Combining evidence for Web retrieval using the inference network model: an experimental study

Theodora Tsikrika and Mounia Lalmas

Department of Computer Science
Queen Mary University of London
London, E1 4NS, UK.
{theodora, mounia} @dcs.qmul.ac.uk

Abstract

In the Web context, link-based evidence is most commonly used in conjunction with content-based evidential information in order to improve retrieval effectiveness. This paper examines the impact the various types of link-based evidence and their combination with content-based evidence have on the retrieval effectiveness for the topic relevance Web task. The inference network model is used in our study, as it supports the combination of multiple document representations and the combination of multiple results produced by different retrieval strategies. Our experiments indicate hardly any improvements in the effectiveness, similarly to previous TREC results for the topic relevance task. However, they allow us to gain an insight into the behaviour of different types of link-based evidence that could be exploited in the context of other retrieval tasks.

Keywords: Web retrieval, topic relevance task, link-based evidence, inference network model, experiments

1. Introduction

The aim of Information Retrieval (IR) systems is to assist users in satisfying their information needs. In the Web context, different types of users' information needs have been broadly classified on the basis of the type of answer expected (Hawking et al., 2001). TREC's Web Track motivates the introduction of appropriate retrieval tasks, in order to represent these distinct types of information needs. So far, the named page/homepage finding, topic distillation and topic relevance (ad-hoc retrieval) tasks have been explored.

Web IR approaches consider various evidence in order to determine the relevant documents in response to users' information needs. Apart from content-based evidence, they exploit the additional sources of information provided by the Web, mainly the *links* connecting the Web documents. Link-based evidence can be utilised in many ways. One of the most common ones is through the analysis of the Web's link structure. This analysis aims at identifying high quality documents by quantifying their relative importance, with respect to either the rest of the Web documents or the documents associated with the results of a given query. The most popular algorithms developed for this purpose are PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1999). Another way of using link-based evidence is to consider the anchor text associated with the incoming links of Web documents, when indexing them.

Web IR methods, which exploit link-based evidential information and use it in conjunction with content-based evidence, can detect high quality documents (Amento et al., 2000; Silva et al., 2000) and improve the retrieval effectiveness for a variety of Web retrieval tasks (Craswell et al., 2001; Kraaij et al., 2002; Ogilvie and Callan, 2003). However, experimental results in TREC-like settings and in the context of the topic relevance task, indicate that the use of several link-based methods has not led to improvements in the retrieval performance

(Hawking et al., 1999b; Hawking, 2000; Hawking and Craswell, 2001). On the other hand, the use of links appears to be beneficial in other types of tasks, such as the named page finding and homepage finding tasks (Hawking and Craswell, 2001; Craswell and Hawking 2002).

Our aim is to examine what different types of link-based evidence can be exploited by a Web retrieval strategy. We consider the following types of link-based evidence: i) different types of links classified according to their syntax and their semantics, ii) Web document representations generated using link-based evidence and iii) the results obtained by Kleinberg's HITS link analysis algorithm. Our objective is to investigate the impact the various types of link-based evidence have on the retrieval effectiveness, especially when combined with content-based evidence. In this study, we concentrate on the topic relevance task, but the next step of our investigation is to examine similar objectives in the context of other retrieval tasks.

To explore these objectives, we need a framework that allows us to carry out these investigations, by supporting the explicit combination of multiple sources of evidence. We have chosen to employ the inference network model (Turtle, 1990; Turtle and Croft, 1991), in order to examine the combination of content-based and link-based evidence for the topic relevance task in the context of Web retrieval. This model is a well-established probabilistic IR model, which views IR as an inference or evidential reasoning process (van Rijsbergen, 1986), where the probability, that a user's information need is satisfied, can be estimated given a document as evidence. It is based on Bayesian inference networks (Pearl, 1988), which provide a sound framework and allow for the combination of distinct sources of evidence. The main contribution of the inference network model has been the development of a formal retrieval model based on a sound theoretical framework that allows the combination of multiple document representations and information need representations, as well as the combination of the rankings produced by different retrieval algorithms. Moreover, it has been turned into an efficient implementation by the InQuery system (Callan et al., 1994) and has shown significant improvements in the retrieval effectiveness over conventional models, demonstrating, in that way, both its practical and theoretical value.

This paper is organised as follows. Section 2 discusses the related work on Web retrieval and combination of evidence. Section 3 describes the various ways link-based evidence can be utilised. In Section 4, the underlying framework of our approach, based on the inference network model, is presented. Section 5 describes our experimental methodology, evaluation and system details. Section 6 presents the experimental results and their analysis. The paper concludes in Section 7.

2. Related Work

Combination of evidence has been applied in IR, in order to improve retrieval effectiveness and can be distinguished into two separate approaches. The first refers to the development of IR models that explicitly support the combination of multiple document representations and/or information need representations under a single framework. The second refers to the development of models and methods that support the combination of different retrieval systems. This refers to the combination of post retrieval results produced by distinct search strategies, which are usually based on different IR models. This second approach is referred to as *data fusion* in traditional IR environments (Belkin et al., 1995; Savoy et al., 1995) and as metasearch (Selberg and Etzioni, 1997) in the Web context. An overview of the methods of combining document representations and of combining search systems is presented in (Croft, 2000).

The main advantage of the first approach is its flexibility, as it allows the combination of evidence to be performed at document term and/or query term level, rather than only after the retrieval is performed. This permits direct involvement in the generation of document and query representations, the selection of document term and query term weighting schemes and the selection of the ranking algorithm. This is a complete contrast to the data fusion approach, where the inner workings of the different search strategies producing the results to be combined, cannot be modified even if they are known.

First, we focus on the approach of combining multiple document representations. Multiple document representations can be generated in various ways. For instance, manually assigned index terms from controlled vocabularies (Rajashekar and Croft, 1995), passages (Callan, 1994) and phrases (Callan et al., 1995) can be considered for producing alternative document representations. The citations of documents (Turtle, 1990) can also be used. This last approach is of particular interest, since it forms the basis of the methods applied in hyperlinked environments, such as hypertext (Frisse and Cousins, 1989) and the Web. In the context of the Web, this method is equivalent with using the anchor text of the incoming links of a document to index it and has been adopted by a number of Web IR systems (Brin and Page, 1998; Ogilvie and Callan, 2003). Experiments using this link-based representation indicate that this method is particularly beneficial for certain types of Web retrieval tasks, such as named page finding (Craswell and Hawking, 2000; Hawking and Craswell, 2001; Ogilvie and Callan, 2003).

Combination of document representations can be achieved in the context of a single model. In a traditional IR environment, (Rajashekar and Croft, 1995) used the inference network model as the underlying framework for combining multiple document representations, generated by using automatic and manually assigned terms. Their results indicated improvements in the effectiveness. More recently, and in the context of the name page and homepage finding Web retrieval tasks, (Ogilvie and Callan, 2003) investigated the pre-conditions for successful combination of content and link-based documents representations. The combination was achieved using a mixture-based language model, which performed the combination at query term level.

The combination of content and link based evidence can also be performed in the context of the combination of retrieval results from link analysis algorithms and content-based representations. The most popular link analysis algorithms are: PageRank (Brin and Page, 1998) and HITS (Kleinberg 1999). PageRank computes a query independent measure of a Web document's citation importance that corresponds well with people's subjective idea of importance and authority. PageRank can also be thought of as a model of user behaviour, that reflects the probability that a random surfer visits a particular Web document, in the Markov chain induced by the Web graph. HITS algorithm classifies Web documents into two categories, the hubs and the authorities, with respect to a particular topic. A Web document is an authority, a high quality document, if it contains clear, accurate and useful information on the topic and a hub if it links to many good authorities. Both these approaches have been extensively studied and modified (Bharat and Henzinger, 1998; Chakrabarti et al., 1998; Haveliwala, 2002). The aim of all these link analysis algorithms is to combine the results they produce with those obtained from content-based methods in order to improve precision in the top retrieved documents.

