A Personalized Web Search Engine Using Fuzzy Concept Network with Link Structure

Kyung-Joong Kim, 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
There has been much research on link-based search engines such as google and clever. They use link structure to find precision result. Usually, a link-based search engine produces high-quality result than text-based search engines. However they have difficulty to produce the result fit to a specific user's preference. Personalization is required to support more appropriate result. Among many techniques, the fuzzy concept network based on user profile can represent user's subjective interest properly. This paper presents another search engine that uses the fuzzy concept network to personalize the results from a link-based search method. The fuzzy concept network based on user profile reorders five results of the link-based search engine, and the system provides personalized high-quality result. Experimental results with three subjects indicate that the system developed searches not only relevant but also personalized web pages on user's preference.

1. Introduction

There is quite a bit of recent optimism that the use of link information can help improve search quality [1,2]. Text-based search engine ranks using both the position and frequency of keywords for their heuristics: The more instances of a keyword, and the earlier in the document those instances occur, the higher the document's ranking. For example, if user wants the most important web site about "physics," text-based search engine returns a web site which has the best frequency of "physics." This web site might be different from user's expectation. Also, keyword spamming lets web page designers trick the algorithm into giving their pages a higher ranking. For example, ranking spammers often stuff keywords into invisible text and tiny text. Hidden from most web users but visible to spiders, such text brims with repeated instances of keywords, thereby elevating a site's ranking relative to more scrupulous sites that restrict such keywords to legitimate usage [3]. Link-based search engine finds most authoritative site, so that these problems can be solved.

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 it represents the knowledge of user [4,5,6]. The construction of fuzzy concept network is based on user profile. Search engine selects fitting web sites for user by processing fuzzy document retrieval using fuzzy concept network as a 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, the current status of search engine is introduced. In Section 3, we propose an architecture of personal web search engine using link structure and fuzzy concept network. In Section 4, we show search results and personalization results. Conclusions are discussed in Section 5.

2. Search Engine

Usually, a search engine consists of crawler, indexer, and ranker. A crawler retrieves web documents from the web [7]. Search engines create a map of the web by indexing web pages according to keywords. From the enormous databases that these indexes generate, search engines link the page contents through keywords to URL's. When a user who seeks information submits a keyword or phrase that best describes user's need, the database of search engine ideally returns a list of relevant URL's.

Search engines such as AltaVista, Lycos, and Hotbot use crawler, also referred to as robots or softbots, to harvest URL's automatically. In directory-based search engines, such as Yahoo and AliWeb, webmasters and other web page creators manually submit the vast majority of indexed pages to the search engine's editors [8,9]. A directory-based search engine receives URL's from web page creators for possible inclusion in its database. Someone who wants a page recognized by Yahoo, for example, must submit the page's URL and background information to a human editor, who reviews it and decides whether to schedule the page for indexing. The indexing software retrieves the page scheduled for indexing, then parses and indexes it according to the keywords found in the page's content. For directory-based search engines, human gatekeepers hold the keys to inclusion in their indexed databases.

Currently on the Web, there are different techniques adopted by the search engines to help the user in shifting large sets of retrieved results. Nearly every search engine uses ranking to present its result in their order of relevancy, to the user. Google and Clever Search use link structure to present its results in their order of relevancy [10,11].

Google is a system of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the web efficiently and produce much more satisfying search results than existing systems. The Google system with a full text and hyperlink database of at least 1.3 billion pages. Google includes URL's that are not crawled. So search result can contain web pages that are broken. However, Google excludes broken web page by computing web page's PageRank value.

Clever Search does not service commercially, but it is a promising next-generation search engine researched by IBM. The Clever Search engine incorporates several algorithms that make use of hyperlink structure for discovering high-quality information on the Web. Clever Search includes enhancements to HITS (Hypertext-Induced Topic Search) algorithm, hypertext classification, focussed crawling, mining communities, and modeling the web as a graph. A number of algorithmic methods to improve the precision and functionality of the basic HITS algorithm is researched in Almaden and elsewhere [12]. Using hypertext classification and topic distillation tools to focus a crawler to work within a specific topic domain, ignoring unrelated and irrelevant material is published [13].

