Regular paper

Journal of the KIECS. pp. 587-594, vol. 14, no. 3, Jun. 30. 2019, t. 95, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2019.14.3.587

사용자 개인 프로파일을 이용한 개인화 검색 기법

윤성희*

Personalized Search Technique using Users' Personal Profiles

Sung-Hee Yoon*

요 약

본 논문은 사용자의 검색 의도와 개별 관심을 반영한 순위화된 검색 결과 문서를 제공하는 개인화 검색 기법을 제안한다. 개인화 검색에서는 사용자의 개별 관심사와 선호도를 정확하게 판별하기 위한 사용자 프로파일을 생성하는 기술이 개인화 검색의 성능을 좌우한다. 개인 프로파일은 사용자의 최근 입력 질의어들과 검색 과정에서 참조했던 문서들에 나타나는 주제어들의 가중치와 빈도가 기록된 데이터 집합이다. 사용자 프로파일은 웹 검색에 앞서 사용자의 입력 질의어를 개인화된 질의어들로 확장하기 위해 사용된다. 중의적 질의어의 정확한 의미를 결정하기 위해서 워드넷을 사용하여 프로파일에 등록된 단어들과 의미 유사도를 계산한다. 검색 시스템의 사용자 측에 질의확장 모듈과 순위 재계산 모듈을 확장모듈로 구축하여 진행한 실험에서 개인화 검색 기술을 적용한 실험 결과가 상위문서들에 대해서 정확률과 재현률이 크게 향상된 성능을 보이고 있다.

ABSTRACT

This paper proposes a personalized web search technique that produces ranked results reflecting user's query intents and individual interests. The performance of personalized search relies on an effective users' profiling strategy to accurately capture their interests and preferences. User profile is a data set of words and customized weights based on recent user queries and the topic words of web documents from their click history. Personal profile is used to expand a user query to the personalized query before the web search. To determine the exact meaning of ambiguous queries and topic words, this strategy uses WordNet to calculate semantic similarities to words in the user personal profile. Experimental results with query expansion and re-ranking modules installed on general search systems shows enhanced performance with this personalized search technique in terms of precision and recall.

키워드

Personalized Search, User Preference, User Profile, Query Expansion, Re-Ranking
개인화 검색, 사용자 선호도, 사용자 프로파일, 질의 확장, 재랭킹

Ⅰ. Introduction

An information retrieval system is a software tool that takes a user's input query, searches a vast number of documents, selects the documents that fit the objective of the query, rank those documents, and return a set of candidate answers. The quality of an information retrieval system is measured by the degree to which it selects documents appropriate to the intent of the query, and the trustworthiness of its ranking.

* 교신저자: 상명대학교 소프트웨어학과
• 접 수 일 : 2019. 05. 16
• 수정완료일 : 2019. 05. 31
• 게재확정일 : 2019. 06. 15
• Received : May. 16, 2019, Revised : May. 31, 2019, Accepted : Jun. 15, 2019
• Corresponding Author : Sung-Hee Yoon
Dept. of Software, Sangmyung University
Email : shyoon@smu.ac.kr

Research on user query behavior in web searching show that users generally query common words rather than technical terms, and use approximately one or two simple words as rather than complex search formulas[1]. By contrast, accurately detecting the meaning of the user's search query and distinguishing users' personal interests is a very important step in improving the performance of search results. In particular, ambiguity of query words leads to document selection simply by matching the lexical patterns of the query words, rather than meaning, which is one of the major reasons for the failure to reflect the user's specific query intent in the search process. The same query word "apple" or "virus" undoubtedly has very different meaning for a user interested in technologies or softwares compared to a user interested in farming or biology. Thus, users' expectations of the improvement of search engine performance is increasingly turning towards high-quality intelligent features that tailor search results to show only documents that fit the individual's interest fields. Individualized or personalized search techniques are not only an important improvement in information retrieval, but also a key technology with great value for personalized advertisements and marketing.

This paper proposes a personalized web search method that takes users' search history and the topic words of referenced documents to infer users' tendencies and interest fields, builds a personalized user profile based on this information, and searches documents that fit individual interests. Query expansion and re-ranking modules are installed on user local site for personalized search using this profile, and the results are evaluated in terms of precision and recall as a traditional evaluation methods for information retrieval.