The experimental results from combination of content and link based evidence in the form of the combination of retrieval results from link analysis algorithms and content-based representations, especially in TREC-like settings, indicate that the additional incorporation of link-based evidence is not beneficial for the topic relevance task (Hawking et al., 1999b; Hawking, 2000; Hawking and Craswell, 2001). It does, however, improve the effectiveness for named page/homepage finding task (Hawking and Craswell, 2001; Craswell and

Hawking, 2002). The results for the topic distillation task are still inconclusive (Craswell and Hawking, 2002).

The combination of evidence can be performed using various underlying frameworks (Croft, 2000). We are focusing on those based on Bayesian inference networks (Pearl, 1988). In this study, we employ the inference network model (Turtle, 1990; Turtle and Croft, 1991), discussed in Section 4. However, there are two more main retrieval models based on Bayesian networks. The first one is the belief network model (Ribeiro-Neto and Muntz, 1996), which is derived from probabilistic considerations over a clearly defined sample space, consisting of the keywords in the collection. This model can subsume any of the classic models and its main strength is the easy incorporation of additional sources of evidence, such as knowledge gathered form past user sessions (Ribeiro-Neto et al, 2000). It has also been applied in the Web (Silva et al, 2000), where the results from Kleinberg's HITS algorithm constituted the additional sources of evidence. The second model is the Bayesian network model (de Campos et al., 2002) used to compute the probability of relevance in a collection given a query. This model is accompanied by an inference mechanism for exact propagation of probabilities.

In this paper, our aim is to have a closer look at the combination of content and link-based document representations and at the combination of the results from link analysis algorithms and content-based representations, in the context of the relevance topic task. In particular, we aim at investigating what types of link-based evidence can be used and how they can be exploited. We use the inference network model to carry out this investigation, as it supports the explicit combination of multiple sources of evidence.

3. Link-based evidence

One of the most important features of the World Wide Web is the presence of links between Web documents. These links can be considered as sources of evidential information that, in conjunction with content-based evidence, could be utilised in support of Web IR methods. This section discusses the link-based evidence examined in this work. Section 3.1 presents a syntactic and semantic classification of Web links. This classification is employed in the generation of different link-based document representations, which are described in Section 3.2. Finally, Section 3.3 describes Kleinberg's HITS algorithm, which performs analysis of the Web link structure.

3.1 Web links: syntax and semantics

The source of link-based evidential information is the mere presence of a link between two Web documents, which signifies the existence of some kind of relationship between them. The precise nature of this relationship, however, is not usually examined. In this way, the considerable amount of human judgement and intent encoded in the creation of Web links is overlooked. Our aim is to exploit this additional source of information by considering a classification of Web links, reflecting the type of the relationship that exists between the documents they connect.

Two taxonomies are described, distinguishing Web links with respect to either their *syntax* or their *semantics*. The syntactic analysis is based on the relationship of the source and target Web documents in terms of their relative location in the Web structure, i.e. if they belong to the same domain or to the same site, etc. The semantic analysis, on the other hand, aims at capturing the reasons a link exists and how the observation of the target document affects our knowledge or understanding of the source document.

The syntactic analysis of links presented here is based on the one discussed in (Géry and Chevallet, 2001). Links can be classified as follows:

- *Same page* : when the link points to the Web document it belongs to.
- *Hierarchical* : when the source and target Web documents belong to the same directory path. These types of links can be further categorised as:
    - *Horizontal* : when the source and target documents belong to the same directory.
    - *Up* : when the source document is *deeper* in the directory path than the target.
    - *Down* : when the source document is *higher* in the directory path than the target.
- *Transversal* : when the target Web document is neither in the ascendant nor in the descendant directories; however, it still belongs to the same site as the source Web document.
- *Inter-host* : when the source and target documents belong to different Web domains.

The semantic analysis allows us to reflect the relationships these links establish between the Web documents they connect. Following (Géry and Chevallet, 2001), we consider three fundamental relationships between Web documents:

- *Composition* : this relation expresses the forming of an entity consisting of simpler components. For instance, a book consists of many chapters, which in turn consist of sections and subsections. If a Web document corresponds to a *section*, a *chapter* can be defined as a composition of several Web documents, usually organised in a hierarchical manner.
- *Sequence* : this relation expresses the order in which some Web documents are organised. For instance, *chapter two* of a book precedes *chapter three* and follows *chapter one*. This order provides the users a path to follow in order to achieve better and more coherent understanding of the document in question.
- *Reference* : this relation expresses the fact that the source and target documents are similar in some way. For instance, the author of the source document might have found the target document to be valuable source of information on a topic of interest. On the other hand, though less likely, a document might be referenced, if it provides a counter-argument to what is claimed in the source document.

The syntactic relation a Web link captures can be easily detected, whereas its semantics cannot be so readily determined. Therefore, the aim is to be able to define a link's semantics given only its syntactic specification. To achieve that, we follow the approach and assumptions adopted in (Géry and Chevallet, 2001). Therefore, we consider the composition relation to be expressed by the *hierarchical down* links and the sequence relation by the *hierarchical horizontal* links[1]. Finally, we consider the notion of the broader reference relation to be captured both by the *inter-host* links and by the *transversal* links.

3.2 Link-based document representations

Different document representations can be generated by considering, each time, a different subset of the available information regarding the document, for instance only the title of the document or its abstract. Web documents, however, are associated with much more information, namely the links pointing to them, referred to as inlinks, and the links contained in them, referred to as outlinks. This section presents how several document representations can be generated by exploiting the inlink and outlink-based information.

[1] In reality, true sequence, between Web document belonging to the same Web site, is a much narrower relation than that expressed by the *horizontal* links. Therefore, additional information obtained, for instance, from a traversal algorithm applied to the Web site's graph, would be required in order to determine, in a stricter manner, which of the *horizontal* links actually reflect the sequence relation.

To represent a Web document using its inlinks the *anchor text*, which is the text associated with each link can been used. The basis of this approach lies on the assumption that the anchor text usually provides a more accurate and concise description of the Web document that it is associated with, than the actual Web document itself, by probably using more significant terms than those contained in the document (Bharat, and Henzinger, 1998). This assumption forms the foundation of many approaches (Amitay, 1998; Amitay, 2000; Brin and Page, 1998; Chakrabarti et al., 1998) and it has been tested empirically on a relatively large Web corpus (Davison, 2000). The results of these experiments indicated that this idea holds true and therefore anchor text can be used to describe the page it references. Furthermore, instead of considering only the anchor text, its surrounding text can also be taken into account. This technique is applied in order to deal with situations where the anchor text is not sufficiently self-descriptive. It is implemented by taking into account the text inside a window of W bytes around the anchor text, called the *anchor window* (Chakrabarti et al, 1998). Typical values of W are 50 – 100 bytes.

As discussed in the previous section, links express different kinds of relationships and can be classified according to their semantics. This additional source of evidence can be exploited by generating various inlink-based document representations, each considering only a specific type of link. Therefore, we will use in total four inlink representations, one considering all incoming links, denoted as *inlink* representation and three considering only inlinks of a specific type, denoted as *inlink composition*, *inlink reference* and *inlink sequence* representations.

In a similar manner, each Web document could be represented by the anchor text of the outlinks it contains. The rationale of this approach lies on the assumption that the text associated with the outlinks of a document could reflect its major topics. As before, four outlink based representations can be generated. These are denoted as *outlink*, *outlink composition*, *outlink reference* and *outlink sequence* representations.

3.2 Link analysis

(Kleinberg, 1999) proposed the HITS algorithm, which exploits the information inherent in the Web links to find the most "important" documents among the thousands of relevant Web documents returned in response to broad-topic queries. There are two measures reflecting the importance of a Web document: its degree of *authority* and its degree of *hub*. The hubs are Web documents pointing to many useful authorities on a specific topic, whereas the authorities are Web documents pointed by many hubs on the same topic. Moreover, there is a mutual reinforcement relation between authorities and hubs: good authorities are pointed to by good hubs and good hubs point to many good authorities.

