ISSN 2229-5518

Design and implementation of Personalised Information Retrieval Using User Profile and Ontology

Aditi Sharma

Abstract— In present scenario, with the unprecedented development of electronics and tremendous growth of the Internet, World Wide Web (WWW) has become common and popular source of information for varied category of users. Web search engines provide interface between users and web to extract the desired information from WWW. The user possess query to search engine and in response the search engine return useful information. It is observed that typical search engines return the same result for the query submitted by different users irrespective of specific user need. Generally, each user has specific information requirement. It is always desirable that search result should satisfy the exact user requirement. We have proposed approaches to make the search adapting to satisfy the user need. The approaches discussed here are based on ontology and dynamic user profile.

Index Terms— World Wide Web (WWW), ontology, dynamic user profile, user possess query, search engines, exact user requirement, adapting

1 INTRODUCTION

WWW is a vast resource of information growing continuously. This information makes search more and more difficult with traditional search entities as they return large data for a given query consisting of relevant as well as irrelevant information. This not only results in wastage of resources and user time but also lead to information overload problem. For example, for the query "Java," some users may be interested in documents dealing with the programming language, "Java," while other users may want documents related to "Coffee." To circumvent this problem, the web search needs to be adaptive and personalised. Personalisation needs user profile and to build a user profile, some source of information about the user required to be collected. This information may be collected – explicitly and implicitly. Commercial systems, such as My Yahoo, explicitly ask the user to provide information to build user profile. Explicit profile creation is not preferred as it puts an additional burden on the user. Other issues related to explicit profile creation are - the user may not accurately report their interests; the profile, so created, remains static while the user's interests may keep changing over time. Hence, the user needs to update the profile. This forced the researchers to opt for implicit profile building based on observations of the user's actions is used in many Researches describes model considers the frequency of visits to a page, the amount of time spent on the page, how recently a page was visited and whether or not the page was bookmarked. In this chapter, an approach is proposed that can be used to make the search adaptive according to each user's need using ontology. Our approach is distinct because it allows each user to perform more fine-grained search by capturing changes of each user's preferences without any user effort. Such a method is not performed in typical search engines. Performance of the proposed approach is evaluated.

- *Aditi Sharma is currentlyworking as a Assistant Professor in Shri Ramswaroop Memorial Group of Professional Colleges, Lucknow, India, PH-+918756151403. E-mail: ksaditi2@gmail.com ,paper id I018724*

2. Ontology

Ontology is formal description of knowledge. It is a set of vocabulary and the semantic interconnection constructed by some rules of interference and logic for a general purpose or a particular domain with a set of specific topics. Ontology defines a set of concepts based on the interrelation ships existing among the concepts. In the Artificial Intelligence and Web Intelligence community, ontology is a set of objects and their conceptual relationships expressing possible facts in a domain. Ontology is an explicit specification of concepts and relationships that can exist between terms. The set of query terms and the relationships among them are reflected in the representational vocabulary with which query expansion is performed. The set of relations such as subsumption IS-A and metonymy PART-OF describe the semantics of the domain. Depending on the knowledge stored, ontology can be categorized into two types: Domain ontology and generic ontology. Domain ontology expert classified information for a domain provides detailed description for the concepts in the domain. It is the set of domain terms and a set of domain knowledge. The domain terms are generated from the abstract description of the domain knowledge, Domain ontology is designed to represent knowledge relevant to a certain domain type ,e.g. medical, mechanical, university etc. The size of the domain ontology depends on the domain it specifies. The contents of the domain ontology need to be updated regularly by the way of domain knowledge updates. A generic ontology stores the lexical relations of the concepts in natural language. It is for general purposes and normally in large size. Sometimes a generic ontology can be extended to domain ontology.

2.1 Key Components of Ontology

Ontology consists of a finite list of terms and the relationships between them. The terms denote important concepts (classes of objects) of the domain and the relationships include hierarchies of classes.

2.2 Purpose and Benefits of Ontology

Fundamentally, ontology is used to improve communication between either humans or computers. The main purpose of ontology is to create a shareable and agreeable semantic resource over a wide range of agents. Building scalable ontology will effectively be a group effort, with ontology growing over time. Therefore, ontology is *shared and scalable computer-based resources*. The ontology can be used as an interchange format by translating between different modeling methods, paradigms, languages and software tools to achieve inter-operability among computer systems.

