Conceptual Information Extraction with Link-Based Search

Kyung-Joong Kim and Sung-Bae Cho

Department of Computer Science, Yonsei University,
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
uribyul@candy.yonsei.ac.kr sbcho@csai.yonsei.ac.kr

Abstract. Link-based search provides a new vehicle to find relevant web documents on the WWW. Recently, there is quite a bit of optimism that the use of link information can improve search quality as in Google. Usually, text-based search engine returns web sites which have simply the best frequency of user query, so that the result might be different from user's expectation. However, hypertextual search engine finds the most authoritative site. Proposed search engine consists of crawling, storing of link structure, ranking, and personalization processes. User profile encodes different relevances among concepts for each user. For conceptual information extraction from link-based search engine, fuzzy concept network is adopted. Fuzzy concept network can be personalized using the profile information and used to conduct fuzzy information retrieval for each user. By combining personal fuzzy information retrieval and link-based search, proposed search agent provides high-quality information on the WWW about user query. To show the effectiveness of the proposed search engine, a subjective test for five persons is conducted and the result is summarized. The result for five persons shows the usefulness of the proposed system and possibility for personalized conceptual information extraction.

1 Introduction

Search engine is one of the important services of web, and search engines such as Yahoo, Lycos and Altavista are mainly used. Recently, Google and Clever Search are considered as a promising next-generation search engine, which have a common feature of using link structure. While the computation of web document's importance and ordering of search result are based on link structure, link information distills valuable documents that cannot be found using text information. Search result must be the most reliable site that people expect. Google solves the problem of slow speed by computing the importance of web documents before searching takes place [1]. Clever Search distills a large search topic on the WWW down to a size that makes sense to a human user. It identifies authoritative and hub sources about user query [2]. While authoritative and hub sources are calculated using link information, authoritative sources are the most reliable web site about specific topics and hub sources are documents that link to many authoritative sources [3].

N. Zhong et al. (Eds.): WI 2001, LNAI 2198, pp. 367–372, 2001.

This paper proposes a system that searches web documents based on link information and fuzzy concept network. We can expect more quality results because it searches using link structure, and more personalized results because it utilizes the fuzzy concept network for more satisfaction to user. Fuzzy concept network calculates the relevance among concepts using fuzzy logic and represents the knowledge of user [4,5]. The construction of fuzzy concept network is based on user profile. Search engine selects the web sites appropriate for user by processing fuzzy document retrieval using fuzzy concept network as user knowledge. Fuzzy concept network and fuzzy document retrieval system can be used for effective personalization method.

The rest of this paper is organized as follows. In Section 2, we propose architecture of personal web search engine using link structure and fuzzy concept network. In Section 3, we show search results and personalization process. Conclusions are discussed in Section 4.

2 Conceptual Link-Based Search

Fig. 1 shows the architecture of personal web search engine using hyperlink structure and fuzzy concept network. Search engine consists of crawling, storing of link structure, ranking, and personalization processes. It uses only link information to find relevant web pages, so that Store Server stores the link structure of web for efficient searching. Crawler extracts link information from crawled web pages and then sends URL and link information to Store Server. As user submits a query, search engine executes a ranking algorithm, which constructs base set using text-based search engine and finds authoritative and hub sources. Fuzzy document retrieval system based on fuzzy concept network is responsible for personalization process. A fuzzy concept network is generated for each user by the information on user profile. Using the fuzzy concept network generated, fuzzy document retrieval system finds the best documents for user.

2.1 Ranking

1. If i is a document in base set, authoritative weight of i is a_i and hub weight of i is h_i. a_i and h_i are initialized to 1.
2. a_i and h_i are updated by following formula.

$$a_i = \sum h_j \quad (j \text{ links to } i) \tag{1}$$

$$h_i = \sum a_j \quad (j \text{ is linked by } i) \tag{2}$$

3. Normalize weight of authoritative and hub so that the sum of squares is 1.
4. Until authoritative and hub weights converge, repeat 2 and 3.

