ISSN 2349-0292
Impact Factor 2.365

Global Journal of Advanced Engineering Technologies and Sciences

AN IMPROVED PERFORMANCE MEASURE FOR USER PREFERENCE BASED SEARCH

N.S.Sindhu*, B. Vinitha, B. Vanitha, S. Sibi

* Assistant Professor Department of Computer Science and Engineering KPR Institute of Engineering and Technology Coimbatore, TamilNadu, India.
Research Scholar Department of Computer Science and Engineering KPR Institute of Engineering and Technology Coimbatore, TamilNadu, India.
Research Scholar Department of Computer Science and Engineering KPR Institute of Engineering and Technology Coimbatore, TamilNadu, India.
Research Scholar Department of Computer Science and Engineering KPR Institute of Engineering and Technology Coimbatore, TamilNadu, India.

Abstract

Today internet usages are rapidly increasing. Users are more dependent on the web search for needs of the information. In existing system the search engine provides the information that are needed for the user but it does not give the information that are relevant to the user's interest. Search engine does not have much know about users interest. Personalization methods are introduced to improve the information retrieval system. In the proposed system personalization is done by three phases: a) Building user profile b) Applying re-ranking technique c) Performance measure. User Profile is an input for performing personalized web search. The user profile which captures the user interest are built from past search history of the user. User Profile contains basic profile and updated profile. We generate Profile score (PS) from basic profile score (BPS) and converged profile score (CPS). Basic profile score is obtained explicit from the user preferences whereas token extraction and count of hit links are used to calculate the converged profile score (CPS). By identifying the needs of individual user's preference we can greatly improve the search engine performance. We fetch the top N results from the search engine and the weight score (WS) is calculated using term frequency and inverse term frequency (TF-IDF) The cosine score (CS) is calculated using cosine similarity. The cosine similarity is used to measure the similarity between the set of documents. The Re-ranking Score is calculated for providing the better results to the user. It can be calculated as sum of profile score, weight score and cosine score. The relevancy of search results is measured using tanimoto coefficient and levenshtein distance. Thus the qualities of search results are progressed.

Keywords: Personalized web search, User Profile, Re-Ranking, Performance measure.

Introduction

Users are more dependent on the web search for the needs of information. The huge amount of information in internet is rapidly increasing day by day; it creates challenges for web search. If identical queries are submitted by different users, a search engine could return the same result. Search engine does not know what we the user essentially wants. This paper proposes the enhanced personalization technique to satisfy the need of user. The user profile includes two levels basic profile and updated profile. The basic profile contains the primary information given by the user such as name, e-mail id, location and preference. The updated profile contains user preference inferred from search history. Search history is used to retrieve the information about user. The profile score is calculated from basic profile score and converged profile score. Basic profile score is calculated, explicitly from the user interest whereas converged profile score is calculated using token extraction and count of hit links.TF-IDF is used to calculate the weight score(WS), where term frequency measures how frequently the keyword occurs in the search result. Inverse document frequency measures how important keyword occurs in search result. And the cosine score (CS) is calculated using cosine similarity which is used to measure the similarity of two text documents. On applying Re-ranking technique the relevance of search results are improved. The Re-ranking score (RS) is computed for providing the top results, that is calculated by sum of Profile Score, weight score and cosine similarity score. The relevancy of search results is measured using tanimoto similarity and levenshtein distance. The Tanimoto Coefficient is used to measure the overlap

of set of documents. The Levenshtein distance (LD) is a measure of the similarity between the set of documents where, the change in order of the search results can be measured. Thus the optimized results are displayed to the user.

Related work

When the queries are issued; the search engine may return the same results to users. Different users may have entirely different information needs and goals when using exactly the same query. For example, a person may query "cookie" to get information about snacks or junk foods, while programmers use the same query to find information about computer programs. When such a query is issued, search engines will return a list of documents that combine different search results.