The ontology is the basis for a formal encoding of the important entities, attributes,
processes and their inter-relationships in the domain of interest. It is another important goal of building ontology. Ontology can deliver many benefits for Systems Engineering such as it may serve as an index into a repository of information to facilitate information search and retrieval. This thesis focuses on this benefit of ontology.

3. Personalized Information Retrieval

Personalized Information Retrieval (PIR) can be defined as the appropriate information retrieval from a large volume of data or information within a user's context, i.e. preference or profile, and also to present the retrieved information appropriately based on the user's context in generic computing environment where any information could be used by anyone. A search query, in Information Retrieval (IR) systems, often results in a long list of results being returned, much of which are not always relevant to the user's information needs. Reasons behind it are two fundamental issues; information overload and information mismatch.
Indeed, contextual retrieval has been identified as a long-term challenge in information retrieval. Allan et al. defines the problem of contextual retrieval as follows: "*Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs*." In order to make web searching personalized or more precise, to provide more effective information, the search process must incorporate User Profiles (UP) rather than considering only the user queries. Ontology has been a basis for the construction of a user model in several personalized systems ranging from information delivery systems to Intelligent Tutoring Systems.

4. Ontology in Personalized Information Retrieval

Ontology has been a basis for the construction of a user model in several personalized systems ranging from information delivery systems to Intelligent Tutoring Systems The retrieval models are based on keyword or term matching, i.e., matching terms in the user query with those in the documents. However, many concepts or objects can be described in multiple ways (using different words) due to the context and people's language habits. If a user query uses different words from the words used in a document, the document will not be retrieved although it may be relevant because the document uses some synonyms of the words in the user query. This leads to low recall. For example, 'document', 'file' and 'article' are synonyms in the context of *piece of information*. If the user query has the word 'document', relevant results that contain 'article' or 'file' (but not 'document') will not be retrieved. Researcher reported Word Net Ontology Based Model for Web Retrieval in order to solve the above problem.

5. User Profiling for Personalized Information Retrieval

Researchers, investigating personalization techniques for Web Information Retrieval, encounter a challenge; that the data required for performing evaluations, like query logs and click-through data, is not widely available due to privacy issues. Researchers have to perform user study; however, such experiments are often limited to small samples of users, restricting some-what the conclusions that can be drawn.

Researchers in describes the importance of information categorization and user profiles in PIR and suggests generic user profile modelling. An author describe personalising information access in digital libraries through user profiles and discusses various ways to gather data categories and methods to capture user preferences, suggesting three unique ways, namely, the document content category, the document structure category and the document source category. Hochul proposed adaptive web profile using Genetic Algorithm. But, previous methods for building user profiles have some drawbacks, among which users' privacy violation is the main concern. Sugiyama proposed time based user profile considering user's permanent and short-term preferences. Our approach has also taken care of privacy violations.

6. Recommender Systems

Recommender systems are software applications that provide personalized advice to users about products or services they may be interested .They recommend items to users, based on preferences they have expressed, either explicitly or implicitly. Recommender systems accumulate user feedback in the form of ratings for items in a given domain and make use of similarities and dissimilarities among profiles of several users in recommendation of an item.

The two main types of recommender systems are:

1) C*ollaborative filtering systems*: recommended items are based on the similar tastes and preferences liked by people in the past.

This original form of CF-based recommendation systems suffers from three problems:

i) Scalability

ii) Sparsity
iii) Synonymy

2) **Content-*based recommender systems*:** recommended items are based on the past preferences of the user. Each of the above system have some limitations, therefore a hybrid systems is proposed which has empirically demonstrate better effectiveness

Content-based filtering systems are usually criticized for two weaknesses:

i) Content limitation
ii) Over-specialization

6.1 Re-ranking of documents in Personalized Information Retrieval

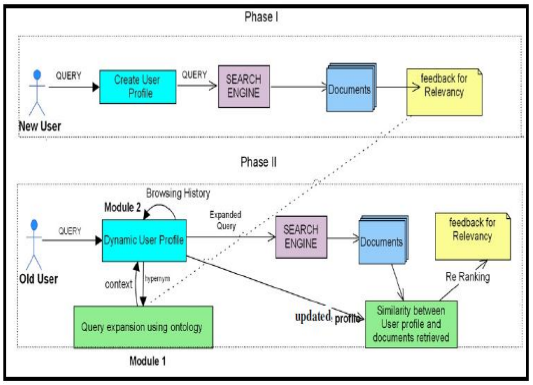
Re-ranking algorithms, query refinement and query suggestion methods, document clustering approaches—these and many other techniques are deployed to provide users of a web search engine better access to the documents relevant for their queries in the context of their information need. Many of these techniques assume that whether or not a document is relevant for a query is determined by its rank in the result list for this query. Naturally, one would expect a document to be the more relevant in the context of a given query the higher it is ranked in the list of retrieved documents. Re-ranking of the results is done using the user profile and profile of others users in the community as selected by the user. Several other works have made use of past queries mined from the query logs to help the current searcher perform collaborative re ranking of results using user and