II. Related Works

2.1 User Search Pattern Analysis

Fundamentally, information retrieval systems decide the appropriateness of a given document by the match between the input query and the indices in the web document. Related researches showed that typically most users input very basic query words, and then search again based on their evaluation of the relatedness of the returned results to find more related documents[1]. Researches also tells the average query length is merely 2.3 words, and consists of single words or compound nouns rather than complex queries such as Boolean expressions. As a result, there is frequent ambiguity in which the user's precise intention is difficult to discern, and the performance is necessarily low. The performance of the search result - that is, how well the set of result documents matches the user's query intent - is typically measured in terms of precision and recall. Simply put, precision is the percentage of result documents that fit the query intent, and recall is the frequency that a document expected to appear in the results is actually returned. Researches show that users typically only look at the top few documents, so precision in search systems appears to be more important than recall.

2.2 Personal Interests and Preferences

A profile is the collection of data about a user's personal preferences or recent interests. There are two ways of collecting this data. The static collection method prompts users to directly input their interests, while the dynamic method infers them by continuously gathering information about preciously searched documents. Some recent research involves collecting keywords on personal interests and creating concept networks between these words as profiles[2-3].

Research that analyzes a cohort of users sharing

a common interest field and reflects that information in personal profiles tend to gather information based on communities such as blogs or member-based internet cafés[4]. These researches introduced methods to connect personal interests with the interests of other users to represent the tendencies of many users in a linked structure. This method uses the tags of folksonomy (also called classification by people) to cluster documents, and takes advantage of the fact that tags are often used as query words[5].

Because the short query words input by users does not provide enough information to produce precise search results, others have proposed search systems that suggest additional related keywords for the users to select, or apply the users's relevant feedback[6]. Theses systems require the users to manually select related query words in the search process, or require search systems with external information to use tags or identify related topic words. Other systems use data-heavy resources such as ontology[7]. But these resources are not suitable to reflect the personal preferences or interests on user local site.

2.3 Semantic Relatedness and Ambiguity

One essential issue that must be resolved in personal web search is identifying the precise meaning of ambiguous words so that they fit the user's intent[8]. For example, the frequent topic words of documents queried and searched by users interested in Apple products compared to users interested in travel and food are much different set. Various resources such as ontology and thesaurus are used in research on resolving ambiguity. One of the useful resource is WordNet as a knowledge base that systemizes the semantic relatedness of vocabulary in a hierarchical method. It was developed and released by Princeton University, and is currently being used widely to research the semantic structure of English vocabulary[9]. It is a massive language database used frequently in natural language processing. Because it represents the concept and relationship between words well, it is often very useful in machine translation for tasks such as choosing target words or substitutes. Much research currently uses Princeton WordNet, as a basis to expand into various world languages, usually by translating the English WordNet then tuning the results[10-11]. This research introduces a personalized web search model that uses a WordNet to calculate the semantic similarity between the query and the topic words in their profile as well as result re-ranking.

Ⅲ. Query-Based Personalized Search

The proposed personalized search model consists of two essential steps: pre-processing to expand the user's query, and post-processing to re-rank the results. Compared to systems that shows related search words, this model expands the query by using implicit search behavior such as query history and document clicks to automatically select keywords related to the search topic. Furthermore, this model uses WordNet to resolve query word ambiguity and to determine the meaning of topic words, and collects the topic words of documents read by the user to update their personal profile.

3.1 Resolving Query Ambiguities

The precise meaning of an ambiguous word can be determined by the proximity to other topic words that reflect the user's interests. In WordNet, the distance between nodes is the semantic distance between words, and represents a quantifiable measure of their conceptual similarity. Although there are many algorithms to calculate the semantic simialirty, this research used a method based on path length, which is the simplest and most commonly used[12-13]. Since users'

individual sets of query words differ based on their personal interests, the precise intent of ambiguous query words is determined by semantic proximity to this set. The semantic distance *Sim(s,t)* between two WordNet nodes *s* and *t* is determined by the following formula (1). Here, *Distance(s,t)* represents the path length between *s* and *t*.

$$Sim(s,t) = 1/\ Distance(s,t) \qquad (1)$$

3.2 Building a Personal Profile

In this research, a personal profile is the set of query words input by the user as well as topic words collected from web documents the user actually clicked on. As search behavior is repeated, the weights are updated with emphasis on query words related to recent interests. When query word ambiguity exists, WordNet is used to calculate the semantic similarity of the topic words. Specifically, a user's personal profile consists of the following information and Fig. 1 shows the personal profile and modules in personalized search.