A new approach for creating and recognizing automatically the user behavior of a computer user is presented in [1]. In this case, a computer user behavior is represented as the sequence of the commands user types during work. This sequence is transformed into a distribution of relevant subsequences of commands in order to find out a profile that defines its behavior. Also, because a user profile is not necessarily fixed but rather it changes, we propose an evolving method to keep up to date the created profiles using an Evolving Systems approach. This paper combines the evolving classifier with a trie-based user profiling to obtain a powerful self-learning online scheme. They also develop further the recursive formula of the potential of a data point to become a cluster center using cosine distance. The novel approach proposed in this paper can be applicable to any problem of dynamic user behavior modeling where it can be represented as a sequence of actions.

The novel approach for measuring personalization of web search [2] to measures the personalization by running multiple searches on the same query and comparing the search results. It shows that, logged in user and IP address of user has considerable impact on the search results rather than search history.

In [3] the author shows that, user behavior data can extensively improve ordering of top results in real web search setting. It is examine alternatives for incorporating feedback into the ranking process and explore the contributions of user feedback compared to other common web search features. It performs a large scale evaluation over 3,000 queries and 12 million user interactions with a popular web search engine, establishing the implicit feedback. They compared two alternatives of incorporating implicit feedback into the search process, namely reranking with implicit feedback and incorporating implicit feedback features directly into the trained ranking function. It is shown that the accuracy of a competitive web search ranking algorithms by as much as 31% relative to the original performance. Personalization process includes negative preferences in [4] personalization strategies. But these are all document based methods, and it could not return users interests.

In Personalized Ranking of Search Results with Learned User Interest Hierarchies from Bookmarks [5], a user profile is build, called (UIH) user interest hierarchy; it uses the web pages in user's bookmarks and the Divisive Hierarchy Clustering (DHC) algorithm. They also proposed a scoring function for personalized ranking with the UIH learned from bookmarks. It can score a page based on the user profile and the results returned by a search engine. The novel approaches for [6], User profile explanation of user interest can be used by search engine to provide personalized search result. In this study, it explores the use of a less-invasive means of gathering user information for personalized web search. In particular we build user profile based on user activity at the search site itself and study the use of these profiles to provide personalized web search results. By implementing a wrapper around the Google search engine, they were able to collect the information about individual user search activities. In particular, they collected the queries for which at least one search results were examined, and the snippets for examined results. User profiles were created by classifying the collected information into concepts in a reference concept hierarchy. These profiles were then used to re-rank the search results and the rank order of the user examined results before and after re-ranking was compared. Sieg et al. [7] also used ODP to learn user profiles for personalized web search. ODP currently contains more concepts/nodes, so they only use a few top levels of categories in the ODP hierarchy. Hence, the user profiles do not envelop the low level categories, which are more specific. Consequently, this may reduce the ranking quality for individuals with more specific interests, not represented as high level categories in ODP.

The novel approaches for [8], with the exponential growth of the available information on the World Wide Web, a traditional search engine, have difficult to analyze the user activities. Consequently, it is very important in many applications for example in an e-commerce Web site or in a scientific one for the search system to find the information which is more relevant to the user. Personalized Web environments that build models of short-term and long-term user needs based on user actions, browsed documents or past queries are playing an increasingly crucial role: they form a

