PERSONALIZED INFORMATION RETRIEVAL USING AGENT BASED APPROACH

Gayathri J #1, Dr. S. Kumar Chandar #2

#1 Assistant Professor, Department of Information Technology, Sri Sai Ram Institute of Technology, Chennai, India
#2 Associate Professor, Department of Management Studies, CHRIST University, Bangalore, Karnataka, India

ghayatirijavaraman@gmail.com
kcresearch@gmail.com

ABSTRACT

The information becomes available on the World Wide Web becomes more difficult to provide effective search tools for retrieving relevant information according to the need of the user. Most of the current information retrieval approaches follow the typical query mechanism in the form of keywords. Locating relevant information on the web is a very complex task because the keyword-matching method usually matches on individual words instead of their meanings. The policy of taking into account context may lead to search for information based on their meaning rather than on the presence of the keywords in the object. Generally, each user has different information needs for his/her query. Therefore, the search results should be adapted to users with different information needs. The proposed approach identifies knowledge fragments consisting of not just entities but also the relations between them which lead to specific information extraction. An agent, in the process of searching relevant information interpret the user request, analyze it, access multiple heterogeneous distributed information sources, extract the desired information and gives them back to the user.

KEYWORDS

Multi Agents, Intelligent Agents, Ontology, Information Extraction, Classification, Context Oriented search, Personalization

I. INTRODUCTION

An information retrieval (IR) system is concerned with retrieving only those documents relevant to the information need of any user. In principle, an IR system should be context aware. The problem with most of the traditional retrieval systems such as search engines is that they are unaware of contextual features. People have difficulty with the typical query systems of search engines. They accept the user query as a set of keywords in the form of a Boolean expression. Unfortunately the keyword-matching method usually returns too many low quality results because the keyword-matching method usually matches on individual words instead of their meanings. Furthermore, the formulation of some complex queries in the form of a Boolean expression is difficult. Limitations of General-purpose search engines are Querying Mechanism, Keyword Exact Matching, Low web Coverage Rate, Limited customization to individual users, Long result list with low relevancy to user query. Therefore better support is needed for expressing one's information need and returning high quality search results. There appears to be a need for systems that do reasoning under uncertainty and are flexible enough to recover from the contradictions, inconsistencies, and irregularities that such reasoning involves.

Taking context into account may lead to search for information based on their meaning rather than on the presence of the keywords in the object. It is the case for information extraction of the majority of from pages about the same topic. The basis of this approach lie in the use of ontologies that make possible to define restricted field and context and intelligent agents to perform information extraction. The following section briefly describes the use of agents and ontologies in the approach to develop intelligent information gathering systems.

Section II describes the use of agent and ontology based approach in information extraction. An overview of hyperlink-based personalized Web search, personalized Web sites, and recommender systems is given in Section III. Section IV describes the architecture with several principal stages in interaction to propose an IR system which retrieves only relevant documents. Section V describes standard measures of precision and recall. Section VI shows the first results of experimentation. Section VII concludes the paper.

II. AGENT AND ONTOLOGY BASED INFORMATION EXTRACTION
The vastness of the Internet and bandwidth limitations makes it difficult for a system to perform direct queries on it efficiently. If the relevant data is already stored in a knowledge base, then it is possible to respond to queries very quickly.

A. Agents

Intelligent agents understand what we mean, find and use relevant information, automate routine tasks, adapt to our needs and desires, cooperate with other assistants. Intelligent Information Agents can be defined as computational software systems that access multiple heterogeneous and geographically distributed information sources in order to help users in the process of searching relevant information. These agents provide transparent access to many different information sources, simplifying the problem of managing a large amount of data. This implies that agents must interpret the user request, analyze it, translate it in the correct way for each source, extract the desired information, and gives the results back to the user.

1) Multi-Agent System:

The capacity of an intelligent agent is limited by its knowledge, its computing resources, and its perspective. By forming communities of agents a solution based on a modular design can be implemented where each member of the agency specializes in solving a particular aspect of the problem. Agents in multi-agent systems (MAS) are autonomous and can engage in interactions which are essential for any ongoing agents’ actions. A MAS approach is thus regarded as an intuitive and suitable way of modeling dynamic systems [3].

The characteristics of MASs are defined as follows:

- Each agent has incomplete information or capabilities for solving the problem and, thus, has a limited viewpoint
- There is no global control system
- Data are decentralized
- Computation is asynchronous

MAS can be used to solve problems that are too large for a centralized agent to solve because of resource limitations and to avoid a one point bottleneck or failure point.

