ISSN 2348 – 7968

Innovative Personalized Architecture in Case of Web Search Users

Y.Raju, Dr. D. Suresh Babu
1Geethanjali College of Engineering and Technology, IT Department, Hyderabad, India
Email: raju.yeligeti@gmail.com
[2] Head, Departments of Computer Science, Kakatiya Government College, Kakatiya University
Email: sureshd123@gmail.com

Abstract

Web search engines provide users with a Large number of results for a submitted query. However, not all return results are relevant to the uses needs. In this paper, we proposed a new web search personalization approach that captures the user's interest and references in the form of concepts by mining search results and they click through. In this paper an effective mixture personalized re-ranking search approach is proposed by modeling user's search wellbeing in a conceptual user profile and then exploiting this profile in the re-ranking process. In this each concept in the user profile consist of two types of documents: categorization document and viewed document Taxonomy is used to represent the user general interest as it contains information from web pages originally associated with open dictionary project category. Viewed documents are used to represent the user's specific interest as it contains information from the web pages clicked by the users. Finally the system create a semantic profile of the user's by monitor and analyze the user's search history. The search results generated will utilize and incorporation of various techniques including clustering, re-ranking and semantic user profile to enhance the performance of the web search engine.

Keywords: user profile, personalization, taxonomy, open dictionary project

I NTORDUCTION

With an enormous growth of the internet most existing search engines such as Yahoo, Google and MSN present user's a single order linear list of pages along with their partial content ranked by the relevance to the search query. Therefore, current retrieval systems are not adaptive enough to satisfy user search needs. Furthermore some keywords[11] have different meanings in the search query such as “Ajax". For this query, users might have a choice of different answers. In case of “Ajax” the Search engine returns "Ajax web based development", "Dutch Foot ball team Ajax Amsterdam " or "cleaning product Ajax". So the webs do not provide adequate information to identify the user names.

Moreover, user's might not choose the right words that best identify their names a recent study demonstrated that user's with more than 7 years of online searching experience obtain much more relevant documents than user's with less experience[1]. The object of web search personalization[5] is to consider the user's search preference and wellbeing in the search process to provide each user with the results that are most relevant to his interests. One of the challenges to personalization is how to identify the user benefit. Another test is to effectively explode these interests in the retrieval system to improve the search results. In particular personalized search engine can be achieved by re-ranking search results returned by a traditional search engine according to the user profile might be constructed from the user's search or browsing behavior.

Another approach is the results are categorize by different topics and user's click results for the user current query are observed to re-order results according to the user's current needs. The user profiles constructed with reference to a topical ontology [8] to categorize user's visited pages then re-ranking is perform by computing the numerical course to check the relevance of search results for a given query against the user profile. The user profile[2] is composed of queries submitted by the user associated with the URL's and topics of the clicked results for each query. Re-ranking is finished by identifying Queries from the user's profile that are similar to the user's current Query then comparing the topics of these related query with the topics of the search outcome .

In this paper an effective hybrid personalized re-ranking[2] search approach is projected by modeling user's search interest in a conceptual user profiles and then exploiting this profiles in the re-ranking process. Finally the amalgam re-ranking process of search results is performed by semantically integrates user's general and specific awareness from the user profile together with the rankings of the traditional search engine.

II Related Work

Most personalization approaches are based on construct a user profile that aim to collect information about the user's topics of curiosity to improve the quality of information retrieval. In order to put up user profile[10] information may be collected either explicitly or implicitly. Explicit information is collected in a straight line by asking the user where as implicit information collected by monitoring the user activities. Profiles that are adapted to the user's changing interest are called dynamic, where as profiles that maintains same information is called static.

In personalized search system user profile can enhance web search quality in one of three phases namely: "part of the retrieval process, query modification, or re-ranking [6, 2] part of the retrieval process Phase: in this Phase user profiles are built into the search process, and are utilized to score web documents. This method of search systems is forced by time constraints. So personalization process as a time consuming process. Query modification Phase: In this user profiles[6, 10] are extended only to the submitted keywords in the query without changing the ranking procedure. Therefore, lists of results are not highly affected by query modification phase. Ranking Phase: When a user submits a query, the search results are obtained from backend search engines. The search results are combined and re-ranked according to the user's profiles trained from the user's previous activities.

One of the forms of representing user profiles is by setting weighted keywords. In keyword profiles, the users can directly provide the system with his interesting keywords or the system can extract keywords from the user's visited pages. The score, number of users interest represents weighted keywords. The main problem with the keyword profiles is the ambiguity exists in words then it might affect the accuracy of keyword profiles.