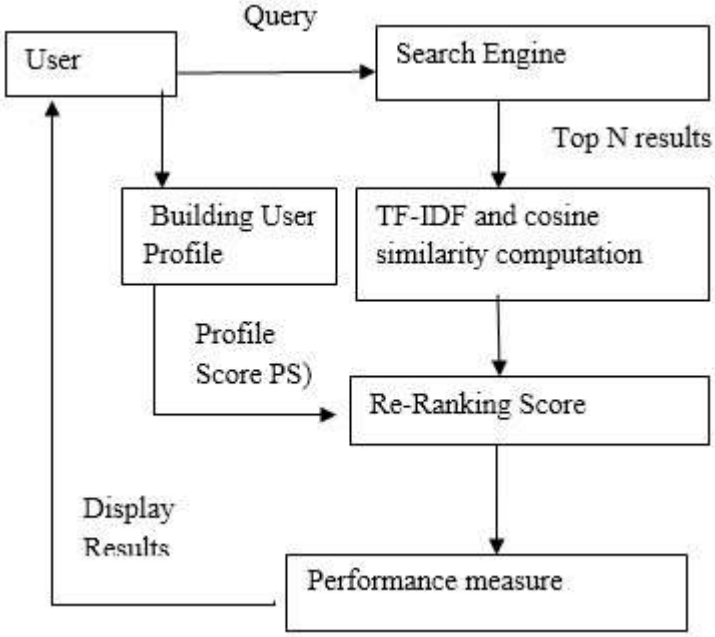
winning combination, able to satisfy the user better than impersonalized search engines based on traditional Information Retrieval (IR) techniques. Several important user personalization approaches and techniques developed for the Web search domain.

In [9], this work presents the algorithm to manage a user's profile in the web search engine. In our approach is aimed to combine several search criteria into a scoring rule that is used to rank documents similar to a query. A profile as a set of user's preferences is used to define the importance of each search criterion in the scoring rule and, thereby, retrieve documents according to the specific search model. The profile management algorithm exploits relevance feedback in order to build profiles. I extend the basic algorithm by clustering techniques that allows constructing of aggregated, "global" profiles. In order to make the approach more flexible and useful for the web I propose the algorithm that is aimed to minimize a number of users' interactions with the system on the assumption of keeping search quality on the high level. Finally, I evaluate the developed approaches on real data and study the effectiveness of the profile management.

Proposed system

The proposed system is mainly for increasing the level of personalization. When the user enters the query, the search engine will return the top N search results. Mean while, the user profile which captures the user interest from the past search history of the user is built. The user profile contains two profiles: a) Basic profile contains basic information about the users. b) Updated profile is inferred from the past search history. We generate Profile score (PS) from basic profile score (BPS) and converged profile score (CPS). Basic profile score is obtained explicit from the user preferences whereas token extraction and count of hit links are used to calculate the converged profile score (CPS). Weight score is calculated for the top N results by TF*IDF and Cosine Score is calculated using cosine similarity. Re-Ranking score is computed by the sum of profile score, weight score and cosine score. To measure the relevancy of search results using tanimoto coefficient and levenshtein distance.

Fig1.Architecture of proposed system:



Components of proposed system:

CREATING USER PROFILE:

A user profile is a collection of data and also be considered as the representation of a user model. User profile refers to the representation of a person's identity and behavior. The user profile which captures the user interest are built from past search history of the user. User profile contains basic profile and updated profile. Basic profile contains basic information about the user and updated profile is inferred from the past search history of the user. Basic profile score is explicitly obtained from the user whereas the converged profile score (CPS) is calculated by token extraction and count of hit links. We generate the profile score (PS) from the basic profile score (BPS) and converged profile score

(CPS) it's applicable for regular user whereas generate the profile score (PS) from the basic profile score (BPS) it's applicable for new user. The count of hit links can be defined as, if p is a page with outgoing-link set O(p) and each outgoing link is associated with a numerical integer indicating visit-count (VC), then the weight of each outgoing link connecting to page p to page o is calculated by,

Weight $_{link}$ (p, q) = [VC (p , q)] / VC (p, q')

Tf-idf and cosine similarity computation:

Term Frequency and Inverse Document Frequency is used for calculating the weight score to obtain the top N results. Term Frequency also known as TF measures the number of times a term (word) occurs in a document. Inverse Document Frequency is to find the relevant documents matching the query.

$w_{t,d}=\log(1+tf_{t,d}) * \log_{10}(N/df_t)$

The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d. df_t is the document frequency of t, the number of documents that contain t. df_t is an inverse measure of the in formativeness of t.