The algorithm computes these two qualitative measures for a set of Web documents related to a specific query. Initially, a set of Web documents called the *root set* is formed. This set contains the k top-ranked documents retrieved by a content-based approach. This set is then expanded by adding the pages pointing to or pointed to by the documents of the root set. The iterative algorithm used in estimating the hub and authority values of a Web document is presented in detail in (Kleinberg, 1999). Briefly, the adjacency matrix A of the graph that corresponds to the expanded set is created and the principal eigenvectors of the matrices AA^T and A^TA are computed. The component of each document in the principal eigenvector of AA^T corresponds to its hub value, while its component in the principal eigenvector of A^TA corresponds to its authority value. Variations and refinements of the initial HITS algorithms have also been applied, including the incorporation of keyword-based evidence in the computation of these two metrics (Bharat and Henzinger, 1998; Chakrabarti et al., 1998).

4. Combination of evidence using the inference network

This section briefly describes the *inference network model*, the framework of our investigation. It determines how the document and queries are represented and how the actual retrieval is performed. More details of the inference network model can be found in (Turtle, 1990; Turtle and Croft, 1991). In addition, our use of the inference network model, as our underlying framework, for supporting the explicit combination of multiple sources of evidence is presented in (Tsikrika and Lalmas, 2002).

The inference network model is a probabilistic model based on Bayesian inference networks (Pearl, 1988). A Bayesian network is a directed acyclic graph (DAG) in which the nodes represent variables and the arcs signify causal dependencies between the variables they connect. The strengths of these dependencies are expressed by conditional probabilities. If there is a directed arc from node *p* to node *q*, then *p* is the *parent* node, *q* is the *child* node and *p* is considered to '*cause*' *q*. This causal dependency is quantified by the conditional probability *P(q|p)*. The nodes that do not have parents are referred to as the *root* nodes of the network.

The inference net model used in our investigation is illustrated in Figure 1, emphasising its ability to support the combination of multiple document representations. It consists of a *document network* and of a *query network*. The document network is built once and represents the collection of Web documents, indexed using a number of representation techniques. The query network is built every time a request is submitted to the system and represents the user's information need expressed as one or more queries. All the nodes on the inference network are either *true* or *false* and they represent the random variables that are associated with the documents, the queries and the concepts used to represent them. The values of the root nodes can be determined by observation, whereas the values of the rest of the nodes are determined by inference.

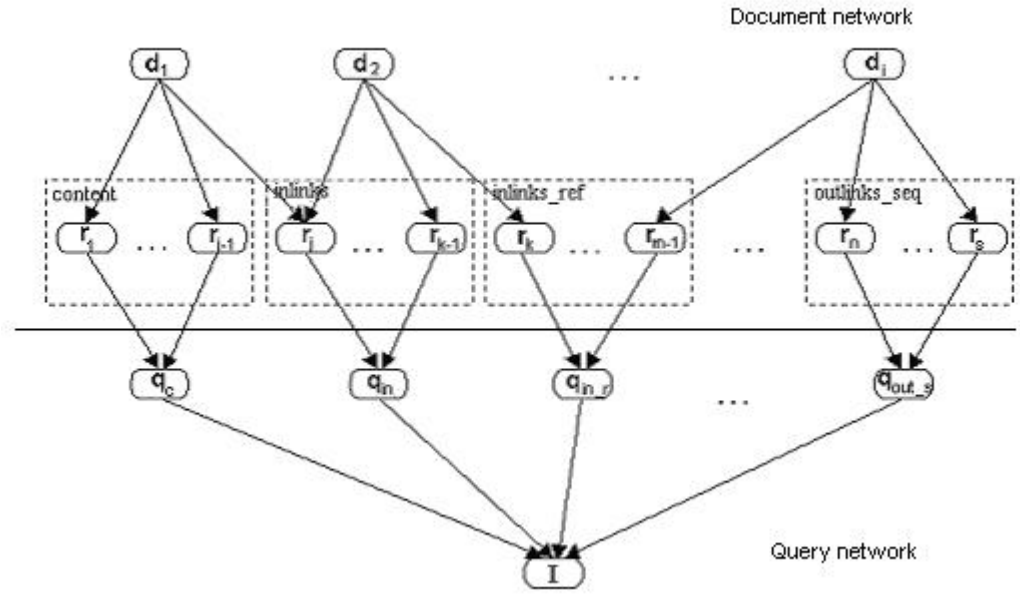

Figure 1: The inference network model

4.1 Document network

The *document network*, shown in Figure 1, is a simple two-level DAG containing two different types of nodes: the document nodes (d_i's) and the concept representation nodes (r_k's). The document nodes correspond to the event that a particular Web document has been observed. The prior probabilities $P(d_i)$ associated with each document d_i are generally set to be equal to $1/n_d$, where n_d the number of documents in the collection. A content representation node r_k represents the proposition that a concept has been observed. When multiple document representation schemes are supported, the set of representation concepts is divided into as many disjoint subsets, as the number of representation schemes (Turtle, 1990; Turtle and Croft, 1991). Each of these subsets is associated with an individual representation technique and contains the representation concepts used to index the documents, using this particular representation scheme. Our approach supports the content-based representation scheme together with the eight link-based representations introduced in Section 3.2. When a concept r_k is assigned to a document d_i, this assignment is represented by a directed arc. The value of the arc is the conditional probability $P(r_k|d_i)$, which is specified given the set of parent nodes and their influence on that concept.

4.2 Query network

The *query network*, as shown in Figure 1, consists of two levels: the queries level and the information need level. A query node represents a specific query formulation and corresponds to the event that this query representation is satisfied. Each query formulation node is associated with one of the representation schemes supported and is constructed using the representation nodes of the corresponding representation technique. If we consider, for example, the *content* and *inlink reference* representation schemes, there will be two query nodes $q_{content}$ and $q_{inlinks\_ref}$. These query formulations are generated using the query terms corresponding to the representation concepts belonging, respectively, to the *content* and *inlink reference* subsets. The combination of these query formulations can be viewed as the combination of the different document representations. Therefore, the information need node, represented by node *I*, corresponds to the proposition that the combined query has been satisfied and the user's information need is met.

The query and document networks are connected by the directed arcs from the content representation nodes to the query nodes. To produce a ranking of the documents in the collection with respect to a given information need *I*, we compute the probability that this information need is satisfied given that document d_i has been observed, $P(I|d_i)$. This is referred to as *instantiating* d_i and corresponds to attaching evidence to the network, by stating that d_i = *true*, whereas the rest of the document nodes are set to *false*. When the probability $P(I|d_i)$ is computed, this evidence is removed and a new document d_j, $j \neq i$, is instantiated. By repeating this computation for the rest of the documents in the collection the ranking is produced.

4.3 Conditional Probabilities

For any of the non-root nodes *A* of the network, the dependency on its parent nodes $\{P_1, P_2, \ldots, P_n\}$, quantified by the conditional probability $P(A| P_1, P_2, \ldots, P_n)$, must be estimated and encoded. In principle, this estimate would require $O(2^n)$ space for all combinations of parent values of a node with *n* parents, since we are dealing with binary propositions. However, by using c*anonical link matrix* forms (Pearl, 1988) to encode this estimate, the space complexity is reduced to $O(n)$ and the derived link matrix is evaluated using closed form expressions, in

order to compute our belief in *A* or *P(A=true)*. The derivations of these expressions and their use in the inference network model are discussed in (Turtle, 1990).

