr>
<tr>
<td>ClueWeb09 (B)</td>
<td>2009</td>
<td>.com</td>
<td>50,220,423</td>
<td>1,180,631,904</td>
</tr>
</tbody>
</table>

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


### 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
<table>
<thead>
<tr>
<th>Year</th>
<th>Task</th>
<th>Name</th>
<th># topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>Large Web (ad hoc)</td>
<td>vlc2</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Small Web (ad hoc)</td>
<td>wt2g</td>
<td>50</td>
</tr>
<tr>
<td>2000</td>
<td>Large Web (online services)</td>
<td>vlc2</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Main Web (ad hoc)</td>
<td>wt10g</td>
<td>50</td>
</tr>
<tr>
<td>2001</td>
<td>Web Topic Relevance (ad hoc)</td>
<td>wt10g</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Home page Finding</td>
<td>wt10g</td>
<td>145</td>
</tr>
<tr>
<td>2002</td>
<td>Topic Distillation</td>
<td>.gov</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Named Page Finding</td>
<td>.gov</td>
<td>50</td>
</tr>
<tr>
<td>2003</td>
<td>Topic Distillation</td>
<td>.gov</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Named Page Finding</td>
<td>.gov</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>Home Page Finding</td>
<td>.gov</td>
<td>150</td>
</tr>
<tr>
<td>2004</td>
<td>Topic Distillation</td>
<td>.gov</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Named Page Finding</td>
<td>.gov</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Home Page Finding</td>
<td>.gov</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Ad hoc</td>
<td>.gov2</td>
<td>50</td>
</tr>
<tr>
<td>2005</td>
<td>Ad hoc</td>
<td>.gov2</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Named Page Finding</td>
<td>.gov2</td>
<td>252</td>
</tr>
<tr>
<td>2006</td>
<td>Ad hoc</td>
<td>.gov2</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Named Page Finding</td>
<td>.gov2</td>
<td>181</td>
</tr>
<tr>
<td>2009</td>
<td>Ad hoc</td>
<td>ClueWeb09</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Diversity</td>
<td>ClueWeb09</td>
<td>50</td>
</tr>
</tbody>
</table>

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) pre-computes 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:

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

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 $x(p_i)$ and a hub score $y(p_i)$. 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:

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

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 $x_0(p_i) = y_0(p_i) = 1$. The local graph $G$ 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, Pehchevski 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.