7. The Proposed Approach

The proposed approach is based on using Dynamic User Profile and Ontology. Overall structure of the system consists of two phases . The first phase includes the standard information retrieval while the second phase uses the relevant documents retrieved in first phase and steps forward following two modules

i. Ontology for Query Expansion
ii. Dynamic user profile

The proposed approach, User profile is built and algorithm finds the context of a user query using relevance feedback and Ontology. In addition, this approach uses a time-based automatic user profile updating with user's changing behaviour.



Personalised Information Retrieval using Dynamic User Profile and Ontology for Query Expansion

Here, the basic terminology and notations used is presented.

A set of *m* finite number of users is termed as *U*. An ith user (*u*i) is indicated as a person who poses the question / query to search engine through web browser. Web User is synonym to user.

New User is a user who poses the query first time using the employed search engine.
New user set *NU* Í *U;*
Old User is the user who has posed the query earlier on the search engine. Hence *OU* Í *U;*
Active User (denoted as *a*) is the user who is currently working; so active user, at time, is either a new user or an old user

ui Î *U {ui: 1 _i _ m}*

and *U = OU* È *NU*

Query Topic (denoted as *QT*, also termed as 'query') is a search query that comprises of one or more keywords/ terms. Length/ size of query are number of terms present in it. *New Query* is a query posed by the user first time. *Old Query* is a query that has already been searched by a user. *Wt(u, j)* is weight given to the jth query topic for the user *u*.
Ontology is an explicit specification of concepts and relationships that can exist between terms. The set of query terms and the relationships among them are reflected in the representational vocabulary with which query expansion is performed.The set of relations such as subsumption IS-A and meronymy PART-OF describe the semantics of the domain. Rather than creating own ontology, existing Ontology WordNet is used. Context is the description of a user's aim / need for information retrieval. In this paper, context is implicitly defined which are updated over time to reflect changes in user interests/ needs. Contexts are extracted from WordNet in terms of concepts.

7.1 Ontology for Query Expansion

The goal is to identify the user context accurately, so search results can be personalized by re-organizing the results returned from a search engine for a given query. In this research, context is extracted from Ontology in terms of concepts. Ontology is used to identify topics that might be of interest to a specific user. For example, the query 'java' will be expanded with "programming language", for the users interested in computer programming language, and with "coffee beans', for the users interested in "tea / coffee", and "island" for the users interested in "islands/ Indonesia".

7.2 Dynamic User Profile

It is thus evident that in order to make search personalized, user profile is essential. As same query, for example "Java" may be asked in different context like Programming Language, Island, and Coffee. However, all the techniques that are used for this purpose have some drawbacks, among which users' privacy violation is the main concern. This research deals with this issue by not exploring the private information such as Social Security No., Name, Age, Gender, Address, Credit Card Nos. and others. This research focuses on how to update the user profile over a period of time using the user's past search history. User profile, representing the user interests is updated with changing context over a period of time. In this method, it is considered that the preferences of each user consist of the following two aspects: Short term preferences and Long term preferences.
In short term preferences, the information used to construct each user profile is gathered only during the current sessions and kept for a limited period of time. As soon as the interest is deviated, it is discarded after executing some adaptive process
aimed at personalizing the current interaction. Conversely, in long term preferences, the user profile is incrementally developed over time and it is stored for use in later sessions. User Profile for user *u* consists of tuples
u(n)<j, Context(u,j), Wt(u,j)>, <k, Context(u,k), Wt(u,k)> ..... (1)

Where for any item j has context (u,j) and computed weight is Wt(u, j) and so on. User Profile P is a vector of weight of all terms of user. In addition, *Wt*(*u*) is the mean term weight for user *u*. When the user poses the old query or similar query
topic, the context retrieved from ontology is added in query topic in order to expand the query. Similar query topic is defined as concept at the same hierarchy level in ontology for example 'Java' and 'C++' are treated as similar as they are at same level just one level down to programming language. Subsequently, the expanded query topic is searched for information retrieval. This approach also updates the user's profile whenever the user's relevant retrieved page changes, in terms of recent context of query terms. At the same time, when user's older interest / context gets deviated over a period of time (threshold defined is 7 days) then user profile is updated with new context.