Therefore, first of all, we need to provide an estimate that characterises the dependence of the representation concepts associated with a particular document representation, upon the Web document containing them. Experiments reported in (Turtle, 1990; Turtle and Croft, 1991) indicated that a good belief estimate is achieved by employing the *tf, idf* weighting strategies. This estimate is given by:

$$P(r_k \mid d_i = true) = 0.4 + 0.6 \times tf_{r_k,d_i} \times idf_{r_k}$$

$$P(r_k \mid all\ parents\ false) = 0.4$$

where $tf_{rk,di}$ and idf_{rk} are calculated within the context of an individual document representation using any standard *tf, idf* weighting scheme.

Next, we need to encode the dependency of each individual query formulation q_s, upon the representation concepts associated with a particular representation scheme. We use canonical link matrix forms to encode this probability and more specifically the *weighted-sum* canonical link matrix form, as described in (Turtle, 1990; Turtle and Croft, 1991). This allows us to assign a weight w_{r_j} to each of the *n* parents of the query node q_s, reflecting their influence on q_s – the parents with larger weights have more influence on our belief. The belief in q_s is then determined by the parents that are true and evaluated as:

$$P_{wsum}(q_s = true) = \frac{\sum_{j=1}^{n} w_{r_j} p_{r_j}}{\sum_{j=1}^{n} w_{r_j}} \qquad \text{where } p_{r_j} = P(r_j = true) \qquad (1)$$

This allows us to estimate our belief in each of the query formulation nodes, associated with each individual representation scheme.

Finally, the information need node *I* combines all of the evidence from the query representation nodes and can be viewed as a way of forming a query that is a composite of the individual query formulations connecting to it. Since a link matrix form can be considered as an operator for combining evidence, this combination can be performed using a *weighted-sum* link matrix as before. The weights w_{q_j} express the importance of each query representation and consequently reflect the importance of the representation scheme associated with that particular query. The belief in *I* is then evaluated as:

$$P_{wsum}(I = true) = \frac{\sum_{j=1}^{n} w_{q_j} p_{q_j}}{\sum_{j=1}^{n} w_{q_j}} \qquad \text{where } p_{q_j} = P(q_j = true) \qquad (2)$$

Instead of using canonical link matrix forms, the full link matrix could be employed, if the number of parents is small and the dependence of a node on its parents does not fit a canonical form.

5. Experimental set-up

In the section, we describe the set-up of our experiments. Our experiments are performed using the InQuery retrieval system (Callan et al., 1994), an efficient and flexible implementation of the Bayesian inference network framework. InQuery has been extensively

documented in the literature (Allan et al., 1999; Broglio et al., 1994; Callan et al., 1994). We have customised and used InQuery v3.2, made available to us by the University of Massachusetts. The test collection used is discussed in Section 5.1. The indexing, retrieval and evaluation methodologies adopted in our experiments are described in Sections 5.2 to 5.4, respectively.

5.1 Test collection

There are two standard test collections that can be employed for performing experiments to evaluate Web IR approaches with respect to the topic relevance task. These are the WT2g and the WT10g test collections, both constructed and used in the context of recent research exercises organised by TREC's special interest Web Track. WT2g consists of 0.25 million documents (2GB) and was used in TREC-8 (Hawking et al, 1999b), whereas WT10g contains 1.69 million documents (10GB) and was employed in TREC-9 (Hawking, 2000) and TREC 2001 (Hawking and Craswell, 2001). For our series of exploratory experiments, we are using the WT2g test collection. WT2g has been criticised (Singhal and Kaszkiel, 2001), for what are considered to be its deficiencies, namely its small size compared to that of the publicly indexable Web and the low number of inter-host links. Despite these shortcomings, the experimental results using WT2g provide us with indications on the behaviour of various types of link-based evidence and their combination with content-based evidential information.

5.2 Indexing

The 247,491 Web documents of the WT2g corpus are indexed using InQuery and a separate index is generated for the content-based and for each of the eight link-based representation schemes introduced in Section 3.2. These representation schemes are listed in Table 1, together with the abbreviations we use to refer to them in the remainder of the paper. The table also lists the number of documents indexed by each scheme and their proportion with respect to the total number of documents in the collection. Finally, the last column lists, separately for the inlink and outlink representations, the proportion of the documents of each of the link-based representations generated by considering only a specific type of links, with respect to the total number of documents in the corresponding link-based representation considering all types of links.

Table 1: List of the nine individual document representations together with number of documents indexed by each representation, and their percentage with respect to the whole collection.

Representations	**Abbreviations**	**Documents in WT2g**		
Content	C	247,491	100.00 %	
Inlink	IN	221,885	89.65%	100.00 %
Inlink composition	IN_{comp}	34,069	13.76%	15.35%
Inlink reference	IN_{ref}	161,350	65.19%	72.72%
Inlink sequence	IN_{seq}	55,097	22.26%	24.83%
Outlink	OUT	213,058	86.09%	100.00 %
Outlink composition	OUT_{comp}	11,031	4.46%	4.77%
Outlink reference	OUT_{ref}	141,062	57.00%	66.21%
Outlink sequence	OUT_{seq}	142,006	57.38%	66.65%

For each of the eight link-based document representations, additional separate indexes are created by considering an *anchor window* of *W* bytes of surrounding text. We experiment with the following values of *W*, *W* = 25, 50, 75 and 100 bytes. The link-based document representation generated by considering an anchor window of *W* is denoted as *representation_W* (for instance: IN_50).

InQuery uses the following belief function (Allan et al., 1999) to estimate the belief in concept *t* within document *d*:

$$w_{t,d} = 0.4 + 0.6 \times \frac{tf_{t,d}}{tf_{t,d} + 0.5 + 1.5\frac{length(d)}{avglen}} \times \frac{\log\frac{N+0.5}{n_t}}{\log N + 1}$$

where n_t is the number of document containing term *t*, *N* is the number of documents in the collection, *avglen* is the average length (in words) of documents in the collection, *length(d)* is the length in words of document *d* and $tf_{t,d}$ is the number of times term *t* occurs in *d*. The same approach is used for all the representations, so that the only difference among them is the use of either content-based or link-based evidence.

The indexing phase also allows us to estimate statistics regarding the different types of links of the WT2g dataset. Table 2 presents the results of the syntactic and semantic analysis of the different types of links, for outlinks whose source is within WT2g, and for inlinks, whose both the source and target lie within WT2g. The low number of inter-host inlinks, already reported elsewhere (Hawking et al., 1999) is evident, whereas the *horizontal* and *transversal* appear to be the most common types of links. For the semantic analysis, same page inlinks and outlinks are ignored, since we are examining the semantic relation expressed by inter-page links only. The *composition* relation is expressed by the *down* links, when considering the outlinks of a document and by the *up* links when considering its inlinks.

Table 2: Percentage of links according to syntactic and semantic analysis

		Outlinks		Inlinks	
		Syntax	Semantics	Syntax	Semantics
	Same page	14.04 %	-	19.45 %	-
Sequence	Horizontal	32.03 %	40.07%	40.87 %	56.55%
Composition	Down	3.25 %	4.07%	8.28 %	-
	Up	6.03 %	-	3.16 %	4.37%
Reference	Transversal	29.42 %	55.86%	27.76 %	39.08%
	Inter-host	15.22 %		0.48 %	
		100.00%	100.00%	100.00 %	100.00%

5.3 Retrieval

We use the 50 TREC topics employed in TREC-8 and associated with the WT2g dataset, topics 401-450. Only the *title* field of the TREC topics is used in the experiments reported here, since it is considered to be representative of real Web search queries (Hawking, 2000; Hawking and Craswell, 2000).

The content representation is used as a baseline for all the conducted experiments. The individual document representations are combined as explained in Section 4.3. There are several factors that could affect the effectiveness of the combination. Here we focus on three of them: i) the parameters of the combination, ii) the number *X* of top-ranked documents retrieved by each document representation and included in the combination and iii) the combination method employed. We examine each factor separately.

