A VARIETY OF RETRIEVAL PERFORMANCE BY USING DIFFERENT TYPES OF LINK

Won-Kyun Joo, Myung-Seok Yang, Min-Ho Lee, Doo-Suk Jin, Yun-Soo Choi,
Jerry-Hyeon Seo, Beom-Jong You, Hyeon Kim
KISTI(Korea Institute of Science and Technology Information
P.O.BOX 122, Yusong, Taejon,
Republic of Korea

Abstract

Since hypertexts are ubiquitous, it has become more and more important to make use of hyperlink information for various applications such as information retrieval. We present in this paper a scheme for using various types of hyperlinks for the purpose of improving retrieval effectiveness. The scheme is based on the distinction between incoming and outgoing links as well as the distinction between direct and indirect links. We also differentiate whether or not anchors of links are query or non-query terms. We ran experiments using a test collection made of encyclopedia containing a large number of hyperlinks. The results show that various types of hyperlinks have different impact on enhancement of retrieval effectiveness and that the improvement can be as much as 44.8% in 11-point average precision when a right combination of weights are given to the different link types.

Key Words

Internet Search Technologies, Using Link Types, Re-ranking, Analysis of link types

1. Introduction

Information retrieval has dealt with a collection of documents whose contents are predominantly texts, a linear sequence of characters, words, and sentences. With growing popularity with World Wide Web (WWW) where texts are inter-linked in a complex way to form a hypertext, however, it has become a common practice to access texts in a non-sequential way by means of navigation or browsing. Nonetheless, search engines on the WWW employ traditional information retrieval techniques that work under the assumption that documents are independent units. Word frequencies in individual documents and in the entire collection are the basis for determining the degree of association between a word and a document.

In this paper, we take a stance that the meaning of a document is not just restricted to the sequential text itself but can be affected by and derived from its relationship with others. We hypothesize that when a source text (or a node in hypertext) is linked to a destination text by means of an anchor in the source and a link pointing to the destination, they each reveal the content of the other to some extent. The main thrust of this paper is to present a way of making use of this link information for retrieval purposes and show that this information in fact improves retrieval effectiveness.

Hypertexts allows for browsing, an effective way of acquiring information from a structured information space and of finding unexpected items in serendipity. However, since browsing is not an efficient means with a large search space, researchers have attempted to combine browsing and searching functions in a single retrieval environment. Lucarella (1990), for example, proposed a model for retrieval of hypertexts and its implementation[3]. More recently researchers have attempted to integrate browsing and searching for hypermedia documents. Croft and Turtle (1989) developed a method by which hypertext links are regarded a evidence in answering a query[1]. Savoy (1996) suggested a scheme where a vector space model was extended to include citation links to improve overall retrieval effectiveness[5]. Recently, Brin and Page(1998) have proposed a ranking measure based on a node-to-node weight propagation scheme and its analysis via eigenvectors[6]. In order to measure the relative importance of web pages, they propose PageRank, a method for computing a ranking for every web pages based on the graph of the web. It is the first application of these approaches to www search. Kleinberg(1999) has developed a model of the web as Hubs and Authorities, based on eigenvector calculation on the co-citation matrix of the web[7]. Our work is a generalization and an extension of these works.

Links in hypertexts have many different characteristics and can be associated with many different meanings. Depending on the nature of the objects at the source and destination points, first of all, links can connect a document with another document or a term with a document. On the other hand, the directionality of a link can determine whether it is an outgoing or incoming link with respect to a given text. Although not common, it is also possible to build a one-to-many link so that more than one destination can be associated with the source. Furthermore, we can define an indirect link between two objects when they are pointed to by a single source, or when they point to a single destination. Finally, we can attach an explicit meaning such as 'example-of' or 'definition-of', or a numerical weight along a link to indicate its strength.

While our long-term goal is to investigate on the roles of various types of links and their properties in improving retrieval effectiveness, our current work has a focus on directionality and the distinction between direct and indirect links. In this paper we suggest ways to make use of such link information in computing the retrieval status value (RSV) or similarity value between a document and query. We also report the results from our experiments with which we tested the value and impact of different link information.

