International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 6, August - 2012

Personalised Information Access Based on Ontology and Collaborative Filtering

Aditi Sharma

Jaipur National University,Jaipur

Abstract

It is now very easy to access information from internet via World Wide Web. If we access anything via search engine they do not deliver the relevant information because they are programmed as one size fits all. Therefore it is very logical and important to personalize the retrieval system according to preference of user. a retrieval system based on user interests andpreferences play an important role to enhance effectiveness of information retrieval. These systems can also distinct short term and long term preferences of user on the basis of frequency of user interest. The aim of this thesis is to refine access of information in the web information retrieval towards personalization by using dynamic user profile,ontology based query expansion and collaborative filtering technique. This thesis work contribute in improving the accuracy of information retrieval to personalize the user profile by combining the dynamic user profile with ontology and dynamic user profile with collaborative filtering .

1. Introduction

World Wide Web (W.W.W) is a magnificent and vast portal for acquiring information from the infinite information present on the web. It is an enormous, assorted, and dynamic information resource to cope up with the abundance of available information. The explosive growth of documents in the Web makes it difficult to determine the most relevant documents for a particular user, given a general query. User performs searching to access relevant document for himself. Therefore, a specific system should be there to personalize the information retrieved by user according to his preference.

Motivation

Information retrieval technologies have been growing and search engines do a significant job of indexing content available on the web and making it available to its users. Search engines often return a bulk of information, more than the user could possibly use, while overlooking the background of a user during the search.If the same query is posed by a school student or a Software Engineer, generally the returned results remain the same. Let us assume that the school student focuses the search about the fruit “Apple”. In order to do this, the user introduces the query “Apple.” Expecting to find documents about that fruit, the user actually finds out that the first results obtained with that query correspond to documents that do not contain that concept at all. Instead, all sorts of web pages related to the well- known Software Company with the same name “Apple” and its products are displayed to the user. The first result concerning the “Fruit” is far away from the top of the list.This underlines the need to provide users with information personalized to their needs. Personalised search has got significant attention addressing the challenges in the web search community. Hence in recent years, the huge efforts have been done in developing techniques for effective and efficient information retrieval to reduce ambiguities. Search engines occupy a significant place in information retrieval technologies by doing a job of indexing contents available on the web and making it available to its users. Search engines retrieve a bulk of information, beyond the capacity of user to go through it. Despite that a search engine may satisfy the search criteria, it often fails to meet the user’s search intention. To provide more effective information, the search process must incorporate user profile rather thanconsidering only the user query.

Research Problems

The search engines are concerned about certain things of information retrieval proved but still there is a scope of improvement which makes this research very significant. On an important note we have considered following problems in this work as prime factors:

1. keyword-matching approach, adopted by search engines is not able to circumvent information overload and information mismatch:

Web searching is based on traditional information retrieval techniques and is typically based on Boolean operations using keywords.Most search engines use only user queries for performing their searching tasks. A user query refers to a list of keywords (with some additional operators). A user poses a query to the web and the query response is the location of the set of documents ranked depending on their similarity to the query. Search engines retrieve documents based on keyword-matching method which returns too many results to the user. Most of the retrieved documents are either irrelevant or contain information hidden in a mess of other data. A user must scroll down a long list of documents and find only a few of them relevant. It is often very time consuming to conduct a useful web search. Most web users do not have the time to examine hundreds of thousands of retrieved pages provided by search engines.

2. No attention to the user's profile In traditional information retrieval system causes the unavailability to provide rich of semantics to facilitate IR processes by matching users' expectations.

As most of the traditional web searching pays no attention to the user's profile, they are not able to consider the complexity of user's intention completely. User profile can be an important source for Information Retrieval (IR) processes. In order to make web searching more precise and to provide more valuable and relevant information access, the search process must incorporate user profiles rather than considering only the user queries. It is very difficult to acquire user profiles automatically because the user profiles are very dynamic and flexible. In traditional IR, user profiles are often represented by keyword / concepts space vectors or by some predefined categories.

Ontology

Ontology is formal description of knowledge. It is a set of vocabulary and thesemantic 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 interrelationships 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 andmetonymyPART-OF describes 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.

Key Components of Ontology

Ontology consists of a finite list of terms and the relationships between them. Theterms denote important concepts (classes of objects) of the domain and the relationships include hierarchies of classes. In general, Ontology is organized in taxonomies and contains modelling primitives such as classes, relations, functions, axioms and instances.