The combination parameters refer to the weights w_{q_j} of formula (2), reflecting the contribution of each document representation. They are chosen in an ad-hoc manner and

therefore, we do not claim them to be optimal. Instead of reporting on the actual values of these parameters, we present their ratio. For instance, when combining the Content with the IN representation with $w_{Content}$ and w_{IN} their respective weights, we report on their ratio $w_{Content}$: w_{IN} as being, for example, 10:1. Furthermore, the combinations are performed using the *X* top ranked documents from the result list generated by each document representation, where X = 10, 20, 30, 50, 100 and 1000.

The combination itself is performed by applying formula (2). There are two combination approaches that could be followed: one that considers the retrieval scores of the retrieved documents and one that considers their ranks. In the first case, the beliefs p_{qj} correspond to the retrieval scores of the documents and the combination can be performed either by normalising them, if we assume that they not are comparable (Lee, 1997), or not. In the second case, we use the ranks instead of the retrieval scores, under the assumption that the retrieval scores obtained are not comparable, since they are probably calculated using different ranking functions on different corpus statistics. The values used in the combination are either the reciprocal ranks of the documents (Zhang et al., 2002) or the normalised ranks of the documents (Lee, 1997). The former method is biased towards the higher-ranked documents, whereas the later method produces a more uniform distribution. In summary, there are four combination methods applied, each considering either i) the unnormalised retrieval scores of the documents or ii) the normalised retrieval scores of the documents or iii) the reciprocate ranks of the retrieved documents or iv) the normalised ranks of the retrieved documents. The normalisation is linear, so that the first document in each list has a score of 1 and the *X*th document has a score of zero.

The application of HITS algorithm results in the estimation of the hub and authority values of the documents initially retrieved by the *content* representation scheme. The root set is constructed by considering the top 50, 100 and 200 documents. The expansion step considers all the outlinks and 50 of the inlinks of each document in the root set. Three different approaches are applied when performing the expansion. In the first case, all the types of inlinks and outlinks are taken into account, in the second case we restrict our expansion by using only the inter-host type links, whereas in the third case we use the reference (inter-host and transversal) type links. The combination of the results of the content-based approach and the hubs and authorities is performed using the normalised scores combination method.

5.4 Evaluation

For the evaluation of the results of the experiments, we focus on precision, since in the context of the Web, users are not interested in recall (Hawking et al., 1999a), especially since thousands of documents are usually returned in response to a broad topic query. For each of the results we report *precision at n retrieved documents* (P@n), where n = 5, 10, 15 and 20.

6. Results and analysis

Extensive experiments were carried out evaluating the effectiveness of all individual representations and of all their possible combinations. The most important findings are reported and analysed in this section. Section 6.1 and Section 6.2 present the results and analysis of the individual document representations and of the combination of document representations, respectively. Section 6.3 presents the results of the application of HITS algorithm and their combination with the content-based representation. (Values in bold in the following tables signify improvements in the effectiveness).

6.1 Individual representations

We start by analysing the retrieval results of each of the nine document representations schemes presented in Table 1. For each individual representation, Table 3 lists the number of queries (out of the total 50 submitted queries) for which at least one document is retrieved and also the numbers of retrieved, relevant and retrieved relevant documents corresponding to these queries.

Table 3: Number of queries, retrieved, relevant, and retrieved and relevant documents for the nine individual document representations.

	Queries	**Retrieved**	**Relevant**	**Retrieved relevant**
Content	50	103839	2279	1700
IN	50	31282	2279	353
IN_{comp}	48	3340	2182	20
IN_{ref}	49	6924	2273	67
IN_{seq}	50	19587	2279	239
OUT	50	50747	2279	173
OUT_{comp}	48	3827	2182	8
OUT_{ref}	50	27564	2279	61
OUT_{seq}	50	26467	2279	63

By examining this table, it is apparent that the content appears to be the stronger evidence for the topic relevance task, since the content representation retrieves a much higher number of relevant documents than any other representation. By comparing the number of relevant documents retrieved by the IN and OUT representations, there is a first indication that the inlinks are stronger evidence than the outlinks for the generation of link-based representations. Regarding the document representations generated by considering the different types of links, three out of six (IN_{comp}, IN_{ref}, OUT_{comp}) fail to return documents for all the submitted queries. Furthermore, the low numbers of relevant documents retrieved by IN_{comp} and OUT_{comp} demonstrate that composition type links are probably the weaker form of evidence out of the three types of links considered. Following a similar analysis, the sequence type links appear to be the strongest evidence, especially in the case of the inlink representation, with the reference type links coming next.

Table 4: Precision values for the nine individual document representations

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
IN	0.1840	0.1340	0.1160	0.0940
IN_{comp}	0.0458	0.0271	0.0181	0.0146
IN_{ref}	0.0653	0.0449	0.0354	0.0265
IN_{seq}	0.1680	0.1180	0.0947	0.0800
OUT	0.0960	0.0620	0.0507	0.0480
OUT_{comp}	0.0125	0.0083	0.0056	0.0042
OUT_{ref}	0.0400	0.0260	0.0187	0.0140
OUT_{seq}	0.0200	0.0240	0.0227	0.0190

The above observations are confirmed by examining Table 4, which presents the effectiveness of each of the aforementioned representations. It is evident that the content representation performs significantly better than any other of the individual representations. We could compare its retrieval performance with that of the runs submitted by other research groups in TREC-8, where WT2g was used. However, this comparison would not be meaningful, since

most of the participants in TREC-8 utilised more than the title field of the queries (namely the title + description fields) and applied relevance feedback and query expansion techniques (Hawking et al., 1999b).

The IN representation performs significantly better than the OUT representation. This indicates that the anchor text is more descriptive for the pages it references than for the pages that contain it. The representations generated using the inlinks are superior to those generated using the outlinks, not only when all links are taken into account (IN vs. OUT), but also in the individual cases when specific types are considered (IN_{comp} vs. OUT_{comp}, etc.).

The IN_{seq} representation performs significantly better than IN_{comp} and IN_{ref}. The same applies for the outlink representations generated when considering the types of links. This indicates that maybe the sequence type links provide the strongest evidence. By examining Table 2, which presents the percentage of types of links in the collection, we observe that the sequence type are the most prevalent inlinks and the second most prevalent outlinks. This does not allow us to draw any conclusions in terms of whether the most common types of links also constitute the strongest evidence. However, the fact that sequence type inlinks are the most commonly observed within WT2g may be due to the low number of inter-host inlinks (Hawking et al., 1999b), as illustrated in Table 2. It would be interesting to see whether sequence type links are still the strongest evidence even in collections where the reference (inter-host + transversal) type links prevail.

Next, we examine how the retrieval effectiveness is affected when taking into account an anchor window of *W* bytes surrounding the anchor text. Table 5 presents the results for both the inlink and outlink representations when 25, 50, 75 and 100 bytes of surrounding text are considered. The link-based representations generated when only the anchor text is considered are our baselines and are denoted by setting $W = 0$.

Table 5: Effect of window size of the anchor text

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
		IN		
W = 0	0.1840	0.1340	0.1160	0.0940
W = 25	**0.1880**	**0.1480**	0.1107	**0.0960**
W = 50	**0.1960**	**0.1460**	**0.1187**	**0.0990**
W = 75	**0.1920**	**0.1400**	**0.1173**	**0.1000**
W = 100	**0.1920**	**0.1440**	**0.1200**	**0.1040**
		OUT		
W = 0	0.0960	0.0620	0.0507	0.0480
W = 25	**0.1120**	**0.0720**	**0.0560**	**0.0480**
W = 50	**0.1400**	**0.0860**	**0.0733**	**0.0610**
W = 75	**0.1280**	**0.0940**	**0.0733**	**0.0690**
W = 100	**0.1400**	**0.1000**	**0.0840**	**0.0700**

Table 6: Retrieved relevant documents for different window sizes

	Retrieved relevant	
	IN	**OUT**
W = 0	353	173
W = 25	400	255
W = 50	430	300
W = 75	454	361
W = 100	473	409

The effectiveness improves as the size of the anchor window increases, particularly for the case of the outlinks. Furthermore, more relevant documents are retrieved when more of the

surrounding text is considered, as illustrated in Table 6. Analogous results, not reported here, are also observed for the inlinks and outlinks representations generated by considering the specific types of links. These results indicate that the standard technique applied in Web IR of considering also the surrounding text is justified. Therefore, for all the combination experiments discussed next, we decide to use the most effective inlink and outlink representations and these are the ones for which *W* is set to 100.

