ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Improving Personalized Web Search Quality in Information Retrieval

L. Shobby Neeta Fancy[1], A.Rajamurugan[2]

PG Scholar, Dept of Information Technology, Regional Centre of Anna University, Coimbatore, Tamil Nadu, India [1]

Teaching Fellow, Dept of Information Technology, Regional Centre of Anna University, Coimbatore, Tamil Nadu, India [2]

ABSTRACT: Personalized Web Search (PWS) is the fine grained search which is mainly used in the field of Information Retrieval (IR). One of the major barriers in the personalization of web is improving the quality of web search. The proposed system mainly focuses on web search quality improvement by implementing profile based personalization in both active and passive manner. After the effective retrieval of Search Engine Result Pages (SERPs), the ranking of SERPs can be performed by the Personalized PageRank(PPR) algorithm.

KEYWORDS: Personalized Web Search, Information Retrieval, PageRank , HITS, Personalized PageRank.

I. INTRODUCTION

As a result of huge piled up collection of information in the web, retrieving the needed information is critical. For the effective retrieval of information from the web, many web search techniques can be used. Normally, the generic web search returns more than thousands of **S**earch **E**ngine **R**esult **P**ages (SERPs) to a query which may contain many irrelevant results.

To avoid the retrieval of irrelevant results, the intention behind the user's query is needed. For this, **P**ersonalized **W**eb **S**earch (PWS) has been introduced as the future of web mining. PWS provides the different SERPs or reorganizes the SERPs differently for the same query issued by the user who have different interests and different preferences.

For example, consider a query "mouse". A biological researcher may be interested to retrieve the information related to "rat". But an engineer shall like to retrieve the information about the computer peripheral.

One of the major issues in PWS is improving the web search quality. The quality of the web search is defined as, the relevant, needed result of the user being retrieved as a top result.

A. Motivations

The existing system focuses to prevent the privacy of the user who is involved in the personalized web search.

1) Disadvantages of the existing system

- It does not automatically construct the user profile. It just mapped the user's preferred topic with some taxonomy of knowledge.
- It only focuses the passive personalization technique. It does not implement the active personalization.
- In passive personalization, the user involvement to the system is not needed. Even though it implements passive personalization, the interaction to the user is needed to get the sensitive topics and the sensitivity level for each topic.

B. Contributions

The proposed system mainly focuses on two techniques namely active and passive personalization to personalize the web search for individual user. The main objective of the proposed system is to provide the web content personalization which provides the relevant SERPs to the user according to his/her interests and preferences.

1) Advantages of the Proposed System

- The proposed system contains both active and passive personalization techniques.

- Distinct from the existing system, the proposed system does not need the user interaction during passive personalization.
- It proposes the Personalized PageRank(PPR) algorithm for the effective preference ordering of the SERPs.

II. Related work

Garvie Brown K. Wilson R. and Shamos M. [1], described the techniques of personalization. The main objective of their work is, to learn about web content personalization and popularly used personalization techniques namely,

- Active Personalization
- Passive Personalization

Personalizing a website means providing content which is specifically relevant to the user. In active profiling website users are asked to complete a registration forms that requests their personal information and specific interests. In passive profiling, user profile can be constructed based upon the queries which can be previously issued by the user and the clicked URLs.

Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli [2], recommended the user profiles for personalized information access. This existing work surveys some of the most popular techniques for collecting users information and build the user profiles.

In this system user profile is represented in a following manner.

- Keyword Profiles
- Semantic Network Profiles
- Concept Profiles

Fang Liu, Clement Yu and Weiyi Meng [4], proposed a technique in personalized web search for improving retrieval effectiveness. Web search personalization is to carry out retrieval for each user incorporating his/her interests. This work proposed a novel technique to learn user profiles from users search histories. The user profiles and general profiles were then used to improve retrieval effectiveness in Web search.

Lidan Shou, He Bai, Ke Chen and Gang Chen [5] have suggested a technique for supporting privacy protection in personalized web search. This existing system focused on profile based personalization in passive profiling technique. It has collected the implicit information from the user side and maps this into the Open Directory Project (ODP). The implicit information depicts the user’s interests and preferences. It may be the previous query issued by the user, desktop files, amount of time spent on the SERPs, clicked Uniform Resource Locators (URLs) and so on.