B. Ontology

Ontology in a particular domain is a description of the concepts and relationships that can exist in the domain [5]. One of the primary purposes of constructing an ontology is to provide a standard, unambiguous representation of a particular domain of knowledge. Ontologies have been built and used successfully in constructing multi-contextual knowledge bases. The use of ontologies brings many advantages

1. Retrieving the appropriate information from documents by providing a structure to annotate the contents of a document with semantic information.
2. Integrating the information from various sources by providing a structure for its organization and facilitating the exchange of data, knowledge and models.
3. Supporting inference to derive additional knowledge.

In information gathering on the Web, the combination of the multi-agents approach with declarative knowledge, leading to the ontologies use, is relevant for the intelligent information gathering systems development. Each information agent is specialized to a single application domain and provides access to the available information sources within that domain. The domain model of an agent defines its area of expertise and the terminology for communicating with it. It provides an ontology to describe the application domain. This ontology consists of descriptions of the classes of objects in the domain, relationships between these classes and other domain-specific information.

III. PERSONALIZED SEARCH SYSTEMS

There are several types of search systems that provide users with information more relevant to their individual needs. For example, hyperlink-based personalized Web search, personalized Web sites, and recommender systems are reviewed [7].

In personalized Web searches, the hyperlink structures of the Web are also becoming important. The use of personalized Page Rank to enable personalized Web searches was first proposed in, where it was suggested as a modification of the global Page Rank algorithm, which computes a universal notion of importance of a Web page.

A. Link Personalization

This scheme involves selecting the links that are more relevant to the user and changing the original navigation space by reducing or improving the relationships between Web pages. E-commerce applications use link personalization to recommend items based on the buying history of clients or some categorization of clients based on ratings and opinions.

B. Content Personalization
In general, content personalization is done when pages present different information to different users. My Netscape4 [7] filters the information that is relevant to the user, showing only sections and details in which the user may be interested. The user may explicitly indicate his/her preferences, or preferences may be inferred (semi) automatically from his/her profile or from his/her navigation activity. At these sites, users choose a set of “modules” from a large set including weather, news, music and so on, and further personalize these modules by choosing a set of attributes of the module to be perceived.

C. Recommender Systems

Recommender systems [7] collect user feedback in the form of ratings for items in a given domain and exploit similarities and differences among profiles of several users in determining how to recommend an item. There are two prevalent approaches to constructing recommender systems – collaborative filtering-based and content-based recommendation.

IV. PROPOSED ARCHITECTURE

The main objective of this proposed method is to retrieve documents relevant to an information need. One approach to retrieve information in context is by using user-defined evidence from document clusters, query histories, and profiles. The development of highly autonomous and adaptive Information retrieval (IR) systems which can automatically select relevant information items on behalf of the users. This approach is proposed to achieve result list with high relevance to user query.

Information Retrieval often refers to the situation that an information seeker takes an active role to specify his queries. Information Filtering is concerned with the removal of irrelevant information from an incoming stream of information, based on the information seeker’s long-term and recurring retrieval goals stored in a user profile. Contextual knowledge refers to the semantic relationships among concepts. Personalization would involve the following steps:

1. Collecting and representing information about the user, to understand the user’s interests.
   Using this information to filter the results returned from the initial retrieval process, or directly including this information into the search process itself to select personalized results. The proposed method introduces several principal stages in interaction to propose an adaptive IR system which retrieves only relevant documents.

A. Document Characterization

IR systems take as input a set of documents and a query. In terms of document representation, a document is preprocessed according to traditional IR techniques to extract a set of tokens (e.g., stems, n-grams, or phrases) as its characterization.

Preprocessing of Documents consist of the following operations: (i) All Stop words (a, an, the, , ..... etc.) are removed from text documents (ii) Then the resulting message after removal of stop words is tokenized into a set of strings (iii) The resulting set of tokens are stemmed to their roots. The algorithm works through a string in a series of steps as follows:

Step 1: It deals with removing plurals and past Participles.
Step 2: Rule for removing the suffix will be given in the form (condition)S1--→S2
Step 3: Porter’s does pattern matches on some common suffixes. Removes suffixes and replaces them with their roots

![Figure 1. Preprocessing](image)

The outline of the preprocessing phase is given in Fig.1. After the preprocessing phase, word-order information is lost and the remaining words are called terms. Then, a term x document matrix is created, where the documents are the rows of the matrix and the terms are the columns of the matrix. A set of d documents and a set of t terms can be modeled with each document as a vector v in the t dimensional space. The jth coordinate of v is a number that measures the association of the jth term with respect to the given document: It is simply defined as,

\[ v_j = \begin{cases} 1, & \text{if document contains the term} \\ 0, & \text{otherwise} \end{cases} \]
Let \( v_j \) be the term frequency, that is, the number of occurrences of term \( t_i \) in the document or the relative term frequency, that is, the term frequency versus the total number of occurrences of all the terms in the document. Three types of tokens can be extracted: Positive tokens represent what items the information seeker would like to retrieve; negative tokens indicate what the information seeker does not want; neutral tokens are not good indicators of information needs.