Another form of user profiles representation is the semantic network -based profiles in this each node represents a concept which represents the user's specific interest in a collection of words and hits synonyms. However, constructing search semantic network profiles is not easy because terms that represent each concept are not predefined. Another efficient model for representing user's interest is the concept profile. These profiles are constructed with a predefined matching between concepts and vocabulary. In the concept based profiles, nodes take action not represent specific words or synonym words instead of this concept profiles represent abstract concepts (Topics) that are interesting to the user.

Another personalization [5] method represented two types of user profiles adopted to the users changing interests. The first type is long term profiles that store visited pages topics as part of the Google directory together with the number of visits for each topic. The second type is short term model that stored user's history of recently visited pages. Considering the entire search history, re-ranking is achieved by computing the similarity between the user profile hierarchical and current search results topic. However, not all information in the user profile reflects the user's current search interest for given query. Furthermore proposed a concurrent re-ranking of search results with no need to store users search history. As the user selects a result, the information included on that page is used to identify user's search needs. However, it has been proven that search strategies used for immediate updates not matched by the users interest, Even though they give more accurate results.

Now a recent study proposed the user profile based on concepts which are groups of words that co-occur frequently in web snippets of visited web pages, Here concepts are organized in the profile as a tree with the relationship between these concepts. these relationship include similar or parent-child relationship. Now weights are assigned, re-ranking is done by assigning scores to current web snippets for given Query based on the aggregation of hits concepts weights.

III Proposed Architecture

In this document, we propose a personalized search system that involves creating concept based user profiles from user search history with reference to ODP concept hierarchy. In the proposed approach, the user profile is enriched with two dissimilar types of information for each concept: taxonomy document, and viewed document. The taxonomy document includes keywords from documents originally associated with topics from the ODP directory. The re-ranking is based on user's general interests and matches in certain query' topic as well as considering the ranks of the non- personalized search engine.

The proposed system consists of four main modules as shown in Figure 1:

Module 1: Preparing the reference taxonomy (or concept hierarchy)

Module 2: collect user information

Module 3: Learning and constructing the user profile

Module 4: Search personalization by exploiting the user profile to re-rank search results.

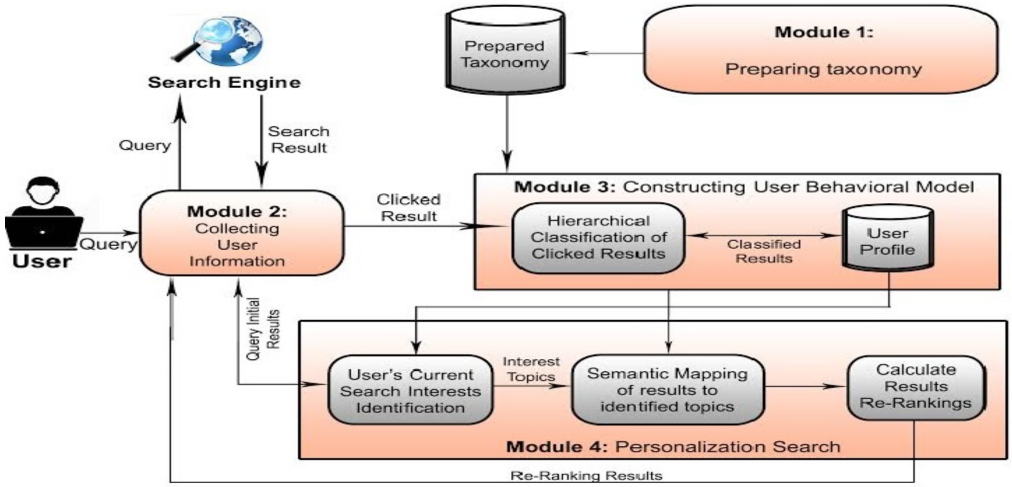
Search Engine
Query
Search Result
Prepared Taxonomy
Module 1:
Preparing taxonomy
User
Query
Module 2: Collecting User Information
Clicked Result
Module 3: Constructing User Behavioral Model
Hierarchical Classification of Clicked Results
Classified Results
User Profile
Query Initial Results
User's Current Search Interests Identification
Interest Topics
Semantic Mapping of results to identified topics
Calculate Results Re-Rankings
Module 4: Personalization Search
Re-Ranking Results

Figure 1: Personalized Search Engine Architecture

3.1 Module 1: Preparing Reference Taxonomy

In this paper, the user profile is constructed with reference to a concept hierarchy or taxonomy of topics. For this purpose, Open Directory Project (ODP) [3] is utilized as our reference taxonomy. The Open Directory Project is an open content directory of the web that is produced and preserved by a group of volunteer editors. Topics in the ODP and web pages that belong to these topics are organized using hierarchical ontology schema as shown in Figure 2.