Mirco Speretta and Susan Gauch [6], proposed a technique for personalizing search based on user search histories. The main objective of this technique is retrieving the interested results of the users by click-log based methods. This existing work built a system that creates user profiles based on implicit collection of information, specifically the queries submitted and the user-selected results.

III. Pws Quality Improvement Procedures

The proposed work is having the following modules namely,

- Inverted Index Construction
- Active Profile Construction
- Passive Profile Construction
- Query-Concept Detection
- Re-Ranking

Figure 1 clearly depicts the overall functionality of the proposed system.

A. Inverted Index Construction

Inverted Index is used for the indexing purpose in web search. It allows fast search in the information retrieval process. It is named as inverted index because, it takes {document-term} collection as an input and turns it into a {term-document} collection as an output. So that, it is called as an inverted index. It maps the words to the collection of documents in which the words appears.

There are two main variants of inverted index.

- Record Level Inverted Index
- Word Level Inverted Index

1) *Record Level Inverted Index*

A record level inverted index contains a list of references to documents for each word. This is also called as inverted file index or inverted file.

2) *Word Level Inverted Index*

A word level inverted index additionally contains the positions of each word within a document. This is also called as full inverted index or inverted list.

The proposed system is making use of the record level inverted index. In record level inverted index, every word contains a list of references for documents.

B. *Active Profile Construction*

The active profile is constructed based on the active involvement of the users. That is, the web surfer involves eagerly refining their search results. Here, user preferences can be gathered and that can be represented as a vector. This vector contains a list of terms which is the list of preferences from the web surfer's side.

Web surfer's explicit feedback is also very helpful for the active profile construction. Relevance feedback mechanism can be implemented by the proposed system. The initial response from the search engine may not satisfy the user's information need if the query is ambiguous. For this, the query refinement will be needed. Relevance feedback automates the query refinement process.

1) *Process of Relevance Feedback*

Initial responses of the search engine for a given query can be presented together with a rating form. This rating form contains a binary options i.e. SERPs are useful/useless. After that, one of the options will be selected which is presented as a second round of input from the web surfer for correcting the ranks of the SERPs based on the web surfer's preferences.

C. *Passive Profile Construction*

Passive profile will be constructed by the implicit feedback mechanism. That is, without the user intervention, the profile will be constructed automatically. To construct the user profile, the following factors will be considered.

- Desktop files of the user
- SERPs which has been clicked by the user.
- Amount of time spent by the user with this SERP
- Query which will be issued by the user.

The above factors are called as the usage patterns. It depicts the user access and navigational patterns from the web. After accessing these patterns a hierarchical user profile will be implemented to provide the SERPs relevant to the user.

D. *Query Concept Detection*

In this module, the user's query is mapped with the inverted index to retrieve the relevant web pages. This is the procedure of the normal web search. But in the personalized web search, the retrieved SERPs are mapped with the profile of the user to provide the appropriate, intended and preferred results to the user. This is the process of query concept detection.

E. *Re-Ranking*

After the query concept detection, the relevant responses are ranked based on ranking algorithms and re-ranked according to every user profile. Google uses PageRank (PR) algorithm for the ranking purpose. The proposed system uses the PPR algorithm for ranking the search engine response pages. Some of the re-ranking algorithms are listed below.