3. Personal Web Search Engine

Figure 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.

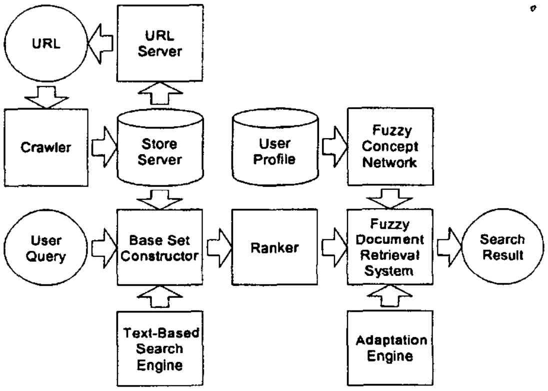

Figure 1. Personal link-based search engine.

3.1 Ranking

Authoritative and hub documents are defined for searching based on link information. Authoritative document contains the most reliable contents about a specific topic. Hub document contains many links to authoritative documents. Text-based search engine is used for constructing root set about user query. Root set contains 200 URL's which are used for expanding to base set. Root set is expanded by including forward link and back link from itself. By iterative weight updating, authoritative and hub rank of web document is decided. Figure 2 shows the construction of base set from root set.

Root set from text-based search engine does not contain all authoritative and hub sources about user query. By expanding root set, base set might contain authoritative and hub sources which are not included in root set. Base set contains enough authoritative and hub sources about user query. To find authoritative and hub sources in base set, iterative weight updating procedure is needed. The procedure is as follows.

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.

   $a_i = \sum h_j$ (j links to i)

   $h_i = \sum a_j$ (j is linked by i)
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 [2].

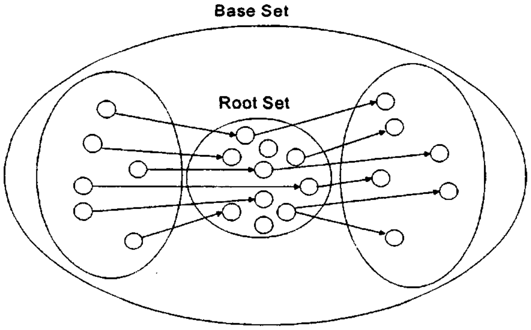

Figure 2. Construction personal link-based search engine.

3.2 Personalization

Lucarella proposed fuzzy concept network for information retrieval [14]. 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.

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

If there are many routes from C_i to C_j, $f(C_i,C_j)$ is decided with maximum value. Figure 3 shows an example of fuzzy concept network. In this figure, $f(C_3,C_2) = \max(0.4,0.3,0.2)$ and finally becomes 0.4.

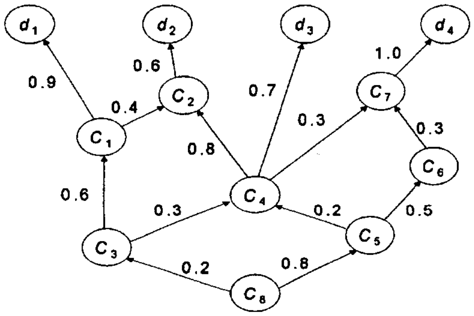

Figure 3. Fuzzy concept network.

Using fuzzy concept network, the document descriptor about $d_1, d_2, \cdots, d_n$ documents can be defined. Fuzzy document retrieval system can decide the importance of document using fuzzy concept network. If a user query is equal to concept C_i, it chooses the best relevant document about concept C_i among $d_1, d_2, \cdots, d_n$. Because this method takes long time to produce search result, fuzzy concept matrix is used.