7.3 User Profile Construction

It is broadly said as profile *P* of any user *u* is
P= (C1* PLong-Term, C2 * PShort-Term) Where *C1* and *C2* are constants denoting the weightage of long term preferences and short term preferences respectively, satisfying
C1 + *C2* = 1. *C1*=0.75 and *C2*=0.25

Is used in these experiments as the long term preferences have higher precedence
proved by . *PShort-Term* is represented as a vector of all short-term preferences

PShort-Term = (PShort-Term-1, PShort-Term-2… PShort-Term-n) ..... (2)
Each element *k*, PShort-Term-k is defined as

$$P_{\text{Short-Term-k}} = \frac{1}{N}\sum_{i=1}^{N}\frac{tf(k, W_i)}{\sum_{j=1}^{T} tf(j, W_i)} * rel(Wi) \quad \text{..... (3)}$$

Where Wi is the ith web page of result set R retrieved from Search engine, tf(k,Wi) is term frequency of term k in ith result page Wi, T is number of terms in the result page, N is total number of pages browsed by the active user. rel (Wi) a binary function on the relevance of a given Wi. It will take 0 or 1 depending on the feedback of user relevant or non-relevant. PLong -Term is represented as vector of all Long-term preferences.
PLong-Term= PLong-Term-1, PLong-Term-2….PLong-Term-m) …….... (4)
Each element k is defined as

$$P_{\text{Long-Term-u}} = \frac{1}{N}\sum_{i=1}^{N}\frac{tf(k, Wi)}{\sum_{j=1}^{T} tf(j, Wi)} * rel(Wi) * e^{-\frac{\log 2}{hl}(d2-d1)} \quad \text{..... (5)}$$

Where Wi is the ith web page of result set R retrieved from Search engine, tf(k,Wi) is term frequency of term k in ith result page Wi, T is number of terms in the result page, N is total number of pages browsed by the $e^{-\frac{\log 2}{hl}(d2-d1)}$ active user. Is a decay factor under the assumption (on the basis of observation throughout experiments) that user's interest is deviated if not browsed within a week. In this factor, d1 is the day when term k last occurs, d2 is the day following to d1 and hl is half-life parameter, set to 14. If (d2-d1) > 7 days then hl is set to 7. rel (Wi) a binary function on the relevance of a given Wi. It will take 0 or 1 depending on the
Feedback of user relevant or non-relevant.

7.4 Data Set used in proposed methods

Two datasets are used for evaluation of the proposed method.

i. Generated Data Set
ii. FIRE 2010 Data Set

The first dataset is manually generated based on the Web that Google has indexed. It is generated by web interactions of 20 users, who used the Google search engine for 30 days, an average of three query topics per day from a collection of 60 query topics. The query topics have an average query length 2.2. The queries used in our experiments were intentionally designed to be short after removing stop words to reflect the general trends in user search queries. The set of predefined query topics is collected from various users with similar as well as non-similar backgrounds. Although query topics were created manually however users were carefully inquired from different background and having different context. In these experiments, users were asked to provide the relevance feedback without much interfering them. All the relevant documents were processed and user profiles were created. The second dataset used for evaluation of the proposed approach is FIRE 2010 dataset. In FIRE 2010 data set consists of a collection of 50 Query topics with description and narration. In this evaluation process, 20 users were asked to interact with

Terrier search engine by undertaking phase 1. Since second data set has predefined context of query topics, so it is considered that all users had same context with each query topic. Some users posed few overlap query topics also and provided relevance feedback. These data sets are used throughout this thesis for all approaches.

7.5 Evaluation of Proposed System

A number of studies have been conducted to measure the performance of the system. Some criteria of evaluation have been proposed by several researchers in the area of the evaluation of information retrieval systems. These criteria include: coverage of the system, form of presentation of the search output, user effort, the response time of the system and recall & precision. Main objective of this research is to achieve personalisation of retrieval activities. Personalised Retrieval effectiveness is defined in terms of retrieving relevant documents and not retrieving non-relevant documents. Two traditional factors of measuring effectiveness are Recall and Precision.