6.2 Combination of document representations

This section discusses the combination of content-based and link-based evidence in terms of the combination of content-based and link-based representations. We examine the effect this combination has on the effectiveness compared to that of the content only approach.

Initially, we concentrate on the two combinations of the content representation with the best performing inlink and outlink based representations, namely IN_100 and OUT_100. The parameters in these combinations determine the relative weights assigned to the individual document representations. We experiment with various values for the ratios $w_{Content}$: w_{IN} and $w_{Content}$: w_{OUT}. Here, we report on some indicative values of 10:1, 5:1 and 1:1. Table 7 lists the results for the two combinations for one specific case, when the top 20 ranked documents are considered and the combination method utilises the unnormalised scores of the retrieved documents.

The ratios 10:1 and 5:1 perform better than the 1:1 indicating that stronger evidence should be placed on the content-based representation compared to the link-based ones, which by itself is not surprising. Further examination of experimental results not reported here, for all possible two-way combinations of content with any link-based representation, for different numbers of top ranked documents and different combination methods, reveal that the 10:1 ratio is the best. Therefore, it will be the one considered for the combinations investigated in the rest of our analysis.

Table 7: Combination of Content+IN document representations for X = 20 top-ranked documents using the unnormalised scores

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
C+IN_100 (10:1)	**0.3960**	0.3620	**0.3560**	**0.3360**
C+IN_100 (5:1)	**0.3880**	0.3620	**0.3560**	**0.3360**
C+IN_100 (1:1)	0.3320	0.2960	0.2920	0.2730
C+OUT_100 (10:1)	0.3360	0.3400	0.3427	**0.3360**
C+OUT_100 (5:1)	0.3240	0.3380	0.3427	**0.3360**
C+OUT_100 (1:1)	0.2840	0.3040	0.3160	0.2360

Next we investigate how the effectiveness is influenced when different numbers of top ranked documents are included in the combinations. Tables 8, 9, 10 and 11 present the effectiveness of the C + IN_100 and C + OUT_100 combinations when the combination method considers the unnormalised retrieval scores, the normalised retrieval scores, the reciprocate ranks and the normalised ranks, respectively, of the retrieved documents. As discussed above, the ratio of combination weights is set to 10:1, and the comparisons are performed for the values of precision at 20, unless otherwise stated.

By comparing tables 8, 9, 10 and 11, the first observation is that, when the 10 top ranked documents are considered, the precision at 5 documents generally improves or remains the same as the baseline, at 10 documents it drops slightly, and at 15 and 20 documents, it drops dramatically. This is due to the fact that since only 10 documents are considered from each

representation, the combined result list is more likely to contain, after rank 10, the results from the weaker link-based representation. For the rest of the analysis, the combinations where only the 10 top ranked documents are considered, are ignored.

Table 8: Content+IN and Content+OUT with 10:1 ratio of combination weights using the unnormalised scores

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
		C + IN_100		
Top 10 docs	**0.3880**	**0.3680**	0.2773	0.2170
Top 20 docs	**0.3960**	0.3620	**0.3560**	**0.3360**
Top 30 docs	0.3640	0.3440	0.3440	0.3340
Top 50 docs	0.3640	0.3280	0.3200	0.3130
Top 100 docs	0.3640	0.3340	0.3093	0.2840
Top 1000 docs	**0.3680**	0.3540	0.3027	0.2630
		C + OUT_100		
Top 10 docs	**0.3840**	**0.3680**	0.2707	0.2100
Top 20 docs	0.3360	0.3400	0.3427	**0.3360**
Top 30 docs	0.3320	0.3060	0.3227	0.3240
Top 50 docs	0.3200	0.2820	0.2920	0.2990
Top 100 docs	0.3080	0.2540	0.2267	0.2307
Top 1000 docs	0.3400	0.2760	0.2307	0.2110

Table 9: Content+IN and Content+OUT with 10:1 ratio of combination weights using the normalised scores

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
		C + IN_100		
Top 10 docs	**0.3680**	0.3320	0.2587	0.2170
Top 20 docs	**0.3680**	0.3660	0.3480	0.3100
Top 30 docs	**0.3720**	0.3660	**0.3547**	0.3330
Top 50 docs	**0.3800**	0.3660	**0.3573**	0.3330
Top 100 docs	**0.3840**	**0.3680**	**0.3547**	0.3350
Top 1000 docs	**0.3800**	0.3660	**0.3573**	**0.3370**
		C + OUT_100		
Top 10 docs	0.3640	0.3240	0.2480	0.2100
Top 20 docs	0.3640	0.3660	0.3493	0.3130
Top 30 docs	**0.3680**	0.3660	**0.3547**	0.3340
Top 50 docs	**0.3680**	0.3680	**0.3547**	0.3340
Top 100 docs	0.3640	0.3660	**0.3533**	0.3340
Top 1000 docs	0.3560	0.3580	**0.3520**	0.3240

For both the combination methods that do not rely on normalised values, the best results are observed when the 20 top ranked documents are considered. This is observed for both C + IN_100 and C + OUT_100 combinations. Furthermore, for both these methods, the precision drops as more documents are included and this drop is more dramatic for the case of unnormalised scores. For the normalised ranks combination method (Table 11), on the other hand, the best results are obtained for X = 100 for the C + IN_100, and for X = 30 for the C + OUT_100. Finally, for the normalised scores (Table 9), C + IN_100 performs best for X = 1000 and C + OUT_100 for X = 50 documents. Generally, there appears to be a different trend for each combination for the methods performing normalisation before combining the

results. Precision increases as more documents are taken into account for the content + inlink combination, whereas for the content + outlink combination the best results are obtained when X is set to 30 or 50 documents. This maybe due to the fact that since the inlink representation retrieves more relevant documents than the outlink representation, it is more likely to contain more relevant documents further down the rankings.

Table 10: Content+IN and Content+OUT with 10:1 ratio of combination weights using the reciprocate ranks

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
		C + IN_100		
Top 10 docs	**0.3680**	0.3620	0.2760	0.2170
Top 20 docs	**0.3680**	0.3660	0.3493	0.3310
Top 30 docs	**0.3680**	0.3640	0.3467	0.3310
Top 50 docs	**0.3680**	0.3640	0.3453	0.3270
Top 100 docs	**0.3680**	0.3660	0.3467	0.3260
Top 1000 docs	**0.3680**	0.3620	0.3467	0.3250
		C + OUT_100		
Top 10 docs	0.3600	0.3600	0.2707	0.2100
Top 20 docs	0.3600	0.3520	0.3427	0.3290
Top 30 docs	0.3600	0.3520	0.3440	0.3260
Top 50 docs	0.3600	0.3520	0.3440	0.3250
Top 100 docs	0.3600	0.3520	0.3440	0.3250
Top 1000 docs	0.3600	0.3520	0.3440	0.3250

Table 11: Content+IN and Content+OUT with 10:1 ratio of combination weights using the normalised ranks

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
		C + IN_100		
Top 10 docs	**0.3680**	0.3620	0.2760	0.2170
Top 20 docs	**0.3680**	**0.3680**	**0.3520**	0.3250
Top 30 docs	**0.3680**	0.3660	**0.3533**	0.3350
Top 50 docs	**0.3840**	0.3620	**0.3547**	0.3340
Top 100 docs	**0.4040**	0.3640	0.3493	**0.3370**
Top 1000 docs	0.3440	0.3020	0.2747	0.2460
		C + OUT_100		
Top 10 docs	**0.3680**	0.3600	0.2707	0.2100
Top 20 docs	0.3560	**0.3680**	0.3493	0.3250
Top 30 docs	0.3600	**0.3680**	0.3493	0.3350
Top 50 docs	**0.3760**	0.3620	0.3480	0.3330
Top 100 docs	**0.3680**	0.3480	0.3387	0.3300
Top 1000 docs	0.2640	0.2060	0.1667	0.1590

Even though the combination of content and link-based document representations does not appear to be beneficial for the retrieval effectiveness, three out of the four combination methods (all except the unnormalised scores) generally provide results close to the baseline, regardless of the number of included documents. This confirms the hypothesis investigated in (Ogilvie and Callan, 2003), although in a different context, that compatible outputs of the individual document representations constitute an important factor for successful combination. This compatibility is achieved both by normalising the retrieval scores and by considering the ranks. In addition, we cannot assume compatibility of the unnormalised

scores, since despite the fact that the same ranking algorithm is employed by all representations, the term statistics differ greatly for the nine representations.