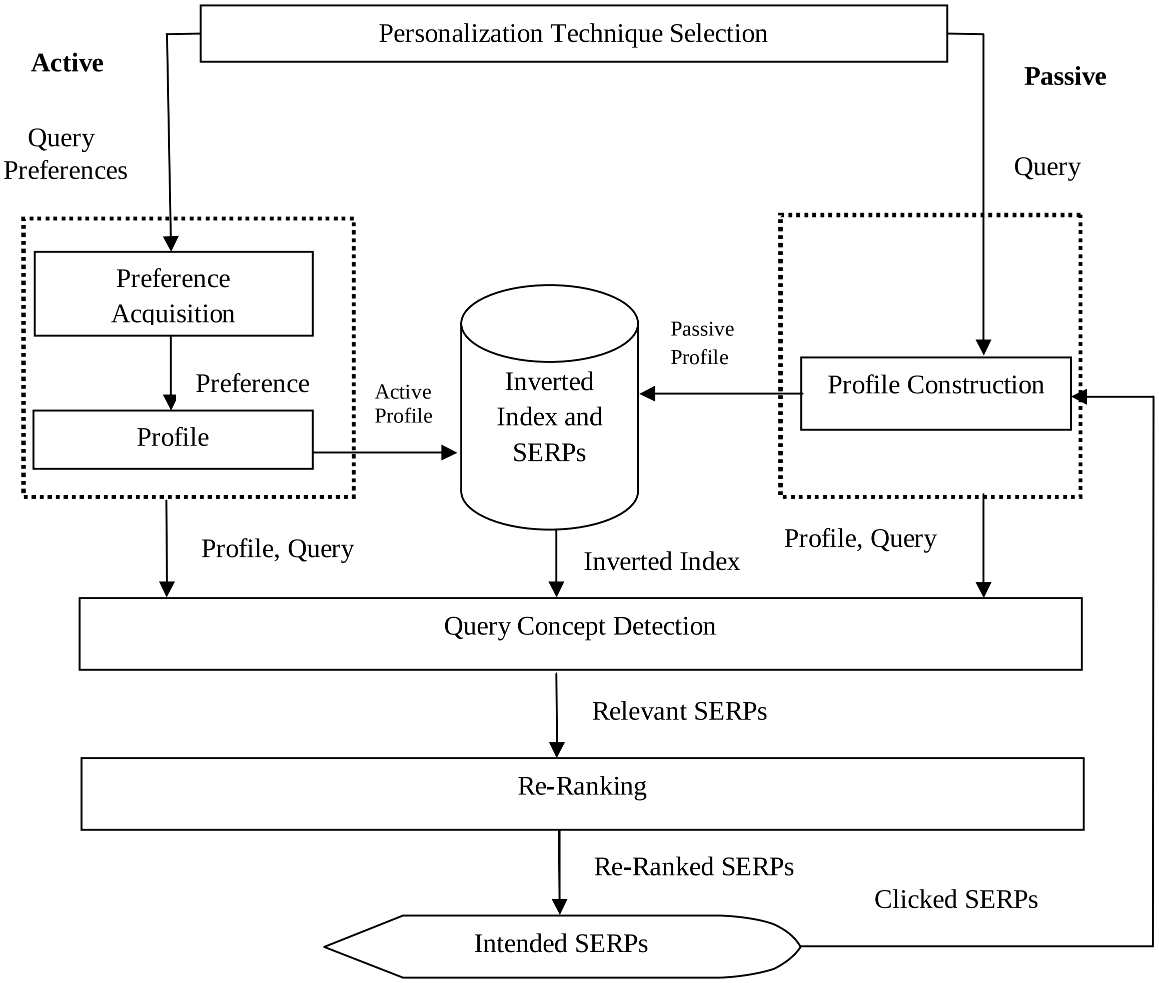


Figure 1. System architecture of PWS quality improvement

- Sorting based on overall clicks a document has got historically.
- Click-Sorted ranking Scheme.
- **P**age**R**ank (PR)
- PPR
- Dynamic Ranking
- **H**yperlink **I**nduced **T**opic **S**earch (HITS)

Both PageRank and HITS are called as link based ranking schemes. In PR and HITS, the Web is modelled as a directed graph.

- In PR, the incoming links are the good first order indicators to assign priority scores to the web page. PR is assigned to every web page independent of the query.
- In HITS, two popular and prominent pages are the central components for ranking of SERPs. They are,
  Authorities - It is the root in which multiple pages are linked and referred that authority.
  Hubs -It contains a comprehensive set of links to authorities.

According to HITS, each and every page contains these two popular measures. HITS mainly depend on the query issued by the user to rank the SERPs. The PPR algorithm is discussed in Section IV.

IV. EXPERIMENTAL METHODOLOGIES

In this section the algorithms and techniques which have been used for PWS quality improvement are presented.

A. *Inverted Index Construction*

Technique 1: Inverted_Index (Document-Term)

Input : {Document-Term} Collection
Output : {Term-Document} Collection

1. The initial step is tokenizing the given document. For a text document, tokens may be regarded as any non empty sequence of characters not including spaces and punctuation.
2. Find the frequency of occurrences of each and every word i.e. term.
3. Update the term, documents in which it occurs and the frequency of each term into the database.
4. The final output {term, documents} collection will be obtained after the construction of an inverted index.