In order to get a precise concept hierarchy, some changes should take place because some parent-child links are not conceptual. For example, some topics are divided geographically, while others are divided alphabetically to separate content. Furthermore, some topics may have fewer children while others may have hundreds. Additionally, some topics may be associated with many web pages, while others may have fewer pages. Therefore, in order to improve the profiling accuracy, parent-child topics that are not conceptually related are eliminated together with those topics that have too few Web pages linked to them, In order to represent the reference taxonomy, we choose the first 30 URLs for each concept based on the order in which they are represented by ODP. Terms from the 30

pages are collected in one document for each concept. The (Term Frequency –Inverse Document Frequency, TF-IDF) mechanism is then used to
weigh each term from 0 to 1 in each document Eq.(1) which is then normalized by the
vector magnitude because documents are not the same length Eq.(2)

Term weight, $tc_{ij} = (tf_{ij} * idf_{i})$ (1)

Where tf ij is the frequency of term i in document j,

Idf_{i} =Log (Number of documents in D / Number of documents in D that contain ti)

D = the collection of documents that represent the ODP concepts i.e. one document for each concept.

Normalized term weight, $ntc_{ij} = (tc_{ij}/$ vector_lengthj) (2)

Where vector_ lengthj = Σ tcij (3)

3.2 Module 2: Collecting User Information

In order to implicitly collect information about users, the Google wrapper [12] stores the information such as user's submitted queries, returned search results, and user clicks.

Google wrapper ***performs*** the following:

- Capture the results returned from the search engine,
- Record them together with the query and the user ID,
- Pass the query with the returned results to Search Personalization module to apply the proposed re-ordering method,
- Then show the re-ordered results to the user.
- If a user clicks on a result, the wrapper records the clicked page in conjunction with the user ID in the log, prior to redirecting the browser to the proper web page. This log is then exploited in the User Profile Construction module to update the user profile.

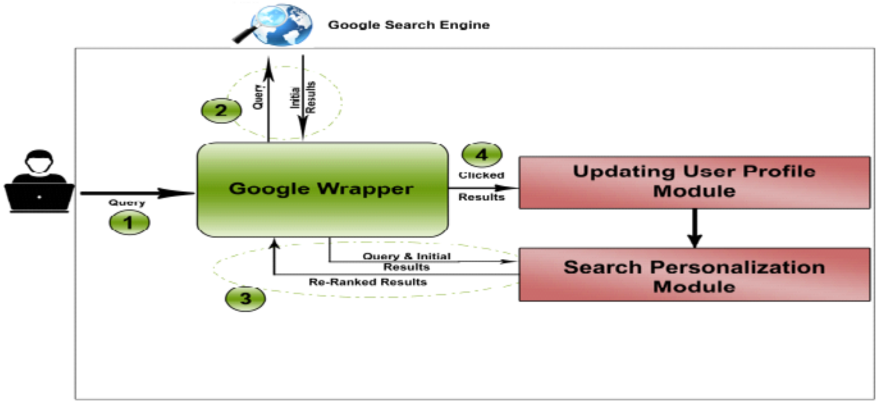

Figure 2: Collecting User Information with Google Wrapper

3.3 Module 3: Constructing the user profile

In this module, data is obtained by observing user search history. This profile is mainly an instance of the ODP reference taxonomy [3]. Specifically, the search results clicked by the user are classified into concepts from ODP which are then used together to build the profile.

Only 0.03% of the pages that are known to the search engines are classified by ODP. So, the hierarchical classification method is used in order to classify clicked search results into ODP concepts. Hierarchical classification starts by matching the document to the best category (concept) at the top level and then “stepping down” the concept hierarchy by matching the document into subcategories of that category only. This method provides better accuracy of the highest matching category.

