ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Indian Journal of Science and Technology, *Vol 9(24), 10.17485/ijst/2016/v9i24/96113, June 2016*

The Potential Knowledge Recommendation System using User's Search Logs

Kinam Park*

Creative Information and Computer Institute, Korea University, Korea;
spknn@korea.ac.kr

Abstract

Background/Objectives: This paper proposes a potential query recommendation system based on the user search history so that information search system users can express their potential information needs in a query, and the information they want can be searched. **Methods/Statistical Analysis:** The proposed system used users' search query to analyze the associative relationship with existing users' search history, and extracted users' potential information needs. The extracted potential information needs are recommended to users in the recommendation query. **Findings**: This paper used 27,656 pieces of search history data for analyzing the utility of the proposed system and conducted a behavioral experiment. The experiment found that the subjects showed a statistically higher level of satisfaction when using the proposed system than when using a general search engine. **Improvements/Applications:** In the future, it will be possible to secure the reliability of recommended queries by expanding and solidifying the search history through researches on personalization.

Keywords: Information Retrieval, Potential Knowledge, Query, Recommendation, Search Log

1. Introduction

Information search refers to a series of processes for analyzing existing data according to users' information needs and searching for information that meets users' information needs. Users' query for expressing information needs in these processes greatly affects search result[1]. If users do not have sufficient understanding and prior knowledge of a certain area, they will have difficulty searching for the information they want using current commercial search engines, and additional efforts to find core keywords for information and express them in the query will be necessary[1,2]. Accordingly, users will waste a lot of time looking for a query suitable for information search, and users who failed to find an appropriate query will modify the query through a repeated search process. In the worst case, users will fail to find desired information, and give up information search.

In general, the query used for search is short and has an ambiguous and implicit meaning on many occasions, and from the viewpoint of systems, if the query is longer than 3 words, search relevance tends to become lower[3,4]. Representative research methods for solving a problem like this are query expansion and relevant keyword[5]. Query expansion is a method in which the information search system performs the search by automatically adding additional information to the user query[1]. Additional information refers to information extracted using the knowledge-based, statistical-based and concept-based method. The knowledge-bases method generally uses the thesaurus to extract additional information. However, the thesaurus is not easy to build, and it is difficult to overcome the sparseness of words, and its accuracy is deteriorated by the ambiguity of the word[3]. The statistical-based method is a method that uses the frequency of co-occurrence of the words in the query, and assumes that

*Author for correspondence

the appearing words are closely related to the same topic. As the words with a high frequency of co-occurrence may appear as the representative keywords in non-relevant documents as well as in relevant documents, however, another problem may occur[3]. Lastly, the concept-based method uses the semantic network that was implemented by extracting the words in consideration of their frequency of occurrence in all documents to be used for the search. However, it is difficult to establish the conceptual relationship between terms. The relevant keyword is a method that analyzes users' query, selects similar search queries in existing users' query history, and recommends them to users. However, this method cannot but be dependent on the frequency of use when selecting words closest to the words in a given query[1]. Accordingly, this paper proposes a potential query recommendation system based on users' search history so that users of the information search system can express their potential information needs in a query, and desired information can be searched.

2. Related Works

As users of the information search system not only express information needs in a query, but also have potential information needs that cannot be expressed in a query, it is not easy for the system to automatically know it. Also, as queries contain many short words with ambiguous meanings, there is not enough information for the system to predict users' information needs[1]. To solve these problems, not only researches on traditional search systems to increase the reliability of search algorithms and information, but also researches to analyze users' search needs are actively conducted.

The Semantic Web Search Engine (SWSE) search engine was developed by the Digital Enterprise Research Institute (DERI). It is a search engine that searches or navigates information expressed in the semantic web standard language from the object-oriented perspective, and supports the interface that searches from the initial keyword for information in units of objects[6]. It expanded the scope to searching Research Data Facility (RDF) resources in units of objects, not the concept of looking for text documents on the web. The SWSE classifies the URIs of about 1 billion RDF documents and collects them, and provides search service. The SWSE handles queries using Simple Protocol and RDF Query Language (SPARQL), which is the W3C standard query, as the internal query engine.