B. *Active Profile Construction*

Technique 2: Active_Profile(User's_preferences)

Input : User's preferences
Output : Active Profile

1. Acquire the user preferences and arranged them in a vector representation.
2. Use the explicit feedback mechanism to refine the search engine response pages.
3. Relevance feedback is used as an explicit feedback mechanism.
4. Provide binary rating form with the initial responses to the given query.
5. Using the preferred vector and binary rating form the active profile is constructed.

C. *Passive Profile Construction [7]*

Algorithm 1: BuildUP(n, D, minsupport, δ)

```
Input: a node n, supporting documents D, thresholds minsupport and $\delta$
Output: A user profile UP
Split( n, D, minsupport, $\delta$)
for each child $c_i$ labeled $t_i$ of node n:
BuildUP($c_i$, S($t_i$), minsupport, $\delta$)
```

Algorithm 2: Split(n,SD(t),minsupport, δ)

Input: a node n labeled term t, supporting documents SD(t), thresholds minsupport and δ

```
Generate the frequent term list {$t_i$} with $D(t_i) \geq$minsupport sorted by the descending order of frequency.
for each term $t_i$
  if $Sim(t_i, t_k) > \delta$, where $k < i$,
    Set the node label as $t_i/t_k$ , and $SD(t_i /t_k) = SD(t_k) \cup D(t_i)$
  else if $P(t_k \mid t_i) > \delta$, where $k < i$,
6         Keep the node label as $t_k$, and $SD(t_k) = SD(t_k) \cup D(t_i)$
7     else
8         Create a new node with label $t_i$, and $SD(t_i)=D(t_i)$
9 Calculate Support($t_i$) for each node with label $t_i$, and that will be presented in descending ordering.
```

D. Query-Concept Detection

Technique 3: Query_Concept(q,UP)

Input : Query ‘q’, User Profile UP
Output: Relevant SERPs

1 For a given query ‘q’, retrieve all the supporting document S(q) from the inverted index.
2 Overlap S(q) with the **U**ser **P**rofile (UP).
3 Display the SERPs which are matched with the User Profile.

E. Re-Ranking [3]

Algorithm 3: PPR(Relevant_SERPs,UP)

- Input : Relevant SERPs, UP
- Output : Re-Ranked SERPs

1 Calculate PPR using the following formula.

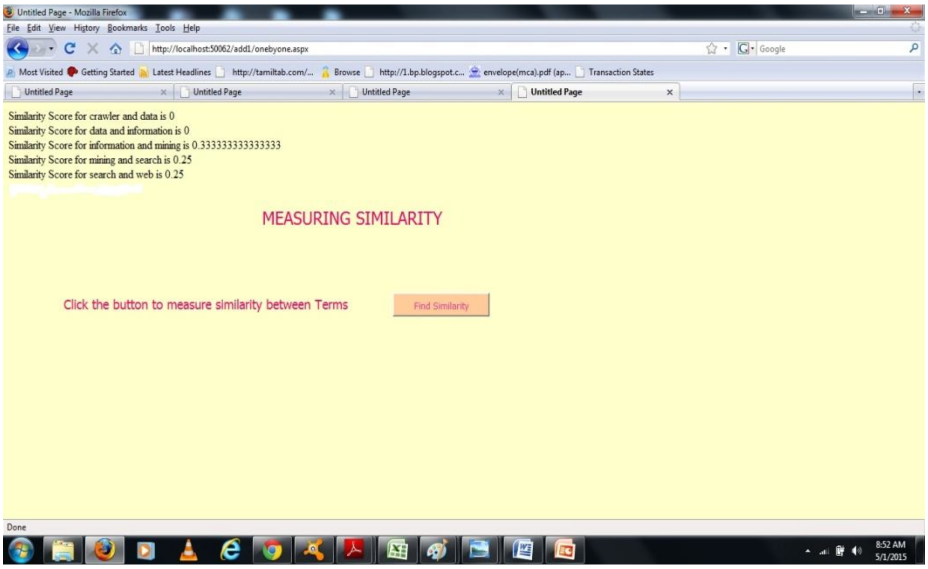
$$Pr_w(A)=(1- d)+[d*(w(T_1)(Pr(T_1)/C(T_1))+\ldots.+w(T_n)(Pr(T_n)/C(T_n))] \qquad (1)$$

- $Pr_w(A)$ => Weighted PR or Personalized PR
- $Pr(T_i)$ => PR of parent page T_i
- $w(T_i)$ => Normalized Weight Factor
- $C(T_i)$ => Number of out links of Parent Page T_i
- T_i => i^{th} parent of Page A
- d => Dumping Factor or Tuned Constant

2 Display the SERPs which are having highest PPR.

V. RESULT AND PERFORMANCE EVALUATION

The figure 2 shows, the similarity score between two terms. It indicates how similar the two terms are. The similarity score takes value between 0 and 1. To find similarity, Jaccord function is used.