**B. User profile Creation**

A retrieval context may consist of an information seeker’s background, long-term search goals, tasks at hand, knowledge about a retrieval domain, etc. User may be interested in more focused information and his/her search goal is to find exact information related to the key word queried. On the other hand, a user may wish to find more general information. Understanding the goals and behavior of information seekers would be useful for building more accurate Web user profiles.

User profile is constructed from the topics of a user’s interest i.e., search intent. The topic in a particular document comprises the terms which represent the subjects [13]. By using the ontological approach; the user profile also includes the topic’s semantic relationship.

![System architecture](image)

Figure 2. System architecture

Hence, this type of user profile is called topic ontology. The topic ontology is constructed from primitive objects (e.g., terms). They consist of primitive classes and compound classes.

IR system should be able to automatically induce a user’s interest based on the user’s interactions with the system. User information is collected explicitly by requesting users enter data by filling in forms and answering questions about them. In general, the adaptive IR system will maintain a separate user profile for each individual information seeker who is interested in a specific topic. Furthermore, these profiles can be aggregated to form a generic profile for a group of information seekers who share similar interests.

Initial contextual knowledge is acquired before the query is received. The initial belief about the information seeker’s background is obtained via an interface which acquires various information from the user. The revised user profile will be used to match against incoming documents. Figure 2 depicts the overall system architecture.

A user profile is an ontology fragment that describes the user’s characteristics in regards to information consumption. Some of the characteristics are relatively static, such as a user’s interests, skills, preferences but many others are highly dynamic like available devices, network bandwidth, current locations, etc., and change significantly depending on the situation.

**C. Adaptive Learning Method**

As user’s information needs as well as the underlying retrieval context may change over time, the learning mechanism of the adaptive IR system will use the information to revise and refine the initial user profile. For example, if the user has requested some documents recently, the retrieved documents will form the basis of supervised learning for the user’s preferences. Supervised learning – also called classification – is based on learning from a training data set.

In Naïve Bayes classifier model of learning it is assumed that text documents are generated from a parametric model [9]. For each new document, using the estimated parameters and Bayes rule, the classifier calculates the probability of generation of document by each class. The classification is selecting the class with highest probability. The classifier parameterizes each class with word frequency and document frequency. Each word occurs in a document independent of all of other words in document and also independent of its occurrence location.

1) **Extraction and Indexing of Information:**

Web pages are obtained by querying external search engines on the Web. Classify the obtained web pages
in such a way that each one is specialized in a specific field.

A concept $c_i$ is a vector representation,

$$C_i = W_{cip1}, W_{cip2}, \ldots, W_{cipn}$$

where $p_1, p_2, \ldots, p_n$ correspond to words in the vocabulary and are called dimensions, $w_{cip}$ denotes the weight of $p_i$ in the vector of $c_i$.

With informational queries, users want to find information on a certain topic. Such queries usually lead to a set of results rather than just one suitable document. Informational queries are similar to queries sent to traditional text-based information retrieval systems. Navigational queries are used to find a certain Web page the user already knows about or at least assumes that such a Webpage exists. Transactional queries can be for the download of a program or file, the purchase of a product, or a further search in a database.

Information objects from specific information sources such as the Web are characterized by a particular indexing scheme and stored in the local cache of the adaptive IR system. The latent semantic indexing method [9] uses singular value decomposition, a well-known technique to reduce the size of the term frequency matrix.

Given a TxD, term frequency matrix representing T terms and D documents, the SVD method removes rows and columns to reduce the matrix to size $K \times K$, where $K$ is usually taken to be around a few hundred for large document collections. To minimize the amount of information loss, only the least significant parts of the frequency matrix are omitted [8].

The latent semantic method consists of the following basic steps:

1. Create a term frequency matrix.
2. Compute singular valued decomposition of frequency matrix by splitting the matrix into three smaller matrices.
3. For each document $d$, replace its original document vector by a new one that excludes the terms eliminated during SVD.
4. Store the set of all vectors, and create indices for them.