2. An algorithm for link-based retrieval

Among many different types of link characteristics mentioned above, we focus on only two aspects: directionality and directness of links. For the former, we make a distinction between outgoing and incoming links

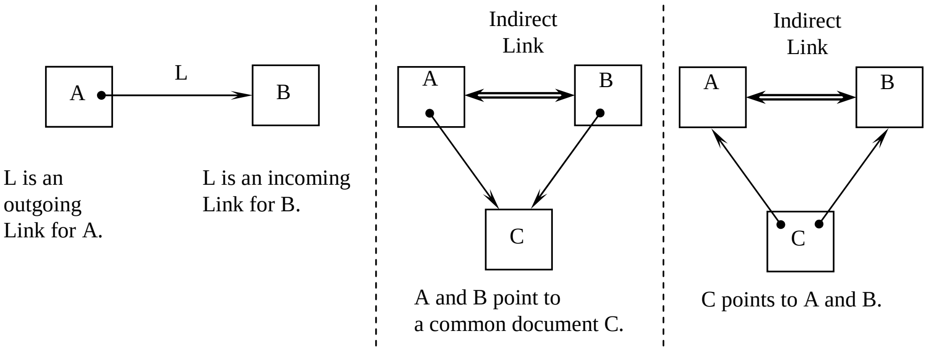
A
L
B
L is an outgoing Link for A.
L is an incoming Link for B.
Indirect Link
A
B
C
A and B point to a common document C.
Indirect Link
A
B
C
C points to A and B.

Fig. 1 Directionality and directness of links

When there is a link from document A to B, it serves as an outgoing link for A and as an incoming link for B. For directness, we define an indirect link that is assumed to exist between documents A and B when they each have an outgoing link to a common document C, or when they each have an incoming link from the same document C. Outgoing and incoming links described above are all examples of a direct link. Fig. 1 illustrates the different types of links.

The overall algorithm for link-based retrieval works as follows.

1. A set of documents (called internal documents) is retrieved with a method based on the vector space model to form an initial retrieval set.
2. By using link information, the set is expanded to include additional documents (called external documents), forming an extended set.
3. The documents in the extended set are re-evaluated for their relevancy based on the link information that exists among the documents

The purpose of the second step is to enhance recall by retrieving those documents that would not be retrieved by the given query but are linked with the initial retrieval set. The new external documents as well as the internal documents are re-evaluated at the third step, where the locations of the source and the destination of the links between any two documents in the extended set are taken into account. Table 1 shows 8 different types of links, depending on whether the source of a link is from a query term or from a non-query term and on where the source and the destination of the link are located. For example, the case 3 refers to a link whose source is a query term in an external document, and whose destination is an internal document. It should be noted that an external document may contain a query term when its rank was lower than the cut-off point and therefore excluded from the initial set.

Table 1 Eight different types of links considered

	Internal External	Internal Internal	External Internal	External External
Query	case 1	case 2	case 3	case 4
Non-query	case 5	case 6	case 7	case 8

Each of the three steps is explained in detail in the following sub-sections.

2.1 The first step: construction of the initial retrieval set

This step is essentially the same as the ordinary process of retrieving documents using a vector space model. We employed the following formula to calculate a weight w_{ij} for a term j appearing in document i:

$$w_{ij} = nt_{ij} \cdot nidf_j = \frac{tf_i}{\max tf_i} \cdot \frac{\log[\frac{n}{df_i}]}{\log(n)} \qquad (1)$$

where n_t is the total frequency of a term normalized by the maximum term frequency in the document, n_{idf} is the normalized inverse document frequency for the term, and n is the total number of documents in the collection. The retrieval status value (*RSV*) for a document i for a query q is defined to be a cosine similarity value:

$$RSV(D_i, Q) = \sum_{j=1}^{q} \frac{w_{ij} \cdot w_{qj}}{|D_i| \cdot |Q|} \qquad (2)$$

When this step is completed, the documents that contain at least one of the query terms and are listed above a certain cut-off point.

2.2 The second step: determination of the extended set

This step handles the cases 1 and 5 in Table 1. The documents in the initial retrieval set are examined for their association with those outside the set, via an outgoing link from either a query or non-query term. Each document connected with a link from a term in one of the internal documents becomes an element in the extended set. This process resembles a form of blind relevance feedback, where top-ranked documents in the initial retrieval are combined with the original query to form an extended query, in the sense that additional documents are retrieved based on their connection to some of the initially retrieved documents. The difference lies in that we use human-generated connections (*i.e. links*) to retrieve additional documents as opposed to word occurrences in initially retrieved documents, which are the basis for adding new terms in the extended query.