From converged weights of authoritative and hub, best authoritative and hub sources are decided.

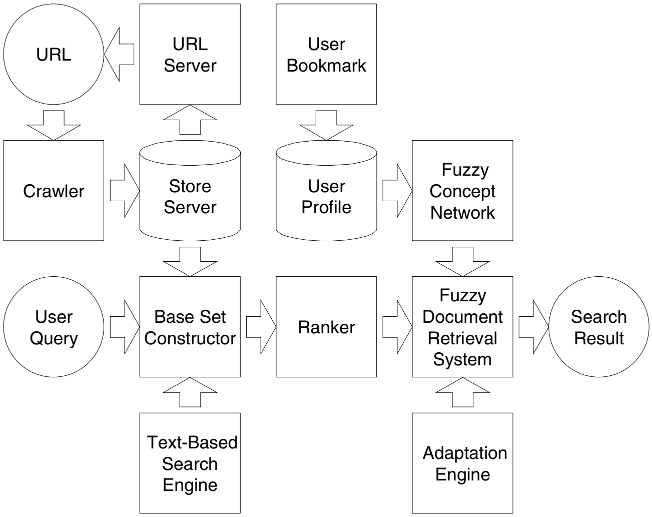

Fig. 1. System overview of the proposed web search engine

2.2 Personalization Process

Lucarella proposed fuzzy concept network for information retrieval [6]. A fuzzy concept network includes nodes and directed links. Each node represents a concept or a document. $C = \{C_1, C_2, \cdots, C_n\}$ represents a set of concepts. If $C_i \xrightarrow{\mu} C_j$, then it indicates that the degree of relevance from concept C_i to C_j is μ. If $C_i \xrightarrow{\mu} d_j$, then it indicates that the degree of relevance of document d_j with respect to concept C_i is μ. $C_i \xrightarrow{\mu} C_j$ is represented with $f(C_i, C_j) = \mu$. Using fuzzy logic, if $f(C_i, C_j) = \alpha$ and $f(C_j, C_k) = \beta$ then $f(C_i, C_k) = \min(\alpha, \beta)$. $C_i \xrightarrow{\mu} d_j$ is represented with $g(C_i, d_j) = \mu$. A document d_j has a different relevance to concepts. A document d_j can be expressed as a fuzzy subset of concepts.

$$d_j = \{(C_i, g(C_i, d_j)) \mid C_i \in C\} \qquad (3)$$

If there are many routes from C_i to C_j, $f(C_i, C_j)$ is decided with the maximum value.

For each document $h \in H$, on the basis of the binary indexing relation I, the document descriptor I_h of h is a fuzzy subset of C defined as follows.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \tag{4}$$

$$d_{ij} = I_{h_i}(C_j)\ ,\ 1 \le i \le m\ ,\ 1 \le j \le n$$

$C = \{c_1, c_2, \cdots, c_n\}$ is a set of concepts. A fuzzy concept matrix K is a matrix where $K_{ij} \in [0,1]$. The (i, j) element of K represents the degree of relevance from concept c_i to concept c_j. $K^2 = K \otimes K$ is the multiplication of the concept matrix.

$$K^2{}_{ij} = \bigvee_{l=1}^{n} (K_{il} \wedge K_{lj})\ ,\ 1 \le i, j \le n \tag{5}$$

$\vee$ and $\wedge$ represent the max operation and the min operation, respectively. Then, there exists an integer $\rho \le n-1$, such that $K^{\rho} = K^{\rho+1} = K^{\rho+2} = \ldots$. Let $K^* = K^{\rho}$. K^* is called the transitive closure of the concept matrix K. Missed information of fuzzy concept network can be inferred from the transitive closure of itself. The relevance degree of each document, with respect to a specific concept, can be improved by computing the multiplication of the document descriptor matrix D and the transitive closure of the concept matrix K as follows.

$$D^* = D \otimes K^* \tag{6}$$

D^* is called the expanded document descriptor matrix.