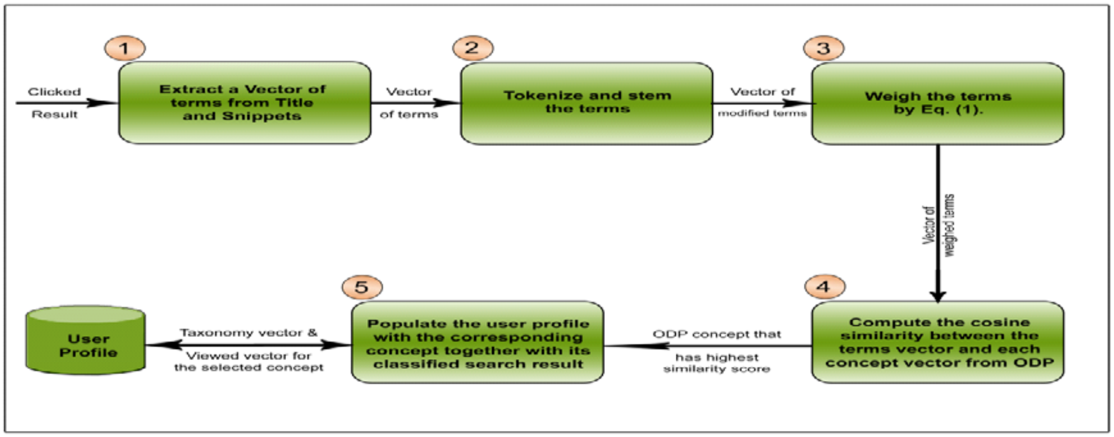
1
Clicked
Result
Extract a Vector of terms from Title and Snippets
Vector
of terms
2
Tokenize and stem the terms
Vector of
modified terms
3
Weigh the terms by Eq. (1).
Vector of weighed terms
4
Compute the cosine similarity between the terms vector and each concept vector from ODP
ODP concept that
has highest similarity score
5
Populate the user profile with the corresponding concept together with its classified search result
Taxonomy vector &
Viewed vector for the selected concept
User Profile

Figure 3: Steps for constructing the user profile for a clicked result

In the second process, Porter Stemmer is used to stem terms of each result. In the next process, the hierarchical classification method is used in order to classify search results into appropriate concepts from the ODP.

In the last process if the concept already exists in the profile, the new classified result is concatenated with the past clicked results under this concept and terms weights are normalized to create a document called **viewed document.**

Taxonomy document-It includes a vector of weighted terms of information originally collected from the reference taxonomy. This kind of document shows an overview of various topics categorized into an ODP concept.

Viewed document-This kind of document represents a user's specific interest at a particular concept.

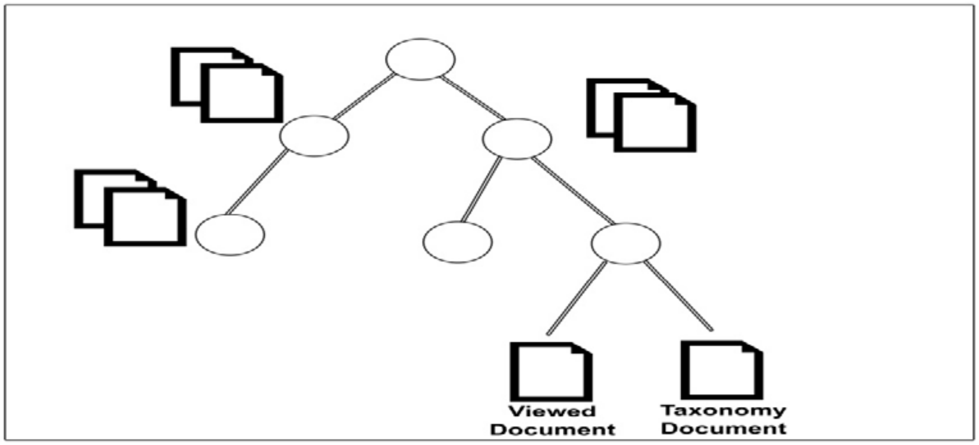

Figure 4: Enhanced Concept-based User Profile

3.4 Module 4: Search Personalization

In this module, a hybrid personalized re-ranking methodology is applied to provide users with more relevant search results for the top. For a given query Search Personalization is achieved in 3 steps:

- Identifying user's topics of interest of current search.
- Semantic Mapping of search results to the identified topics.
- Calculating search results re-ranking sources.

- **Identifying user's topics of interest for current search**

As a first step, the query submitted by the user is matched to the user profile to choose concepts that are highly similar to a user for the current query. For this purpose, the cosine similarity is computed between the query and user's profile taxonomy documents.

- **Semantic Mapping of search results to the identified topics**

After selecting the concepts that represent the user’s query, search results are semantically mapped to these concepts. This step is necessary to measure the relevance of each result with the concepts selected from the user profile.