We cannot conclude which combination method performs consistently better, despite limiting the number of factors affecting the combination i.e. the combination parameters, number of included documents and the combination method employed and examining the impact of each factor separately. However, the normalised scores method appears to be more robust than any other, regardless of the numbers of included documents, whereas normalised ranks achieves similar effectiveness considering less top-ranked documents.

Finally, we investigate the hypothesis that in combining document representations, each individual representation should perform well in order to improve the retrieval performance. To test this hypothesis, we examine the combination of two link-based representations IN_100 and OUT_100, which perform poorly with respect to the content-based baseline. Table 12 lists the results of their combination when the ratio of combination parameters is 1:1 and the normalised scores combination method is employed.

Table 12: Combination of IN+OUT with 1:1 ratio of combination weights using the normalised scores

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
IN_100	0.1840	0.1340	0.1160	0.0940
OUT_100	0.0960	0.0620	0.0507	0.0480
		IN_100 + OUT_100		
Top 10	**0.1960**	**0.1480**	**0.1360**	**0.1100**
Top 20	**0.1960**	**0.1500**	**0.1347**	**0.1140**
Top 30	**0.1920**	**0.1440**	**0.1347**	**0.1180**
Top 50	**0.2000**	**0.1420**	**0.1360**	**0.1200**
Top 100	**0.1840**	**0.1460**	**0.1240**	**0.1170**

These results demonstrate that the above hypothesis is not true. This is consistent with the results of recent experiments both in the field of combination of document representations and also in the field of metasearch (Ogilvie and Callan, 2003). This means that even though we may obtain in further study other link-based representations that do not perform well when used individually, this should not stop us considering combining them with other representations, as increased performance may be achieved.

6.3 Hubs and Authorities

This section discusses the combination of content-based and link-based evidence in terms of the combination of results from content-based representation with results from the analysis of the Web link structure. We examine the effect this combination has on the effectiveness compared to that of the content only approach.

Table 13 presents the results of the HITS algorithm, when the root set is constructed from the 50 top ranked documents of the content representation and the expansion step is performed by considering three different subsets of links. Before analysing these results, one might argue that this evaluation is not meaningful, since different sets of relevance assessments are required for algorithms, such as HITS, which aim at identifying the most important documents or key resources (Craswell and Hawking, 2002) for a specific topic. However, the results still give us an insight to the behaviour of different types of link-based evidence.

The best results for both hubs and authorities are obtained when the expansion is performed when considering only the inter-host inlinks and outlinks. Similar results are also observed,

when the root set is constructed using the 100 or 200 top ranked documents. This confirms the claim that only inter-host links convey information about the authority of the pages they point to and about the degree of hub of the pages they originate from (Kleinberg, 1999). By comparing the results of the experiments when the root set is constructed using the 50 top-ranked documents to those obtained when the top 100 and top 200 are used (these results are not reported here), the first approach achieves the best results.

Table 13: Results for HITS algorithm with the root set constructed using the 50 top ranked documents and the expansion performed using one of the 3 listed sets of links

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
		Authorities		
Inter-host	0.1320	0.1580	0.1893	0.2010
Reference	0.0080	0.0200	0.0213	0.0260
All links	0.0280	0.0280	0.0333	0.0360
		Hubs		
Inter-host	0.1640	0.1820	0.1933	0.2130
Reference	0.0920	0.0660	0.0600	0.0570
All links	0.0760	0.0580	0.0533	0.0450

Table 14 lists the results of the combination of the content document representation with the hubs and authorities obtained through the application of the HITS algorithm, when the root set is constructed using the 50 top ranked documents and the expansion step considers only the inter-host links.

Table 14: Content+HITS with 10:1 ratio of combination weights using the normalised scores. For HITS, the root set is constructed using the 50 top ranked documents and the expansion is performed using inter-host links.

	Precision at 5	**Precision at 10**	**Precision at 15**	**Precision at 20**
Content	0.3680	0.3680	0.3520	0.3360
		C + Authorities (Inter-host links)		
Top 10 docs	**0.3720**	0.3320	0.2733	0.2430
Top 20 docs	**0.3680**	0.3660	0.3440	0.3120
Top 30 docs	**0.3680**	**0.3680**	**0.3520**	0.3300
Top 50 docs	**0.3680**	**0.3680**	**0.3520**	0.3340
		C + Hubs (Inter-host links)		
Top 10 docs	**0.3720**	0.3480	0.2867	0.2500
Top 20 docs	**0.3680**	**0.3680**	0.3507	0.3180
Top 30 docs	**0.3680**	**0.3680**	**0.3520**	0.3320
Top 50 docs	**0.3680**	**0.3680**	**0.3520**	0.3320

The results indicate that the precision values are generally comparable to those of the baseline, but still the incorporation of this type of link-based evidence does not lead to improvements in the effectiveness. Therefore, the incorporation of evidence based on link structure analysis does not appear to be beneficial for the topic relevance task.

7. Conclusions

The aim of this study is to examine the impact the various types of link-based evidence have on the retrieval effectiveness, especially when combined with content-based evidence, for the

topic relevance task in the Web context. We consider three types of link-based evidence: i) different types of links classified according to their syntax and their semantics, ii) Web document representations generated using link-based evidence and iii) the results obtained by the application of Kleinberg's HITS link analysis algorithm. Our investigation is carried out using the inference network model as the underlying framework. We perform experiments using the WT2g test collection, in order to evaluate the effectiveness of the individual link-based document representations and of the combination of content-based and link-based evidence.

Regarding the individual document representations, the content-based one performs significantly better than any of the link-based representations. The inlink-based representations perform significantly better than the equivalent outlink-based ones, demonstrating that the anchor text is more descriptive for the pages it references than for the pages that contain it. Out of the specific types of links considered, the sequence type links appear to provide the strongest evidence. Also, the effectiveness improves significantly and more relevant documents are retrieved, as the size of the anchor window increases.

For the combination of content and link-based document representations, the results indicate that stronger evidence should be placed on the content-based representation. Three out of the four combination methods (all except the unnormalised scores) generally provide results close to the baseline, regardless of the number of included documents, confirming that compatibility of the outputs of the individual document representations constitutes an important factor for successful combination. We cannot conclude which combination method performs consistently better. The normalised scores method appears to be more robust than the others regardless of the numbers of included documents, whereas normalised ranks achieves similar effectiveness considering less top-ranked documents. Finally, by combining document representations that do not perform well when used individually, increased performance may be achieved.