Cosine Similarity is a similarity metric that can be used to measure the similarity of two text documents. Documents can be represented by vectors. Similarity is then measured as the angle between the two vectors. This method is useful when finding the similarity between two text documents whose attributes are word frequencies. A perfect correlation will have a score of 1 (or an angle of 0) and no correlation will have a score of 0 (or an angle of 90 degrees).

$\cos(q_i, d_i) = (q_i . d_i) / |q_i| . |d_i|$

q_i is the tf-idf weight of term i in the query and d_i is the tf-idf weight of term i in the document.

Re-ranking:

User's usually view the first few pages of search result. Re-Ranking score is computed for providing the better results to the user. It can be calculated as, sum of Profile Score (PS), Weight score (WS) and Cosine Score (CS). Re-Ranking score is,

$$RS=PS + WS + CS$$

Re-Ranking score is calculated for reorder the top N results and providing the better results to the user based on the user preference.

Performance Measure

The relevancy of search results is measured using tanimoto coefficient and levenshtein distance. The Tanimoto Coefficient is used to measure the overlap of set of documents. The Tanimoto Coefficient is found from the following equation: In the equation, X and Y are data objects represented by vectors. The similarity score is the dot product of X and Y divided by (the squared magnitudes of X and Y minus the dot product.

$$w_{t,d} = \log(1+tf_{t,d}) \times \log_{10}(N/df_t)$$

Tanimoto (x, y) = x. y / ($\| x \|^2 + \| y \|^2 - x. y$)

Where $\| x \|$ means the length of the vector x.

The Levenshtein distance (LD) is a measure of the similarity between the set of documents where, the change in order of the search results can be measured.

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(q-p)^2}$$

Where p is set of documents and q is a set of documents. Using above method, to measure the relevancy of search results is compared to the original search results and give the better results to the user which is more relevant to the user preference. Thus the optimized results displayed to the user.

Conclusion

In existing system the search engine provides the information that are needed for the user but it does not give the information that are relevant to the user's preference. In proposed system, we improve the level of personalization by building the user profile, on applying re-ranking technique to the top N results and then relevancy of web search results is measured.

ISSN 2349-0292
Impact Factor 2.365

Reference

1. Agapito Ledezma, Araceli Sanchis, Jose Antonio Iglesias, Plamen Angelov (2012), "Creating Evolving User Behavior Profiles Automatically", ieee Transactions on knowledge engineering, VOL. 24, no. 5.
2. Aniko Hannak, Piotr Sapie 'zy' nski, Aresh & Molavi Kakhki, "Measuring personalization of web search", Proc 22nd International Conference on WWW (2013) 527-538.
3. E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. ACM SIGIR, 2006.
4. W. Ng, L. Deng, and D.L. Lee, "Mining User Preference Using Spy Voting for Search Engine Personalization," ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.
5. H. Kim and P. Chan. "Personalized ranking of search results with learned user interest hierarchies from bookmarks". In O. Nasraoui, O. Zaine, M. Spiliopolou, B. Mobasher, B. Masand, and P. Yu, editors, "Web Mining and Web Usage Analysis", pages 158–176. Springer, 2006.
6. M. Speretta and S. Gauch. "Personalized search based on user search histories". In Proc. Intl. Conf. Web Intelligence, pages 622–628, 2005.
7. A. Sieg, B. Mobasher, and R. Burke." A large-scale evaluation and analysis of personalized search strategies". In Proc. CIKM, pages 525–534, 2007.
8. [Micarelli et al. 2007] Micarelli, A., Gasparetti, F., Sciarrone, F., Gauch, S.: "Personalized search on the World Wide Web"; In: Brusilovsky, P., Kobsa, A., Nejdl, W., (ed.), The Adaptive Web, Springer-Verlag, Berlin, Heidelberg, (2007), 195–230.
9. "User Profile Management in a Web Search Engine" by Vladimir Eske in June 2004.