Since external documents do not ordinarily contain query terms, it is not possible to calculate *RSV* for them by simply applying the formula as in (2). Our strategy in assigning a *RSV* for an external document is to let it inherit the *RSV* from the source document containing the anchor term of the link, reflecting the similarity between them. In other words, it is computed by first calculating the similarity between the two (i.e. an external document and the corresponding internal document) and prorating the *RSV* of the internal document accordingly based on the similarity. More formally,

$$RSV(D_e, Q) = RSV(D_i, Q) * Sim(D_i, D_e) \ldots\ldots\ldots\ldots..(3)$$

where $0 \leq Sim(D_i , D_e) \leq 1$. When there are more than one incoming links to the external document, we calculate the RSV's for individual links using the formula in (3) and choose the maximum. It is also possible to combine them with the *Dempster-Shafer* combination rule [4], under the assumption that the more incoming links from the initial retrieval set, the more likely the external document would satisfy the query (i.e. the higher *RSV*). Since this aspect of accumulating evidence is incorporated when the final *RSV* is calculated as described below, we take a conservative approach of giving a *RSV* that is no larger than that for the most relevant document among those that point to the external document at hand.

2.3 The third step: incorporation of link information

After generating all the candidate documents for the extended set, the next step is to re-rank them using all the link information across documents in the set. Not only can new documents (i.e. external documents) be inserted at various places in the list of retrieved internal documents, but also the original ranks of the internal documents can change with newly calculated RSV's. From this step on, the distinction between internal and external documents is no longer necessary.

The basic scheme is to make documents influence each other when they are connected with links. We conjecture that the more links candidate documents share, the more likely they form a coherent body of documents for the given query since the candidates are supposed to be more or less relevant to the query in the first place. Since different link types are likely to have different impact on the bondage between documents, we analyze them separately based on direct and indirect links.

- **Impact by direct links**

A direct link can be further classified based on the directionality (i.e. whether a link is an incoming or outgoing link) and whether the anchor is a query term or non-query term. Considering the four cases, we use the formula (4) below to calculate the effect of direct links on a given document *D*. Here the four terms represent the following, respectively: 1) the effect of outgoing links starting from query terms in *D*, 2) outgoing links starting from non-query terms *D*, 3) incoming links from query terms in other document, and 4) incoming links from non-query terms in other documents. They are combined with the *Dempster-Shafer* combination rule represented by the operator '⊕'. Evidence E_{dl} for a document i by direct links is defined as follows:

$$E_{dl}(D_i) = \alpha_0 E_{ik} \oplus \alpha_1 E_{il} \oplus \alpha_2 E_{im} \oplus \alpha_3 E_{in}$$

$$with \ :$$

$$E_{ik} = \sum_{k=1}^{r} RSV(D_k, Q) \qquad \text{….. (4)}$$

$$E_{il} = \sum_{l=1}^{s} RSV(D_l, Q)$$

$$E_{im} = \sum_{m=1}^{t} RSV(D_m, Q)$$

$$E_{in} = \sum_{n=1}^{w} RSV(D_n, Q)$$

where α_i are the parameters indicating the strength or the importance of the four types of links, and *r*, *s*, *t*, and *w* are the numbers of the four different types of links starting from or pointing to *D*, respectively. Also, *k*, *l*, *m*, and *n* are the each documents corresponding to the four terms representation above.

It should be noted that the symbol ∑ does not have the usual meaning of summation but the meaning of the operation ⊕ based on *Dempster-Shafer* combination rule. This not only ensures that the value of each term never exceeds 1 but also reflects the intuition that the link information provides additional evidence that the document satisfies the query, which should be gathered in a principled way. The new *RSV* value for *D* can be computed in the following way:

$$RSV^{new}(D_i, Q) = RSV(D_i, Q) \oplus E_{dl}(D_i) \text{ ……………..(5)}$$

- **Impact by indirect links**

When two documents A and B have links to the same document, we assume an association between them. Likewise, when a document has separate links to the two documents, we also assume an association between (see Figure 1). In other words, a *bi-directional indirect link* can be established between two documents A and B by a common destination from the links starting from them, or by a common source of two (or more) links starting from another document C, which point to A and B, respectively. In order to compute the strength of the indirect link between D_i and D_j, we need to take into account two aspects. One is how many links leave from D_i and D_j separately and how many of them point to the same destination D_k, the other is how many links arrive in D_i and D_j starting from D_k at the same time. The evidence E_{il} for a document *i* by indirect links can be calculated as follows:

$$E_{il}(D_i) = \sum_{r \in L_{ij}} (\sigma_{i,j} * RSV(D_j, Q)) \oplus \sum_{s \in M_{ij}} (\varphi_{i,j} * RSV(D_j, Q)) \qquad \text{.. (6)}$$

$$\sigma_{ij} = \frac{|L_{ij}|}{|L_i| + |L_j|}, \varphi_{ij} = \frac{|M_{ij}|}{|M_i| + |M_j|}$$

where |•| gives the number of links. L_i and L_j are the each links starting from document *i* and *j*. L_{ij} indicates a pair of two links leaving from D_i and D_j and arriving at the same destination document *k*. Indirect links(M_{ij}) created by a document *k* pointing to D_i and D_j can be calculated in a similar way. *M* is same as *L* except arriving D_i and D_j instead of leaving from these.