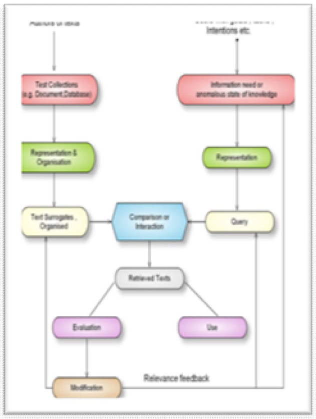
Purpose and Benefits of Ontology

Fundamentally, ontology is used to improve communication between either humansor 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 modelling methods, paradigms, languages and software tools to achieve inter-operability among computer systems.Ontology can deliver many benefits for Systems Engineering such as it may serve asan index into a repository of information to facilitate information search and retrieval. This thesis focuses on this benefit of ontology.

Information Storage and Retrieval Systems

An IR system must support certain basic operations. There must be a way to enter documents into a database, change the documents, and delete them. There must also be some way to search for documents, and present them to a user. Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information. “Information retrieval is concerned with the representation, storage, and organization and accessing of information items.” An IR system matches user queries--formal statements of information needs--to documents stored in a database. A document is a data object, usually textual, though it may also contain other types of data such as photographs, graphs, and so on. Often, the documents themselves are not stored directly in the IR system, but are represented in the system by document surrogates.

1. **Information Retrieval Model**
2. **Vector Space Model**
3. **Information Retrieval Process**
4. **Information Retrieval Techniques**


Intentions etc.
Test Collections (e.g. Document Database)
Information need or anomalous state of knowledge
Representation & Organisation
Representation
Text Surrogates , Organised
Comparison or Interaction
Query
Retrieved Texts
Evaluation
Use
Modification
Relevance feedback

Working model of Information Retrieval

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 contextinto a single framework in order to provide the most appropriate answer for a user’sinformation 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.

Ontology in Personalized Information Retrieval

Ontology has been a basis for the construction of a user model in severalpersonalized systems ranging from information delivery systems to Intelligent TutoringSystems The retrieval models are based on keyword or term matching, i.e., matching terms inthe 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.

WordNet as Ontology

Relation	Description
Synonymy	Symmetric and interchangeable relation between terms
Antonymy	Symmetric relation between terms with opposite

	meaning
Hyponymy / Hypernymy	Transitive relations of A kind of for nouns
Meronymy / Holonymy	Relation of A part of for noun
Troponymy	Transitive relations of A kind of for verbs
Entailment	Relation between two verbs while one logically entails the other

Description of Ontology

An example of WordNet ontology

Jazz is a concept having three subconcepts: Talk; Popular Music and Dance Music. Talk having hierarchal subconcepts like conversation, speech, communication and so on. Similarly, PopularMusic has sub concept music genre.

User Profiling for Personalized Information Retrieval

Researchers, investigating personalization techniques for Web InformationRetrieval, 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. Hochulproposed 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.

Recommender Systems

Recommender systems are software applications that provide personalized advice tousers 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 threeproblems:

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

Re-ranking of documents in Personalized Information Retrieval

Re-ranking algorithms, query refinement and query suggestion methods, documentclustering 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 pastqueries mined from the query logs to help the current searcher perform collaborative reranking of results using user and community profiles built from thedocuments marked as relevant by the user or community respectively. The search process and the ranking of relevant documents are accomplished within the context of a particular user or community point of view.

Research Issues in Personalized Information Retrieval

Issues	Problem	Possible Solution
Cold Start & Latency	Latency (or New Item) problem: lack of rating data for a new item	Use a hybrid recommendation approach that considers both content based and collaborative information during the recommendation process
	Cold Start problem: no rating data for any users of the system at the time of initialization of the system.	i)Use the external ontology in combination with content based and feedbacks from academics asthe starting knowledge for arecommendation system. ii) Initialize the profile of new user by reusing the properties of similar peer profiles in the semanticneighborhood.
Domain Knowledge	Ignoring domain knowledge (semantics)	i) Using the semantic knowledge about items to enhance the recommendation ii) Use of ontological profiles for user using knowledge base semantics
Adapting to User Context & Managing the Dynamics In User Interests	**Adapting to User Context:** Nonconsideration of context as a concept into personalized recommendations.	i)Use of current behavior to discover the user context via nodes in a concept network induced from the original ontology and is updated incrementally based on the user's interactions with the concepts of the ontology. Then use of this context to predict the current behavior of the user. ii) Context may be obtained by the system based on the user's longterm and short term preferences
	Managing the Dynamics in User Interests: User dynamic nature handling through the static user profile	i)Use of a continuous weighting function that associates a higher weight to more recent interactions with a user. ii)Evolution of a population of profiles per user, as interests of users change over the time, profiles that better reflect their current interests become more prominent within the population

Personalized Information Retrieval System Issues

this thesis attempts to overcome the following three issues:

1. Cold start and Latency- This issue relates to the lack of rating data for a new item and users of the system at the time of initialization.
2. Domain Knowledge.
3. Adapting to user Context & Managing the Dynamics in User Interests.