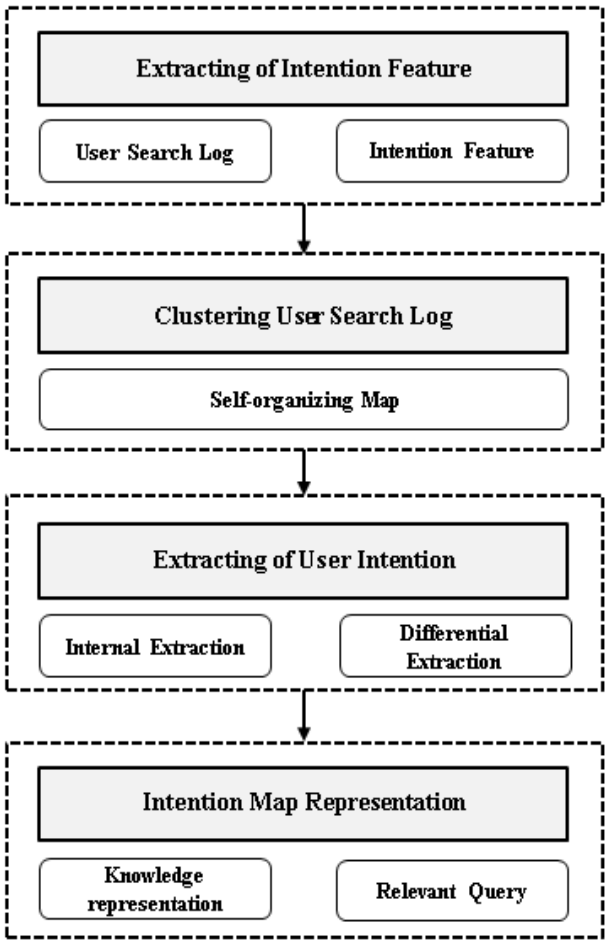

Figure 1. Automatic extraction of user's search intention from web search log.

In Figure 1, a method of extracting user intentions using user search history data is proposed. To improve search performance, the features of users' search intentions were selected using the search history data of existing users, and the clustering algorithms and the users' search intention extraction algorithm were used to automatically extract users' search intentions. As the model provides interface that does not take users into consideration, and uses the frequency-based search algorithm, it cannot secure the reliability of the quality of search results[7].

Reference[8] proposes the Hierarchical Phrase Vector Mode (HPVM) that can satisfy the information needs of personal users with different intentions while emphasizing the importance of user intentions from the viewpoint of personalized search in web-based information search. The HPVM hierarchically expands user intentions through the phrase-based vector model, and uses the Support Vector Machine (SVM) to provide users with documents most suitable for user intentions by sorting positive or negative documents based on learned user intentions. As the HPVM expanded the query through user feedback and took patterns into consideration, however, if there is no user participation, it may affect performance. The initial query pattern learning was not taken into consideration either.

3. Potential Knowledge Recommendation System

To solve problems like existing researches' failure to understand users' information needs and search intentions, and increased costs due to users' information needs, this paper proposes a potential query recommendation system based on user information history. The proposed system used users' search history to extract candidates of users' potential information needs, and judged the association between the extracted candidates and the search history of existing users, and extracted potential information needs. The extracted potential information needs are expressed in a query and recommended to users. In Figure 2, the configuration of the potential query recommendation system based on user search history proposed in this paper. The proposed system extracts query analysis and recommendation queries based on personal history and public history using the queries entered by users to recommend users' potential queries. Frequent items are extracted based on personal history, and reliability is calculated, and frequent items are extracted again based on public history and reliability is measured. That is, if the reliability of personal history-based association rule mining does not satisfy the minimum reliability, the personal history for the query cannot become the potential recommendation query. So public history-based frequent items with similar interests are extracted and reliability is extracted, and then those with more than the minimum reliability are presented as recommended queries.