User's personal profile =
{ recent query words }
∪ *{ topic words of clicked documents }*
∪ *{ meaning of words based on WordNet }*
∪ *{ topic words weights based on frequency }*

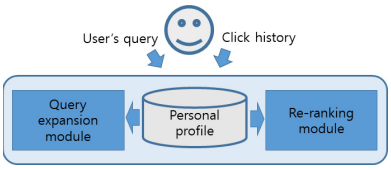

Fig. 1 Personal profile and modules

3.3 Query Expansion and Re-ranking Modules

The user profile is applied twice in the personalized search process. First, the recent interest topic words in the profile are added to the simple input query during the query expansion process. Secondly, it is used to re-rank the result documents in order of interest. The user's query word is expanded into two or three words by adding topic words with high weight in the profile, and inserted into a common search engine. The act of clicking on documents returned through such a search is an implicit indication of the appropriateness of that document. In particular, given research that shows that users tend to focus only on the first 5 or so documents, deciding the order of documents is a very important part of evaluating the performance of the search. In this research, overlapping between the topic words in the result document and the interest topic words saved in the user's profile is calculated. Furthermore, topic words and weights are updated for documents that the user actually clicks on. The rank value of documents in D_{qe}, the set of documents returned by the expanded query, is calculated according to the following formula (2), where d_q^r is the r-th ranked document in D_{qe}.

$$T(d_q^r) = 1\ /\ \log(r+1) \qquad (2)$$

In addition, the following formula (3) is used to measure an overlapping of documents returned after query expansion was incorporated into the re-ranking calculations. Here, d_{qi} is the set of documents returned by the query q_i, which is the original query expanded by the i-th keyword.

$$T(d_q) = \sum_{i=1}^{k} T(d_{qi}^r) \qquad (3)$$

IV. Experiments and Evaluation

The personalized search technique suggested in this paper was built on client-side as expanded module on existing search systems. Fig. 2 shows the architecture.

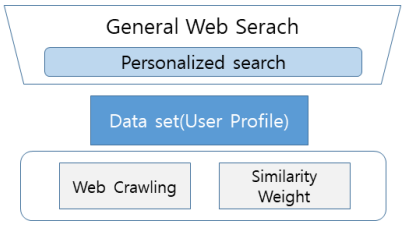

Fig. 2 Architecture of personalized search with user profile

The participants were 30 university students with varied interests such as music, travel, computers, programming, and sports. Initial user profiles were built based on one week's queries and clicked documents. Table 1 shows a general summary of the experiment environments.

Table 1. Experimental environments summary

number of users	30
average number of tested queries	52.3
maxmum length of query expansion	3
size of ranked documents	30
maximum rank movement	6

As the measures of search performance for *top N* ranked documents from the size of *test* set, the precision and recall were calculated according to (4-1) and (4-2).

$$precision(N) = (test \cap top\ N) / N \quad (4\text{-}1)$$
$$recall(N) = (test \cap top\ N) / test \quad (4\text{-}2)$$

Because measuring these two values requires whether each document was relevant and interesting to the user, users themselves manually conducted the evaluation.

To evaluate the performance improvement of the personalized search system, we compared the precision and recall of proposed search system with the existing general search system without the expansion module. The experimental results are summarized in Fig. 3 and Fig. 4. Users' satisfaction with the search results as measured by precision and recall was much improved in the personalized search system, with particularly high satisfaction with the top 5~10 documents.

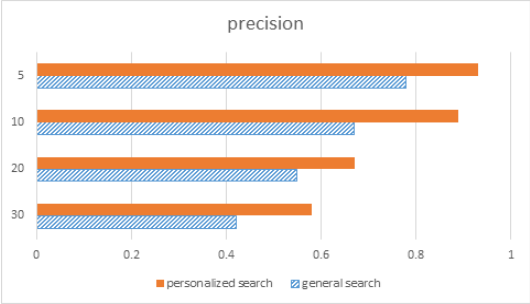

Fig. 3 Enhanced precision of personalized search

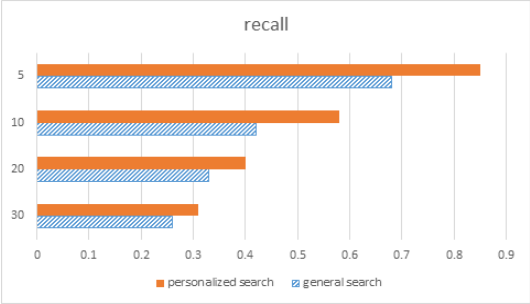

Fig. 4 Enhanced recall of personalized search

Given previous research showing that most users only pay attention to and click on the top ranked documents, this result is clearly meaningful. The importance of performance improvement in the 20~30th documents and lower is comparatively small.