Regarding the application of HITS algorithm, the best results for both hubs and authorities are obtained the root set is constructed using the 50 top-ranked documents and the expansion is performed when considering only the inter-host inlinks and outlinks. In the context of the combination of results from content-based representation with results from HITS, the effectiveness is generally comparable to that of the content-based representation, but still the incorporation of this type of link-based evidence does not lead to improvements. Therefore, the incorporation of these types of link-based evidence does not appear to be beneficial for the topic relevance task, as already observed in TREC-8 (Hawking et al., 1999b), TREC-9 (Hawking, 2000) and TREC-2001 (Hawking and Craswell, 2001).

An obvious continuation of this study would be to consider a different corpus, one that contains more inter-host inlinks. However, we believe that similar observations would be made, as it has been the case for the topic relevance task in TREC-9 and TREC-2001, where the WT10g was used and the incorporation of link-based evidence hardly improved the effectiveness. In the future, we plan to investigate other combination methods based on the inference network model, by considering for instance combination methods that do not rely on canonical link matrix forms. A more interesting line of work is also to apply our approach in the context of other Web retrieval tasks, in order to investigate the impact of the various types of link-based evidence. The insights we have gained into the behaviour of link-based evidence will assist us in our future investigation.

References

Allan, J., Callan, J., Feng, F. & Malin, D. (1999). INQUERY and TREC-8. In Voorhees, E. M. and Harman, D.K. (Eds.), *Proceedings of 8th Text Retrieval Conference (TREC-8)*, NIST Special Publication 500-246, (pp. 637-644). Gaitensburg, MD.

Amento, B., Terveen, L. & Hill, W. (2000). Does authority mean quality? Predicting expert quality rankings of Web documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 296-303). ACM Press.

Amitay, E. (1998). Using common hypertext links to identify the best phrasal description of target Web documents. In *Proceedings of the 21st Annual International ACM SIGIR's Post-Conference Workshop on Hypertext Information Retrieval for the Web*.

Amitay, E. (2000). InCommonSense – Rethinking Web Results. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2000)*, (pp. 1705-1708).

Belkin, N. J., Kantor, P., Fox, E. A. & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3), 431-448.

Bharat, K. & Henzinger, M. (1998). Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 104-111). ACM Press.

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale HyperTextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference*, (pp. 107-117). Elsevier Science.

Broglio, J., Callan, J. & Croft, W.B. (1994). INQUERY System Overview. In *Proceedings of the TIPSTER Text Program*. (pp. 47-67). Morgan Kaufmann, San Francisco, CA.

Callan, J. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp.202-310). ACM Press.

Callan J.P., Croft, W.B. & Broglio, J. (1994). TREC and TIPSTER Experiments with INQUERY. In *Information Processing and Management*, 31(3), 327-343.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P. & Rajagopalan, S. (1998). Automatic resource list compilation by analysing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*.

Craswell, N. & Hawking, D. (2002). Overview of the TREC-2002 Web Track. In Voorhees, E. M. and Buckland, L.P. (Eds.), *Proceedings of 11th Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251. Gaitensburg, MD.

Craswell, N., Hawking, D. & Robertson, S. (2001). Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 250-257). ACM Press.

Croft, W.B (2000). Combining approaches to information retrieval. In Croft, W. B. (Ed.), Advances in Information Retrieval: Recent Research from the Centre for Intelligent Information Retrieval, (pp. 1-36). Kluwer Academic Publishers.

Davison, B. D. (2000). Topical Locality in the Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 272-279). ACM Press.

de Campos, L. M., Fernández-Luna, J. M. & Huete, J. F. (2002) A Layered Bayesian Network Model for Document Retrieval. In *Proceedings of the 24th European Colloquium on Information Retrieval Research (ECIR 2002)*, (pp. 169-182).

Fox, E. A. & Shaw, J. A. (1994). Combination of multiple searches. In Harman, D.K. (Ed.), *Proceedings of the 2nd Text Retrieval Conference (TREC-2)*, NIST Special Publication 500-215, (pp. 243-249). Gaitensburg, MD.

Frisse., M. & Cousins, S. (1989). Information retrieval from hypertext: Update on the dynamic medical handbook project. In *Proceedings of the ACM Hypertext Conference*, (pp. 199-212).

Géry, M. & Chevallet, J. P. (2001). Toward a Structured Information Retrieval System on the Web: Automatic Structure Extraction of Web Pages. In *Proceedings of the International Workshop on Web Dynamics, in conjunction with the 8th International Conference on Database Theory*.

Haveliwava, T. H. (2002) Topic-Sensitive PageRank. In *Proceedings of the 11th International World Wide Web Conference.*

Hawking, D. (2000). Overview of the TREC-9 Web Track. In *Proceedings of 9th Text Retrieval Conference (TREC-9)*, NIST special publication 500-249, (pp.87-102). Gaitensburg, MD.

Hawking, D. & Craswell, N. (2001). Overview of the TREC-2001 Web Track. In *Proceedings of 10th Text Retrieval Conference (TREC-2001)*, NIST special publication 500-250, (pp.61-67). Gaitensburg, MD.

Hawking, D., Craswell, N., Bailey, P. & Griffiths, K. (2001). Measuring Search Engine Quality. *Information Retrieval*, 4(1), 33-59.

Hawking, D., Craswell, N., Thistlewaite, P. & Harman, D. (1999a). Results and Challenges in Web search evaluation. In *Proceedings of the 8th International World Wide Web Conference.*

Hawking, D., Voorhees, E., Craswell, N. & Bailey, P. (1999b). Overview of the TREC-8 Web Track. In Voorhees, E. M. and Harman, D. K., editors, *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, NIST special publication 500-246, (pp. 131-150). Gaithersburg MD.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.

Kraaij, W., Westerveld, T. & Hiemstra, D. (2002) The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 27-34). ACM Press.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 267-276). ACM Press.

Ogilvie, P. & Callan, J. (2003). Combining Document Representations for Known-Item Search. In *Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 143-150). ACM Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc.

Rajashekar, T. B. & Croft. W. B. (1995). Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society of Information Science*, 46(4), 272-283.

Ribeiro-Neto, B. & Muntz, R. (1996) A Belief Network Model for IR. In *Proceedings of the 19th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 253-260).

Ribeiro-Neto, B., Silva, I. & Muntz, R. (2000) Bayesian Network Models for Information Retrieval. In Crestani, F. and Pasi, G (Eds.), *Soft Computing in Information Retrieval: Techniques and Applications*. (pp. 259-291).

van Rijsbergen, C. J. (1986). A Non-Classical Logic for Information Retrieval. *The Computer Journal*, 29(6), 481-485.

Savoy, J., Le Calvé, A. & Vrajitoru, D. (1996). Report on the TREC-5 Experiment: Data Fusion and Collection Fusion. In *Proceedings of the 5th Text Retrieval Conference (TREC-5)*, NIST Publication 500-238, (pp. 489-502). Gaithersburg, MD.

Selberg, E. & Etzioni, O. (1997) The MetaCrawler Architecture for Resource Aggregation on the Web. *IEEE Expert*, 12 (1), 8-14.

Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E. & Ziviani, N. (2000). Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 96-103), ACM Press.

Singhal, A. & Kaszkiel, M. (2001). A Case Study in Web Search using TREC Algorithms. In *Proceedings of 10th International World Wide Web Conference*, (pp. 708-716).

Tsikrika, T. & Lalmas, M. (2002). Combining Web Document Representations in a Bayesian Inference Network Model Using Link & Content-Based Evidence. In *Proceedings of the 24th European Colloquium on Information Retrieval Research (ECIR 2002)*, (pp. 53-72).

Turtle H. R. (1990) Inference Networks for Document Retrieval. Ph.D. dissertation.

Turtle, H. & Croft, W.B. (1991) Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems*, 9(3), 187 - 222.

Zhang, M. Song, R., Lin, C., Ma, L., Jiang, Z., Jin, Y., Liu, Y., Zhao, L. & Ma, S. (2002). THU at TREC-2002: novelty, web and filtering In Voorhees, E. M. and Buckland, L.P. (Eds.), *Proceedings of 11th Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251. (pp. 29-42).Gaitensburg, MD.