Similarity Score for crawler and data is 0
Similarity Score for data and information is 0
Similarity Score for information and mining is 0.333333333333333
Similarity Score for mining and search is 0.25
Similarity Score for search and web is 0.25

MEASURING SIMILARITY

Click the button to measure similarity between Terms

Figure 2. Finding Similarity

The figure 3 compares the performance of active and passive profiling techniques against the time for different queries like mouse, tree and dental hygiene. The X-axis denotes the profiling techniques and the Y-axis denotes the time in milli seconds. The passive profiling technique gives accurate results in the short time comparing to active profiling because in passive profiling there is no need for user interaction to refine the SERPs.

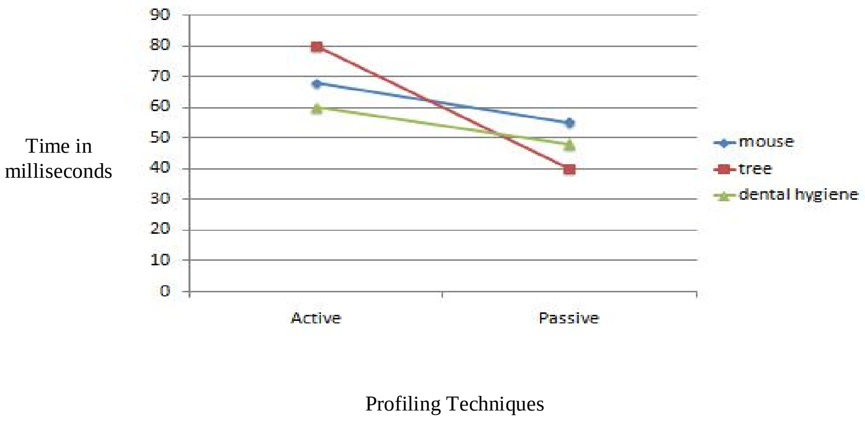


Figure 3.Performance Evaluation

VI. CONCLUSION AND FUTURE WORK

The main contribution of the proposed work is improving the personalized web search quality in the field of information retrieval through profile based personalization. That is, the relevant, needed result of the user must be retrieved as a top result according to their interests and preferences. The direction towards the future work is obtaining the better personalization results by selecting the profile parameters depending on the query characteristics.

REFERENCES

1. Garvie Brown K. Wilson R. and Shamos M. 'Personalization' Available from: <http//www.cs.jyu.fi/ai/vagan/Personalization.ppt>.
2. Gauch S. Speretta M. Chandramouli A. and Micarelli A.,"User Profiles for Personalized Information Access," Springer, pp.54-89,2007.
3. Mehmet S. Aktas, Mehmet A. Nacar and Filippo Menczer," An Application of Personalized PageRank Vectors:Personalized Search Engine".
4. Liu F. Yu C. and Meng W. ,"Personalized Web Search for Improving Retrieval Effectiveness," IEEE Transactions on Knowledge and Data Engineering, vol. 16, No.1, pp.28-40, 2014.
5. Shou L.,Bai H.,Chen K. and Chen G.,"Supporting Privacy Protection in Personalized Web Search," IEEE Transactions on Knowledge and Data Engineering, vol. 26, No.2, pp.453-467, 2014.
6. Speretta M. and Gauch S., "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf., Web Intelligence (WI), 2005.
7. Xu Y. Zhang B.Chen Z. and Wang K. "Privacy-Enhancing Personalized Web Search,"International World Wide Web Conference Committee(IW3C2),2007.

BIOGRAPHY

L.SHOBBY NEETA FANCY has obtained B.TECH. Information Technology in University College of Engineering Tindivanam (A Constituent College of Anna University Chennai) and doing M.TECH. Information Technology in Regional Centre of Anna University, Coimbatore.

Mr.A.RAJAMURUGAN has obtained M.TECH. Mainframe Technology in Regional Centre of Anna University, Coimbatore and he is working as a teaching faculty in Department of Information Technology in the same regional centre.