Ⅴ. Conclusion

This paper suggested a personalized web search technique that detects personal interests to provide personalized search results. Built as an locally-expanded module on the user's existing search

environment, the objective of this system is to provide search results based on recent interests maintained in a personal profile by collecting and continually updating topic words related to the user's personal interest. Specifically, the personal profile consists of recent query words input by the user as well as topic words of documents actually clicked on by the user, where the meaning of ambiguous topic words was determined by calculating semantic similarity to other topic words pertaining to the user's interests. Its performance was evaluated via experiments on real users with varied personal interests, comparing the personalized search system with existing search systems. The personalized system showed improved performance, with 93% precision and 85% recall in the top 5 documents.

We anticipate personalized search method that consider personal interests and preferences to be widely applied in fields such as advertisement or marketing. To improve search result quality and increase user satisfaction, continued research that applies various natural language processing research results, such as ambiguity resolution in web document topic classification, can be expected to contribute greatly to the improvement of information retrieval.

References

[1] S. Park, "Analysis and Evaluation of Term Suggestion Services of Korean Search Portals: The Case of Naver and Google Korea," *J. of the Korean Society of Information Management,* vol. 30, no. 2, 2013, pp.297-315.

[2] S. Koratkar and S. A. Takale, "Deriving Concept Based User Profile for Search Engine Personalization," *Int. J. of Science and Research,* vol. 4, no. 6, 2013, pp.3086-3089.

[3] K. R. Remesh and P. Samuel, "Concept Networks for Personalized Web Search Using Genetic Algorithm," *Int. Conf. on Information and Communication Technologies,* Kochi, India, Dec. 2015, pp. 566-573.

[4] G. Park and S. Lee, "Personalized Search based on Community through Automatic Analysis of Query Patterns," *J. of the Korean Institute of Information Scientists and Engineers: Database,* vol. 36, no. 4, 2009, pp. 321-326.

[5] D. Kim, S. Kan, H. Kim, and B. Lee, "Folksonomy-based Personalized Web Search System," *J. of Digital Contents Society*, vol. 11, no. 1, 2010, pp. 105-116.

[6] S. Yoon, "Using Query Word Senses and User Feedback to Improve Precision of Search Engine," *J. of the Korean Society of Information Management*, vol. 26, no. 4, 2009, pp. 81-91.

[7] B. Kim, "Words Recommendation Algorithm for Similarity Connection based on Data Transmutability," *J. of The Korea Institute of Electronic Communication Sciences*, vol. 8, no. 11, 2013, pp. 1719-1724.

[8] Y. Lee and Y. Chung, "An Experimental Study on an Effective Word Sense Disambiguation Model Based on Automatic Sense Tagging Using Dictionary Information," *J. of the Korean Society for Information Management*, vol .24, no. 1, 2007, pp. 321-342.

[9] Princeton University, "About WordNet," *WordNet online Technical report*, 2010.

[10] Y. Kim and Y. Kim, "A Question Example Generation System for Multiple Choice Tests by utilizing Concept Similarity in Korean WordNet," *J. of the Korean Information Processing Society*, vol. 15-A, no. 2, 2008. pp. 125-134.

[11] I. Lee, D. Hwang, Y. Hahm, and K. Choi, "Open Korean WordNet(KWN): Dictionary-based Semi-Automatic Development," *The 26th Annual Conf. on Human & Language Technology*, Busan, Korea, Oct. 2014. pp. 193-196.

[12] L. Meng, R. Huang, and J. Gu, "A Review of Semantic Similarity Measures in WordNet," *Int.*

J. of Hybrid Information Technology. vol. 6, no. 1, 2013. pp. 1-12.

[13] T. Simpson and T. Dao, "WordNet-based semantic similarity measurement," *Engineering Applications of Artificial Intelligence,* vol. 36, Mar. 2016, pp.80-88.

저자 소개

윤성희(Sung-Hee Yoon)

1987년 서울대학교 컴퓨터공학과 졸업(공학사)
1989년 서울대학교 대학원 컴퓨터공학과 졸업(공학석사)
1993년 서울대학교 대학원 컴퓨터공학과 졸업(공학박사)
1993년 ~ 현재 상명대학교 소프트웨어학과 교수
1999년 ~ 2000년 University of Michigan 방문연구교수
2007년 ~ 2008년 University of Victoria 방문연구교수
※ 관심분야 : 자연어처리, 정보검색 등