Meanwhile, fuzzy document retrieval system uses fuzzy logic to deal with the uncertainty of document retrieval. Fuzzy theory was proposed by Zadeh in 1965 [15]. Fuzzy set theory provides a sound mathematical framework to deal with the uncertainty [16]. Fuzzy document retrieval system is defined as follows [6].

$$< H, C, Q, I, K, \phi, \psi >$$

H : set of documents
C : set of concepts
Q : set of queries
I : binary fuzzy indexing relation from H to C
K : knowledge base
$\phi : Q \times H \rightarrow [0,1]$, retrieval function
$\psi : H \times H \rightarrow [0,1]$, relevance function

For each pair (q,h), $q \in Q$, $h \in H$, $\phi(q,h) \in [0,1]$ is called the retrieval status value. For each pair (h_1,h_2), $h_1,h_2 \in H$, $\psi(h_1,h_2) \in [0,1]$ is called the degree of relevance between h_1 and h_2 or relevance degree between h_1 and h_2. The binary fuzzy indexing relation I is represented as the form of

$$I = \{\mu_I(h,c),(h,c) \mid h \in H, c \in C\}$$

with a membership function $\mu_I : H \times C \rightarrow [0,1]$, indicating for each pair (h,c) to what degree the concept c is relevant to document h. 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.

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

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

$C = \{c_1, c_2, \cdots, c_n\}$ is a set of concepts. A fuzzy concept matrix K is a matrix which $K_{ij} \in [0,1]$. 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.

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

$\vee$ 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 [6].

$$D^* = D \otimes K^*$$

D^* is called the expanded document descriptor matrix.

Fuzzy document retrieval system personalizes the results of link-based search engine. It selects the five best authoritative sources for a user query. These documents are the most reliable about a user query. First, it defines a document descriptor using the frequency of concept in a document. For each document, it counts the occurrence of concepts in user profile and normalizes the count between 0 and 1.

Fuzzy concept matrix is constructed from user profile that contains some of the relevances between n concepts. Figure 4 shows the construction of a fuzzy concept matrix based on user profile. It represents user's interest about concepts. If the relevance between C_i and C_j is recorded in user profile as μ , $<i,j>$ element of the fuzzy concept matrix is decided as μ . If the relevance between C_i and C_j is not recorded in user profile, $<i,j>$ element of the fuzzy concept matrix is decided as 0. Transitive closure of the fuzzy concept network represents all degree of relevances between n concepts.

User Profile

Java	Book	0.7
Java	Car	0.3
Java	WWW	0.9
Java	Ship	0.1
Book	Car	0.3
Book	WWW	0.5
Book	Ship	0.1
Book	Cafe	0.4
Car	WWW	0.7
Car	Ship	0.6
WWW	Ship	0.5
Ship	Cafe	0.3

Fuzzy Concept Matrix

	Java	Book	Car	WWW	Ship	Cafe
Java	1.0	0.7	0.3	0.9	0.1	0.0
Book	0.7	1.0	0.3	0.5	0.1	0.4
Car	0.3	0.3	1.0	0.7	0.6	0.0
WWW	0.9	0.5	0.7	1.0	0.5	0.0
Ship	0.1	0.1	0.6	0.5	1.0	0.3
Cafe	0.0	0.4	0.0	0.0	0.3	1.0

Figure 4. Construction of fuzzy concept matrix based on a user profile.

The expanded document descriptor of the five best authoritative sources can be decided by multiplying document descriptor of these documents and transitive closure of user's fuzzy concept network. Using the expanded document descriptor, new ranking of the documents is generated. The sum of relevance of a document with respect to concepts is used for recording.

4. Experimental Results

Search engine gets 100 URL's from the text-based search engine, say Altavista, about a user query. Root set consists of these 100 URL's. Store Server returns forward link and back link of the root set documents. Base set consists of root set, forward link set, and back link set. Among base set documents, it finds authoritative and hub sources. To regulate the size of base set, it limits forward link and back link of a root set document to 3 and 50, respectively. It selects the first three URL's in a document as forward link. The size of base set is about between 500 URL's and 1000 URL's. Some empirical study says that authoritative and hub weights of documents converge before 5 iterations. Therefore, the iteration number of ranking algorithm is decided as 5. Table 1 shows the search result of a query of "Java."