Considering both the direct and indirect effect from links, we calculate the final *RSV* for *D* as follows:

$$RSV^{final}(D_i, Q) = RSV(D_i, Q) \oplus E_{dl}(D_i) \oplus \alpha_4 E_{il}(D_i) \cdots\cdots \text{ (7)}$$

It should be noted that while the parameter α_4 is explicitly shown in this formula, other parameters are hidden in E_{dl} *(D_i)* as in the formula (4). The exact values for α_i's are determined by experiments.

3. Experiments

We implemented the re-ranking scheme described above and ran a series of experiments to understand the effects of using link information in document retrieval. While a more immediate goal was to demonstrate that utilization of link information improves retrieval effectiveness, we were also interested in understanding the roles of different types of links we considered in the retrieval scheme.

There are a myriad of hypertext documents available in the World Wide Web, which contain interesting links, but it isn't possible to use them for the experiments because there are no relevance judgments for them. Although the CACM collection could be used by treating references as links to other documents as in [Savoy 1996], we felt that they do not serve the purpose of our research because the nature of the links are quite different from typical hypertext links. Our choice was a test collection developed by ETRI (Electronics and Telecommunications Research Institute) in Korea, which consists of encyclopedia documents containing a large number of hypertext links, published by Kyemong Publishing Co. The collection consists of 23,113 documents (9.4 MB) in Korean, which are basically explanations of titles such as "comet" and "gene", and 46 natural language queries. There are 182,844 hypertext links whose anchor words are the titles of other documents. In other words, following a link by clicking on an anchor word will lead the user to a new document whose title is the same as the anchor word.

In order to understand the roles each of the link types plays, we plugged in different numbers for the five parameters and measured the 11-point average precision. In Table 2, the results are summarized for different numbers of link types used in formula (4). Here αi's indicate different link types considered: outgoing links from a query term, outgoing links from a non-query term, incoming links from a query term, incoming links from a non-query term, and indirect links.

Table 2. Comparisons among contributions made by different link types

	Outgoing Query term (α_0)	Outgoing Non-query term (α_1)	Incoming Query term (α_2)	Incoming Non-query term (α_3)	Indirect links (α_4)	11-point average precision
One Link Type	0.2					0.4097
		0.7				0.4282
			0.5			0.4591
				0.1		0.3663
					0.1	0.3623
Two Link Types	0.6	0.8				0.4506
	0.6		0.7			0.5034
	0.7			0.2		0.3984
	0.5				0.2	0.4086
		0.6	0.9			0.4710
		0.7		0.1		0.4247
		0.9			0.3	0.4326
			0.6	0.4		0.4754
			0.5		0.1	0.4510
				0.1	0.1	0.3686
Three Link Types	0.7	0.6	0.7			0.5032
	0.5	0.9		0.1		0.4480
	0.7	0.9			0.1	0.4495
	0.7		0.9	0.4		0.5086
	0.7		0.9		0.3	0.5067
	0.9			0.3	0.4	0.4080
		0.4	0.9	0.3		0.4847
		0.5	0.9		0.2	0.4730
		0.9		0.1	0.2	0.4326
			0.9	0.5	0.2	0.4786
Four Link Types	0.8	0.1	0.9	0.5		0.5123
	0.8	0.2	0.9		0.1	0.5047
	0.7	0.9		0.1	0.1	0.4491
	0.8		0.9	0.2	0.2	0.5108
		0.2	0.8	0.3	0.1	0.4854
Five	0.7	0.1	0.9	0.3	0.1	0.5126
Baseline						0.3540

The first group of rows from the top in the table shows the case where only one of the link types was used. For example, the very first row shows the 11-point average precision when only outgoing links from a query term are considered in re-ranking documents. The values shown in the table (e.g. 0.2 in the first case) are the ones with the best performance in the particular case. For example, when only $\alpha1$ and $\alpha2$ were given a non-zero value (i.e. when only outgoing links were considered), the combination of 0.6 and 0.8 gave the best results among all other combinations of values considered between 0.1 and 1.0 with an interval 0.1.

As can be seen in the table, the effectiveness increased as we progressively added the right combination of link types. When only one link type was considered, which is the incoming links from a query term, the 11-point average precision was 0.4591, but when we added another link type (outgoing links from a query term), the value increased to 0.5034. The best value in each category increased to 0.5086, 0.5123, and 0.5126, showing steady improvements. This indicates that when the right parameter values are used, the five link types together contribute to enhancement of retrieval effectiveness.