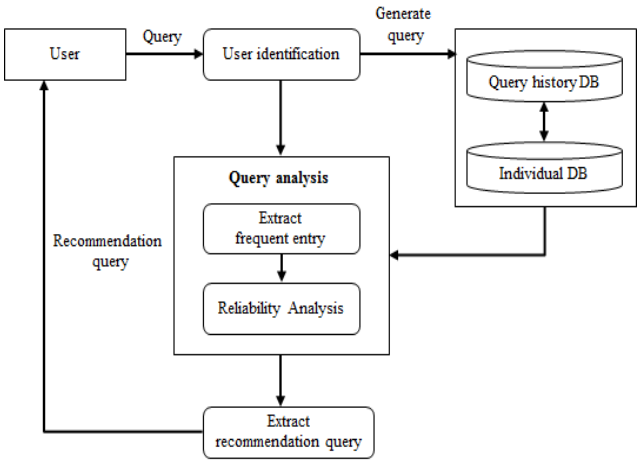

Figure 2. The system architecture.

If both the personal history-based method and the public history-based method exceed the minimum reliability, personal history-based and public history-based queries will be recommended in that order.

3.1 User Indentification

User identification is a method of separating personal history from public history. To this end, the proposed system used the login method and the cookie method. The login method system is not restricted by the environment, and the cookie method has an advantage in long-term data collection. The system saves the queries entered by users in the database as time series. At this time, for user identification, the unique identifier key value is saved together, and if a query the same as the user's initial query appears, it will be weighted.

3.2 Query Analysis and Query Recommendation

This paper analyzed the association between the queries entered by users and the query history to extract recommended query candidates from the search histories saved in the query history database. The association is analyzed through calculation of reliability based on personal history and public history, and daily search histories were extracted and used as the frequent items for calculation of reliability. This paper used the apriori algorithm to analyze the association rule[2]. The apriori algorithm has a weakness, i.e. as the number of comparison items

increases, the computational complexity increases, but it also has strength, i.e. it can be calculated easily and it is easy to understand the result. Therefore, this paper maximized association analysis through daily batch jobs. The query analysis and query recommendation process can be explained by way of an example. Table 1 show the search logs of users "A," "B," "C," "D" and "E," and the search logs consist of the query each user entered initially and re-entered.

User log of input queries calculates the frequency of the same queries of each user to extract frequent item sets. However, the overlapping identical queries of the same user are handled in a single transaction. The group of candidate item sets shown in Table 1 is compared with the search log transaction again, and the process of finding frequent item sets is repeated. Table 2 shows the frequent item sets for the search log history.

Table 1. Query logs

User	Query
A	mango, onion, nintendo, milk, egg, yoplait
B	doll, onion, nintendo, milk, egg, yoplait
C	mango, apple, milk, egg
D	mango, umbrella, corn-flakes, milk, yoplait
E	corn-flakes, onion, onion, milk, ice-cream, egg

Table 2. Frequent entry

Query	Frequency
mango → onion	1
mango → milk	3
mango → egg	2
mango → yoplait	2
onion → milk	3
onion → egg	3
onion → yoplait	2
milk → egg	4
milk → yoplait	3
egg → yoplait	2

Table 3. Final frequent entry

Query	Query
Onion, milk, egg	3
Milk, egg, yoplait	2

The items in excess of the arbitrary weight that were extracted from the frequent item sets for the search log history. This paper set the minimum weight at 3, and values less than the weight were excluded. The final frequent item set is the same as Table 3, and {onion, milk, egg} and {milk, egg, yoplait} show reliability of 3 and 2 respectively, and "milk, egg" can be suggested for the initial query "onion," or "egg, yoplait" for the initial query "milk."

4. Experiment and Evaluation

This paper conducted a behavioral experiment to evaluate the utility of the proposed system, and tested the method of extracting potential queries through analysis. The behavioral experiment measured users' satisfaction with the system. To measure the level of satisfaction, recommended queries were provided to users with information needs and their satisfaction with search was measured.

4.1 Experiment Method

The experiment had 30 subjects. The data used to build the experimental system was the search history logs of the 30 subjects, collected from November 20, 2012 to March 30, 2013, and 27,656 search logs of commercial search systems. As for the experimental methodology, personal history-based and public history-based recom-

Table 4. Evaluation item

Item	Satisfaction measurement
Purpose	Is appropriate information for query provided?
Diversity	Are the answers for the queries rich and various?
Usefulness	Are the answers for the queries complete and helpful?
Distinction	Are the answers for the queries distinguishable from the conventional search systems?
Specialty	Are the answers for the queries valuable above the level of common sense?
Entertainment	How much are you interested in the answers for the queries?
Reliability	Are the answers for the queries realistic and reliable?