It selects "java.sun.com" as the best authoritative site about "Java." Also, it selects famous java sites such as "www.javalobby.org," "javaboutique.intemet.com," "java.about.com/compute/java/mbody.htm," and "www.javaworld.com" as authoritative sites. Table 2 shows the experimental results about other queries related with "Java." It selects "www.jini.org" as the best authoritative site about "Jini."

It selects the five authoritative results as a source of personalization. It makes a document descriptor of these documents. These five documents' ranking is reorded with respect to user's interest. User's interest is recorded on a user profile. 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 documents.

In this experiment, three users evaluate five authoritative documents about "Java." Table 3 shows the ranks three users made. Each user evaluates five documents. Table 4 shows the personalized results of search engine about "Java" for three users. Shade box shows if personalized rank is equal to user-checking's.

	Authoritative result
1.	java.sun.com
2.	www.javalobby.org
3.	javaboutique.internet.com
4.	java.about.com/compute/java/mbody.htm
	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

Table 1. Search Result of "Java"

	Query="Java2"
1.	java.sun.com
2.	www.appserver-zone.com
3.	www.sun.com/service/sunps/jdc/java2.html
4.	jdc.sun.co.jp
	Query="Javaone"
1.	java.sun.com
2.	www.togethersoft.com
3.	www.javacats.com
4.	www.zdevents.com
	Query="Jdk"
1.	java.sun.com
2.	developer.netscape.com/software/jdk/download .html
3.	java.sun.com/products/jdk/1.1/docs/index.html
4.	www.ora.com/info/java
	Query="Jguru"
1.	java.sun.com
2.	www.magelang.com
3.	www.javaworld.com
4.	java.sun.com/products/javamail/index.html
	Query="Jini"
1.	www.jini.org
2.	java.sun.com
3.	www.artima.com
4.	archives.java.sun.com/archives/jini-users.html
	Query="Servlet"
1.	java.sun.com
2.	www.servletcentral.com
3.	java.sun.com/products/servlet/index.html
4.	archives.java.sun.com/archives/servlet-interest.html

Table 2. Authoritative results of java-related queries

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

Table 3. Ranking of three users (Each user evaluates five documents.)

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

Table 4. Personalized search results (Shade box shows if personalized rank is equal to user-checking's.)

5. 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 with respect 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 some better results.

6. References

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *The Seventh International WWW Conference*, 1998, http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

[2] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *IBM Research Report RJ 10076*, 1997.

[3] L. Introna and H. Nissenbaum, "Defining the web: The politics of search engines," *IEEE Computer*, vol. 33, pp. 54-62, 2000.

[4] S.-M. Chen and Y.-J. Horng, "Fuzzy query processing for document retrieval based on extended fuzzy concept networks," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 29, no. 1, pp. 96-104, 1999.

[5] S.-M. Chen and J.-Y. Wang, "Document retrieval using knowledge-based fuzzy information retrieval techniques," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 5, pp. 793-803, 1995.

[6] C.-S. Chang and A.L.P. Chen, "Supporting conceptual and neighborhood queries on the world wide web," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 28, no. 2, pp. 300-308, 1998.

[7] B. Pinkerton, "Finding what people want: Experiences with the webcrawler," *The Second International WWW Conference*, Chicago, USA, 1994,http://www.thinkpink.com/bp/WebCrawler/WWW94.html.

[8] Yahoo, http://www.yahoo.com.

[9] AliWeb, http://www.aliweb.com.

[10] The Clever Search, http://www.almaden.ibm.com/cs/k53/clever.html.

[11] Google, http://www.google.com.

[12] S. Chakrabarti, B.E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Spectral filtering for resource discovery," *SIGIR 1998 Workshop on Hypertext IR for the Web*, Melbourne, Australia, 1998.

[13] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: A new approach to topic specific resource discovery," *The Eighth World Wide Web conference*, Toronto, Canada, 1999.

[14] D. Lucarella and R. Morara, "FIRST: Fuzzy information retrieval system," *Journal of Information Science*, vol. 17, no.2, pp. 81-91, 1991.

[15] L.A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338-353, 1965.

[16] L.A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems*, vol. 1, no. 1, pp. 3-28, 1978.