A close examination of the results indicates that different contribution is made by different link types, however. Incoming links from a query term (signified by a non-zero value for $\alpha2$) are shown to have the strongest impact in all the five groups. Outgoing links from a query term are the next most important type. This can be seen most clearly in the second group where the best score was obtained when $\alpha0$ was given a non-zero value in addition to $\alpha2$. In other groups with three, four, and five non-zero parameters used, the value for α0 is the second biggest one. Similarly, we can see the next most important are incoming links with non-query terms.

In sum, links leaving from a query term are more important than those leaving from a non-query term, and incoming links are more important than outgoing links. That is, it is the safest to increase RSV for a document that is pointed to by links whose anchor is a query term. Indirect links seem to have the least impact on retrieval effectiveness.

Table 3. Comparison between the best case and the baseline

Recall Level	Baseline	Best Case with Links
0.0	0.5847	0.7599 (+ 29.9 %)
0.1	0.5847	0.7599 (+ 29.9 %)
0.2	0.5628	0.7437 (+ 32.1 %)
0.3	0.4792	0.6479 (+ 35.2 %)
0.4	0.4088	0.5538 (+ 35.4 %)
0.5	0.3794	0.5263 (+ 38.7 %)
0.6	0.3381	0.4707 (+ 39.2 %)
0.7	0.2283	0.3775 (+ 65.3 %)
0.8	0.2115	0.3373 (+ 59.4 %)
0.9	0.1682	0.2995 (+ 75.6 %)
1.0	0.1579	0.2995 (+ 89.6 %)
average	0.3540	0.5126 (+ 44.8 %)

There are some idiosyncratic aspects in these experiments, though. Because of the characteristics of the hypertexts in the encyclopedia collection, a link leaving from a query term seldom retrieves a new external document. This is because most of them should have been retrieved by the title word that is the same as the query word. The only documents that have been left out would be those below the cut-off point (30 in the experiments). This observation indicates that most of the improvement we obtained seems to stem from the re-ranking scheme rather than from newly retrieved documents. Nonetheless, we observed many cases where some relevant documents that would never be retrieved by query terms only were actually retrieved because of the link information.

Table 3 shows the comparison between the baseline case and the best case that took into account all of the five different link types, at various recall levels. As shown in the Table, the average improvement we obtained over the baseline was about 44.8 % when we used an appropriate combination of hypertext links.

4. Conclusion and future works

With the growing popularity of WWW documents, hypertext documents are ubiquitous. As such, it becomes more and more important to make use of link information existing in such documents for information retrieval. In this paper, we showed our scheme for using link information to improve retrieval effectiveness and experimental results that proved our hypothesis that such information would be useful. In the experiments using the ETRI-Kyemong test collection containing a large number of hyperlinks, we obtained as much as 44.8 % increase, in 11-point average precision, over the baseline where no link information was used.

We observed that by using links, both incoming and outgoing, we could get documents that would have never been retrieved by matching them against the query terms. At the same time, we could re-rank those that had already been retrieved so that relevant ones could be moved toward the top of the list.

While the results show that our approach to using link information is promising, there are some limitations. First of all, the size of the documents in the collection we used is small, and all the anchors are important key terms that correspond directly to the titles of other documents. Additional valuations should be done with more typical document sets. Even with the same collection, we would like to compare our results with the blind relevance feedback strategy where a set of top-ranked documents from the initial retrieval are combined to the original query to generate an expanded one. Still another extension to the current work is to devise a variety of schemes for indirect links and test their efficacy.

References

[1] W. Bruce Croft and Howard Turtle, A retrieval model for incorporating hypertext links, *Proc. Hypertext'89*, 1989, 213-224.

[2] M. D. Dunlop and C. J. van Rijsbergen, Hypermedia and free text retrieval, *Information Processing & Management*, *29*(3), 1993, 287-298.

[3] Dario Lucarella, A model for hypertext-based information retrieval, *Proc. European Conference on Hypertext(ECHT)*, 1990, 81-94.

[4] Elaine Rich and Kevin Knight, *Artificial Intelligence: 2nd edition* (McGraw-Hill, 1991).

[5] Jacques Savoy, An extended vector-processing scheme for searching information in hypertext systems, *Information Processing & Management, 32*(2), 1996, 155-170.

[6] Brin, s., Page, L., Anatomy of a large-scale hyper-textual web search engine, *Proc. of the 7th international World Wide Web Conference*, Brisbane, Australlia, 1998, 107-117.

[7] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM, 46*(5), 1999, 604-632.