- **Calculating Search Results Re-ranking Scores**

Re-ranking search [6, 2] results is the last step in the proposed personalized web search approach. For example, a user may be interested in certain parts of a concept. In this case, the viewed documents should be considered greatly when re-ranking search results. Nevertheless, personalized search results could be provided only if such viewed documents hold adequate information about users’ interests.

IV Conclusion

Personalized web search provides users with results that accurately satisfy their specific goal and intent of the search. In this paper, a hybrid personalized search, re-ranking approach is proposed based on constructing a conceptual user profile and exploiting it in re-ranking search results. The user profile consists of concepts obtained by hierarchically classifying user’s clicked search results into categories from the concept hierarchy, Open Directory Project. Each concept in the user profile Consists of two types of documents; taxonomy document and viewed document. Taxonomy document is used to represent the user general interests as it contains information from web pages originally associated with such ODP category. Viewed document is used to represent the user specific interests as it contains information from web pages clicked by the user. Finally, for a given query, search results are re-ranked by semantically mapping them to the general user and specific interests from the profile together with rankings of the basic search engine.

References

1. Al-Maskers and Sanderson 2011] Al-Maskari, A., Sanderson, M.: “The effect of user characteristics on search effectiveness in information retrieval”; Journal of Information Processing & Management, 47, 5, (2011), 719-729.
2. Li et al. 2009] Li, L., Yang, Z., Kitsuregawa, M.: “Rank optimization of personalized search”;
3. Open Directory Project, 2012, http://www.dmoz.org

4. Monickaraj, A. Pankaj Moses; Vivekananda, Dr K.; Prabhu, K.. An Upgraded Focus On Link Analysis Issues In Web Structure Mining**.** American Journal of Computer Science And Information Technology (Ajcsit)**,** [S.L.], V. 1, N. 1, P. 01-09, Dec. 2013. Issn 2349-3917.

5. Kumar, R., Sharan, A.: "Personalized web search using browsing history and domain knowledge"; Proc. International Conference on Issues and Challenges in Intelligent Computing Techniques, IEEE, (2014), 493-497.

6. Hawalah, A., Fasli, M.: "A Hybrid Re-ranking Algorithm Based on Ontological User Profiles"; Proc. 3rd Computer Science and Electronic Engineering Conference, 2011

7. Bibi, T., Dixit, P., Ghule, R., Jadhav, R.: "Web search personalization using machine learning techniques"; Proc. Advance Computing Conference, IEEE, (2014), 1296- 1299.

8. Andhare, A.A., Mahajan, N.V.: "Personalization of Web Knowledge Using ontology model"; International Journal of Advance Research in Computer Science and Management Studies, 2, 2, (2014), 348-353.

9. Antoniou, D., Plegas, Y., Tsakalidis, A., et al.: "Dynamic refinement of search engines results utilizing the user intervention"; Journal of Systems and Software, 85, 7,(2012), 1577-1587.Open Directory Project, 2012

10. A., Mobasher, B., Burke, R.: "Web search personalization with ontological user profiles"; Proc. 16th ACM conference on Information and KnowledgeManagement, (2007), 525-534.

11. Journal of Universal computer Science, vol.20, no. 9 2014, 1232-1258

12. Google Search Engine, 2012, http://www.google.com.eg/

13. Google Ajax API, 2012, http://googleajaxsearchapi.blogspot.com/

14. Kumar, A.Sharma "Personalized web search using Browing history and domain knowledge" International Conference on Issues and Challenges in Intelligent Computing Techniques IEEE 2014, 493- 497

15. Kim, Chan, P. K "learning implicit users interest hierarchy for personalized re-ordering of web search results" web intelligent conference 2009, 105 – 116

16. Bipi, T.Dixit, P.,Ghule, R., Jadhav, "Web Search Personalization Using Machine Learning Techniques

Author Biography

Y Raju Working as an Associate professor in IT Dept at GCE, Hyderabad. He received M.Tech from JNTUH. Presently Pursuing Ph.D from JNTUH. He has published papers in international journal and conferences. His main research area includes Data Mining, Information retrieval System and Artificial Intelligence.

Dr.D.Suresh Babu is currently working Head, Department of Computer Science, Kakatiya Government College, Kakatiya University, Warangal India. He has received his Ph.D Degree in Computer science & Engineering from Acharya Nagarjuna University, Guntur, A.P., INDIA. His main research interest includes Data Mining, neural networks, Information retrieval System and Artificial Intelligence. He has been involved in the organization of a number of conferences and workshops. He has been published more than15 papers in International journals and conferences.