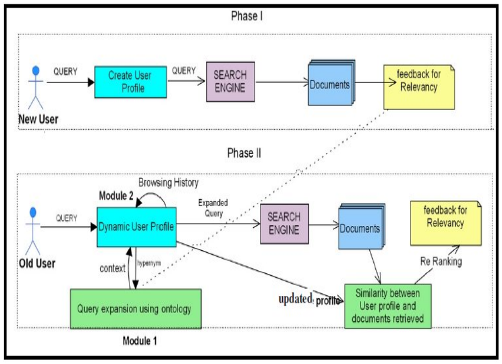
Design and implementation of Personalised Information Retrieval using User Profile and Ontology

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:

1. Ontology for Query Expansion

2. 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 behavior.



Personalised Information Retrieval using Dynamic User Profile andOntology 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*updatedQuery Topic (denoted as *QT*, also termed as 'query') is a search query thatcomprises of one or more keywords/ terms. Length/ size of query are number ofterms present in it. *New Query* is a query posed by the user first time. *Old Query* is aquery that has already been searched by a user. *Wt(u, j)* is weight given to the jthquery topic for the user *u*.

Our goal is to identify the accurate user context to personalize search results by reorganizingthe results returned from a search engine for a given query. Initially, when learning user interests, system doesn't perform good until enough information have been collected for user profiling. Using ontology, the basis of the profile allows the initial user behaviour to be matched with existing concepts in the ontology and relationships between these concepts. In our approach, the purpose of using ontology is to identify topics that might be of interest to a specific user. For example, the query 'jazz' will be expanded with "music", "dance" for the users interested in music or dance, and with "talk', for the users interested in "talk"," conversation" "speech".

Data Set used in proposed methods:

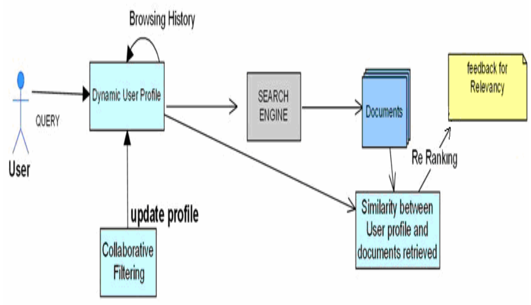
Two datasets are used for evaluation of the proposed method.

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 nonsimilarbackgrounds. 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 I and phaseII of our system. 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.

Design and Implementation of Personalised InformationRetrieval using User Profile and Collaborative Filtering

Collaborative Filtering is an approach which considers not only the profile of theactive user but also considers the neighbourhood of the active user with similar preferences while recommending the items. Collaborative filtering means that people collaborate to help one another in filtering the documents they access, by using their reactions.It is observed that collaborative filtering system algorithms are required to consider the following points to provide useful recommendations
1) similarity between users for cluster formation
2) selecting a sub-set of the neighbourhood
3) prediction for rating of items
In collaborating filtering, cluster of users having similar interests needs to be created. As the number of users in the cluster increases, the performance of the collaborative filtering reduces due to the noise generated by the users' recommendations. Thus, collaborative filtering algorithm chooses the appropriate neighbourhood from the cluster to provide recommendations. In our proposed approach, collaborative filtering recommendations are combined with the dynamic user profile to study the impact of this combination on the performance



Dynamic User Profile

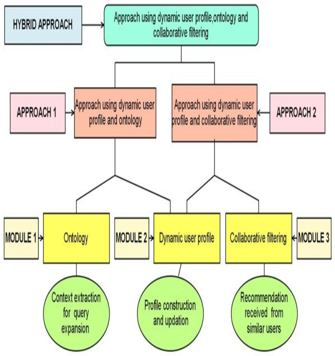
Here the existence of a set of n users is assumed, $U = \{u1, u2... un\}$ and item $i = \{i1, i2... in\}$. User Profile for user u consists of tuples
$u(n) = \{<i1, Wt(u,i1)>, <i2, Wt(u,i2)> ... <in, Wt(u,in)>\}$ .... (1)
where for any item im, the computed weight is Wt(u, im). User Profile P is a vector of
weight of all terms of user.