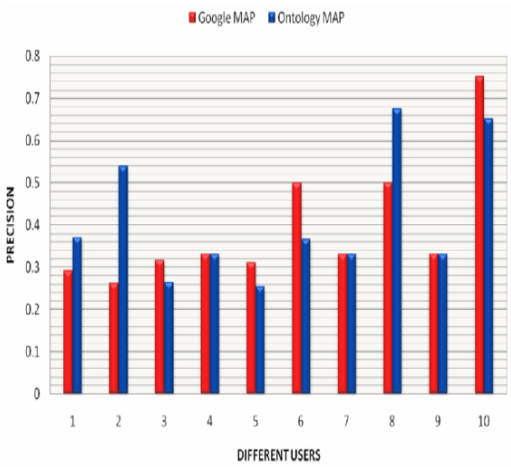
8. Results and Analysis

The results obtained from search engine are compared with results obtained from the proposed approach. The average precision and average recall measures are used to evaluate the retrieval accuracy performance of the proposed method. The definitions of these measures assume that, for a given query, there is a set of documents that is relevant and a set of documents that is not relevant.
MAP, MAR and Mean Average F-Score for each user in the data set are computed as different precision and recall values are obtained for the same query posted by different users in different context. The data sets were used for the performance evaluation of the proposed approach.

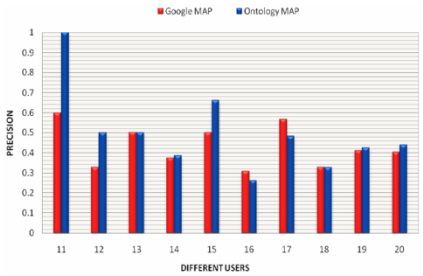
8.1 Generated Data Set Results

The Generated Data Set Results discussed as: MAP results and MAR results.

MAP Results



MAP for Google and Ontology for first 10 users



MAP for Google and Ontology for next 10 users

The users are classified on the basis of their performance. The MAP performance comparison between the Google and the Ontology is calculated as

$$\text{Performance} = \frac{\sum_{i=1}^{20} (OMAP_i - GMAP_i)}{\sum_{i=1}^{20} GMAP_i} \text{ x } 100$$

Where GMAP = Mean Average Precision for Google results; OMAP = Mean Average Precision for Ontology results

Comparison	No. of Users	Performance
OMAP > GMAP	9	35.96%
OMAP = GMAP	5	-
OMAP < GMAP	6	-17.21%

MAP comparison between Google and User Profile +Ontology

9. Summary and Conclusion

The design and implementation of the proposed approach using Dynamic User profile and Ontology. The experiments designed are first discussed, followed by the experiment framework and environment. The overview of the proposed system. In addition, this chapter gives details of the algorithms devised and implemented for query expansion using ontology and re-ranking of documents with using user profile. Evaluation of Context aware applications is quite difficult as they depend on context. The contexts or situations of interest depends on user to user and can't be generalized. Results show the precision of this approach (Ontology + UP) is better by 10.35%
over Google on an average. Similarly, Recall of this approach (ontology + UP) is better by 4.19% over the Google on an average. The results obtained from generated data set were comparable with FIRE data set results.

10. References

[1] Bhowmick P, Sarkar S and Basu A (2010). "Ontology Based User Modeling
for Personalized Information Access". International Journal of Computer
Science and Applications Technomathematics Research Foundation, 7(1), 2010, 1-22.

[2]Gauch S, Speretta M, Chandramouli A and Micarelli A (2007). "User
profiles for personalized information access". Lecture Notes in Computer
Science. 2007, 4321, 54-60.

[3] Gemechu F, Zhang Y and Liu T (2010). "A Framework for Personalized
Information Retrieval Model". Second International Conference on
Computer and Network Technology, 2010, 500 – 505.

[4] Tao X and Li Y (2009). "Concept-based, Personalized Web Information
Gathering: A Survey". In proceedings of the 3rd International Conference on
Knowledge Science, Engineering, and Management, 2009, 215-228.

[5] Tao X and Li Y (2009). "A User Profiles Acquiring Approach Using
Pseudo-Relevance Feedback". In Proceedings of the Fourth International
Conference on Rough Set and Knowledge Technology, 2009, 658-665.

[6] Stefka Toleva–Stoimenova(2010). "Evaluation of Web Based Information
Systems". Users' Informing Criteria. State University of Library Science and
Information Technologies, 7, 2010.

[7] Sugiyama K and Hatano K (2004). "Adaptive web search based on user
profile constructed without any effort from users". 13th international
conference on World Wide Web, 675 – 684.

[8] Tan A H (1999). "Text mining: The state of the art and the challenges". The
Pacific Asia Conference on Knowledge Discovery from Advanced Databases
(KDAD'99).