3 Experimental Results

Proposed system selects the five authoritative results as a source of personalization. It makes a document descriptor of these documents. The ranking of these five documents is reordered with respect to user's interest, which is recorded in a user profile. User profile is constructed from the information of user's bookmark. Crawling URL's in the bookmark, HTML documents are extracted from web. Relevance between two keywords are computed by the value of cooccurrence in the documents crawled. If the cooccurrence number of two keywords is the maximum, relevance value is 1.0. Otherwise, the relevance between two keywords is the proportion of cooccurrence to the maximum. User profile contains 10 concepts as follows: “Book,” “Computer,” “Java,” “Internet,” “Corba,” “Network,” “Software,” “Unix,” “Family,” and “Newspaper.” User profile contains 20 degrees of relevance between 10 concepts. A fuzzy concept network for a user is generated based on 20 degrees of relevance in the user profile. Unrecorded information can be inferred from the transitive closure of the fuzzy concept network. Expanded document descriptor results from multiplication of the document descriptor and user's fuzzy concept network. The sum of the degree of relevances with respect to concepts decides new ranking of the documents.

In this experiment, five users evaluate five authoritative documents about “Java.” Each user evaluates five documents. Table 1 and 2 show the search result of a query of "Java." It selects “java.sun.com” as the best authoritative site about “Java.” Table 3 shows the personalized results of search engine about “Java” for five users. Shade box shows if personalized rank is equivalent to that ranked by user.

Table 1. Search result of java (authoritative result) and comparison with Google

Authoritative result	Google
1. java.sun.com	1. java.sun.com
2. www.javalobby.org	2. java.sun.com/docs/books/tutorial/
3. javaboutique.internet.com	3. softwaredev.earthweb.com/java
4. java.about.com/compute/java/mbody.htm	4. javaboutique.internet.com/
5. www.javaworld.com	5. www.sun.com/java/

Table 2. Search result of java (hub result)

Hub result
1. industry.java.sun.com/products
2. java.sun.com/industry
3. java.sun.com/casestudies
4. industry.java.sun.com/javanews/developer
5. industry.java.sun.com/jug

Table 3. Personalized search result (Shade box shows that personalized rank is equal to user-checking’s.)

User 1	User 2	User 3	User 4	User 5
2	1	2	1	2
1	2	1	3	1
3	3	3	2	3
4	4	5	5	4
5	5	4	4	5

4 Conclusions

To find relevant web documents for a user, the proposed search engine uses link structure and fuzzy concept network. Search engine finds authoritative and hub sources for a user query using link structure. For efficient searching, link structure is stored in advance. Fuzzy document retrieval system personalizes link-based search results with respect to user’s interest. User’s knowledge is represented using fuzzy concept network. Search engine finds relevant documents in which user is interested and reorders them according to user’s interest. Using user’s feedback about search results, it is possible to change the value of fuzzy concept network. This adaptation procedure helps to get better results to fit user’s preference.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. The Seventh International WWW Conference. (1998)
2. The Clever Search, http://www.almaden.ibm.com/cs/k53/clever.html.
3. Kleinberg, J.: Authoritative sources in a hyperlinked environment. IBM Research Report RJ 10076. (1997)
4. Chen, S. -M., Horng, Y.-J.: Fuzzy query processing for document retrieval based on extended fuzzy concept networks. IEEE Transactions on Systems, Man, and Cybernetics, vol. 29, no. 1. (1999) 96-104
5. Chang, C.-S., Chen, A. L. P.: Supporting conceptual and neighborhood queries on the world wide web. IEEE Transactions on Systems, Man, and Cybernetics, vol. 28, no. 2. (1998) 300-308
6. Lucarella, D., Morara, R.: FIRST: Fuzzy information retrieval system. Journal of Information Science, vol. 17, no, 2. (1991) 81-91