Collaborative Filtering

Collaborative filtering (CF) is the process of filtering for information or patternsusing techniques involving collaboration among multiple agents, viewpoints, data sources etc. The algorithm given in is used to make predictions based on collaborative filtering concept. In this algorithm, query is submitted by the user. Query topic is pre-processed by removing Stop Words (semantically non-relevant terms) followed by stemming. The processed query goes through a search engine and documents are retrieved.

Design and Implementation of Personalised Information Retrieval using Hybrid Approach

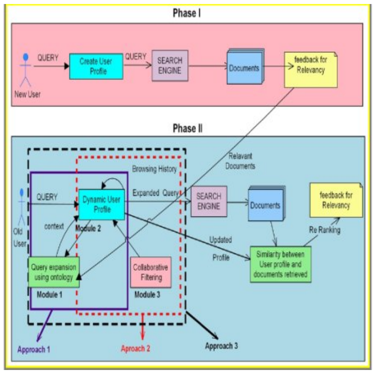
Hybrid system in order to addressthe problems faced in above approaches. The hybrid system finds the context of a user query with least user involvement by using Ontology. In addition, this approach uses a time-based automatic user profile, updating with user's changing behaviour as well as this approach uses recommendations from similar users using a collaborative filtering approach. the hybrid approach which has been developed in this thesis. The chapter contains an overall presentation of the proposed hybrid approach, giving details of each of the steps in the subsequent sections followed by extensive evaluation of the proposed approach using standard IR methods. the concept of proposed hybrid approach.


HYBRID APPROACH
Approach using dynamic user profile,ontology and collaborative filtering
APPROACH 1
Approach using dynamic user profile and ontology
Approach using dynamic user profile and collaborative filtering
APPROACH 2
MODULE 1
Ontology
MODULE 2
Dynamic user profile
Collaborative filtering
MODULE 3
Context extraction for query expansion
Profile construction and updation
Recommendation received from similar users

Concept of Hybrid Approach

The Proposed Hybrid Approach

The overview of the two phase hybrid. The phase Iincludes the standard Information Retrieval while the phase II uses the relevantdocuments retrieved in first phase and steps forward using three modules namely(i) Dynamic user profile (ii) Ontology for query expansion and (iii) Collaborative filtering.In first approach, we combined the module (i) and module (ii). Similarly in second approach, we combined module (i) and module (iii). In third - Hybrid approach, we combined module (i), module (ii) and module (iii).



Integrating Together

the working of the system designed. This figure has two arms –leftand right hand side. In right hand side, proposed approach is used. It is also named as phase II, which has three modules namely: (i) Ontology for query expansion. (ii) Dynamic user profile and (iii) Collaborative filtering. In the left hand side, standard search engine retrieves the information. This known as phase-I.

Conclusion

There are many well known reasons behind this. Firstly, the search engines aremostly based on keyword-matching. Consequently, users suffer from the problems of *Information mismatch* and *Information overload.* Secondly, web users provide only short phrases in queries to express their required needs. In addition, web users formulate their queries differently because of their personal vocabulary, perspective and knowledge. If user information needs can be better captured and interpreted, more useful and meaningful information can be delivered to them. Web users have a perception in their mind by which they can easily decide whether a document is useful to them or not while reading through the documents although they may not be able to justifythe reason. This perception can be obtained from their background knowledge and used to find the information implicitly. Hence more meaningful and personalized information may be retrieved for users

REFERENCES

[1] Aas K (1997). “A Survey on Personalized Information Filtering Systems for
the World Wide Web” December 1997.
[2] Adomavicius G and Tuzhilin A (2005). “Personalization Technologies: A
Process-Oriented Perspective” Communications of the ACM, 48(10),
October 2005.
[3] Allan J and Aslam J (2003). “Challenges in Information Retrieval and
languagemodeling”. ACM SIGIR Forums, 37(1), 2003, 31-47.
[4] Amato G and Straccia U (1999). “User Profile Modeling and Applications to
Digital Libraries”, ECDL '99 Proceedings of the Third European Conference
on Research and Advanced Technology for Digital Libraries, 184-197.
[5] Anand S S and Mobasher B (2005). “Intelligent techniques for web
personalization.” Lecture Notes in Computer Science 3169, 1-36.