mended queries were provided to users with search needs with regard to the same topic and their satisfaction was measured. The level of satisfaction was measured with the Likert Scale. The Likert Scale is a method of measuring attitude that was devised by R. Likert in 1932. This attitude measurement method combined the summated rating scale that pools several opinions about the same contents, puts them on 3~7-point continuous scales, and use the sum as the score for the attitude, and the internal consistency scale based on item analysis. It is possible to measure the level of satisfaction at intervals of one point with 'very inappropriate' being 1 point, and 'very appropriate' being 7 points. The satisfaction evaluation items used for the experiment are as shown in Table 4.

4.2 Result of the Experiment

This paper conducted a paired T-test with regard to the level of satisfaction with the proposed system and the commercial search system to measure the level of satisfaction with recommended potential queries. The test result showed that the average level of satisfaction with the proposed system was 5.73 ("satisfied," SD=.31), and the average level of satisfaction with the commercial system was 4.76 ("so-so," SD=.15). The experimental subjects were more satisfied with the proposed system, and the difference was statistically significant at $p<=0.5$ ($t(12) = -7.392$, $p<= 0.5$).

5. Conclusion and Future Tasks

This paper proposed the potential query recommendation system based on users' search history so that information search system users can express their potential information needs in queries, and desired information can be searched by reducing the difficulty of generating queries. The proposed system used users' search history to extract candidates for users' potential information needs, judged the association between the extracted candidates and existing users' search history and extracted potential information needs. The system was developed so that extracted potential information needs can be expressed as queries and recommended to users. This paper conducted a behavioral experiment to evaluate the utility of the proposed system, and tested the method of extracting potential queries through analysis. The behavioral experiment measured users' satisfaction with the system. Recommended queries were provided to users with information needs, and their satisfaction with search was measured. The result of the experiment showed that the subjects were more satisfied with the proposed system than with general search engines. The difference was statistically significant. The significance of the researches on recommendation of potential queries based on users' search history, which is proposed in this paper, is as follows. For starters, it is possible to automatically extract users' search intentions. It is possible to automatically extract potential information needs based on users' search history, not by extracting keywords appearing in user queries. Also, users' satisfaction with the initial search can be increased by providing a query guide when users generate queries for search. In the future, it will be possible to secure the reliability of recommended queries by expanding and solidifying the search history through researches on personalization.

6. Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education (NRF-2014R1A1A2056200)

7. References

1. Kinam Park, Hyesung Jee, Taemin Lee, Soonyoung Jung, Heuiseok Lim, Automatic extraction of user's search intention from web search logs. Multimedia Tools and Applications. 2012 November; 61(1):145–62.
2. Young-an Kim, Gun-Woo Park, An Efficient Extended Query Suggestion System Using the Analysis of Users' Query Patterns. Journal of the Korean Institute of Communication Sciences. 2012 July; 37(7):619–26.
3. Ji-Hye Kim, Doo-Soon Park. Development of the Goods Recommendation System using Association Rules and Collaborating Filtering. The Journal of Korean Association of Computer Education. 2005 August; 9(1):71–80.
4. Schuemie MJ, Kang N, Hekkelman ML, Kors JA. GeneE: Gene and protein query expansion with disambiguation. Bioinformatics. 2010 January; 26(1):147–48.

5. Xu J, Croft WB. Query Expansion Using Local and Global Document Analysis. Proceeding of the 19th international ACM SIGIR, Switzerland; 1996. p. 4–11.
6. Andreas Harth, Aidan Hogan, Jürgen Umbrich, Stefan Decker. Building a Semantic Web Search Engine: Challenges and Solutions. Proceedings of the 3rd XTech, lreland; 2008.
7. Ashwin Kumaar M, Palani Thanaraj. Feature Extraction of Arterio-Venous Malformation Images using Grey Level Co-Occurrence Matrix. Indian Journal of Science and Technology, 2015; 8(35):1–5.
8. GunWoo Park, JinGi Chae, Dae Hee Lee, SangHoon Lee. User Intention based Personalized Search: HPS (Hierarchical Phrase Serch), Proceedings of the WSEAS, USA; 2008. p. 266–76.