D. Matching Mechanism

The matching mechanism of the adaptive IR system tries to match the user’s information needs with incoming information objects by comparing the corresponding queries and document characterizations. Information Filtering systems determine which documents in the results sets are relevant and which are not. This is usually done by comparing the documents to a list of keywords that describe a user or a set of documents that the user previously judged relevant or irrelevant, respectively. Good systems filter many non-relevant documents and keep the relevant ones in the results set.

IR systems using the vector model employ similarity measures to find relevant documents to the query. A common similarity measure is the cosine measure, where the relevance of a document to the user query is measured by cosine of the angle between the document vector and the query vector. The smaller the cosine, the more relevant the document is considered to the user query.

A retrieval situation refers to a user query and associated retrieval context [2]. The matching measure is noted as,

$$\text{Match}(U, \text{doc}) = \sum \left[ \deg(B, l) - \deg(B, -l) \right]$$

$|S|$ where, $U$ refers to the stored user profile and doc is the retrieved document. The degree function is not a simple model, since the function $\deg(B, l)$ infers about the relevance of a document characterization $d$ with respect to the knowledge base $\text{content}(B)$ which represents the stored user information. The basic idea is that a document $\text{doc}$ is characterized by a set of positive literals $d = \{l_1, l_2, \ldots, l_n\}$. The set $S$ of active information carriers partially characterizing the document $\text{doc}$ is defined by,

$$S = \{l_i \cdot d: \deg(B, l_i) > 0 \lor \deg(B, -l_i) > 0\}.$$

The proposed similarity measure, not only considers the syntactic aspects in information matching, but also takes into account the semantics among information items. Information objects which are found relevant after the matching mechanism by the IR system are dispatched to the user in the form of a retrieval result set.

V. MEASURING RETRIEVAL EFFECTIVENESS

The evaluation of the IR systems is usually done with the standard measures of precision and recall, where:

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$
In the context of information retrieval, precision is defined as the ratio of relevant documents to the number of retrieved documents and recall is defined as the proportion of relevant documents that are retrieved [5]. Recall in the form of finding the most information rich pages, called authority pages and hub pages, i.e., pages that have links to many authority pages are also recognized as being very valuable.

VI. IMPLEMENTATION AND RESULTS
User profile creation and document characterizations modules are implemented. The implementation of matching mechanism stage is under progress.

VII. CONCLUSION
The existing information retrieval system is lack of user modeling and is not adaptive to individual users. The above sections describe the proposed architecture which retrieves only relevant documents adaptive to users. The proposed method is composed of several stages, the preprocessing step which gives document characterization as output. User profile creation is done explicitly by requesting users entering data about them by filling in forms. Information about user’s current context and long and short-term interests are collected from the user profile form. Information objects from specific information sources such as the Web are characterized by a particular indexing scheme and stored in the local cache. The matching mechanism filters the relevant information by comparing the documents to a list of keywords that describe a user. With this architecture, the system could be updated easily in the future by adding new features or improving existing ones.

VIII. REFERENCES

1995.
[9] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers.

AUTHORS PROFILE

J.Ghayathri received the Masters degree in Computer Science and Engineering from Anna University, Coimbatore. She is working as an Assistant Professor in the Department of Information Technology- Sri Sai Ram Institute of Technology, Chennai. She is having a teaching experience of 8 years. Her interests include Software Engineering, Testing, and Knowledge Management. She is an active member of professional bodies like ISTE, IAENG. She can be reached at ghayathirijayaraman@gmail.com

Dr S. Kumar Chandar is working as an Associate Professor in Department of Management Studies – Christ University, Bangalore having 14 years in wide variety of technical, managerial and teaching roles. He received PhD in Management in 2013 from Chhatrapati Shahu Ji Maharaj University (Kanpur University). He has MCA from N.M.S.S.V.N College, Madurai & MPhil (CS) from Manonmaniam Sundaranar University. He has published and presented over 30 research papers in international and national journals in the area of computer science and management. He is an AIMA certified Accredited Management Teacher in the Area of Information Technology, SMFI certified Strategic Management Teacher and has ITIL V3 Foundation-level certificate in Service Management. He is actively involved in the various professional bodies like NHRD, CSI, ISTE, AIMA and ACS. His Areas of Interest includes Business Intelligence, Knowledge Management, Strategic Management, Social Networking Analysis and Artificial Neural Networks.