International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

Information Retrieval Using Customized Ontology Model Based On Clustering Technique

Dr. Santhi Baskaran[1], Tamizharuvi.G[2], Maharajneesha.J[3], Poorni.D.C[4]
[1234] Department of Information Technology , Pondicherry Engineering College, Puducherry, India

Abstract*— The explosion of data leads to the problem on how information should be retrieved accurately and effectively. To address this issue, ontologies are widely used to represent user profiles in personalized web information gathering. Most models use only knowledge from either a global knowledge base or user local information. In this paper, a non-content based customized ontology model is proposed for knowledge representation and reasoning over user profiles. This model generates user Local Instance Repository which includes non-content based descriptors referring to the subjects. The proposed customized ontology model is evaluated by comparing it against the previously proposed content-based ontology model for web information gathering. The result shows that this model has improvement over the former models in the hit/miss ratio, recall and precision parameters.

Keywords*— Ontology, user profiles, non-content based descriptors, local instance repository, global knowledge

I. INTRODUCTION

In recent times, the amount of web information has exploded rapidly. Gathering useful information from the web has become a challenge for the web users. In most of the models that has been developed to solve this issue, user profiles has been created for extracting user background knowledge [1],[5],[9],[10].

User profiles contain the concept model which represents the background knowledge possessed by the users. A superior representation of user profiles can be built by simulating user's concept model. A concept model is implicitly possessed by users and is generated from their background knowledge. While this concept model cannot be proven in laboratories, many web ontologists have observed in user behaviour [10].

For simulating user concept models, ontologies—a knowledge description and formalization model—are utilized in personalized web information gathering. Such ontologies are called ontological user profiles or personalized ontologies.

The user background knowledge can be analysed through global and local analysis. Global analysis uses the existing global knowledge bases for user background knowledge representation. Local analysis is used for extracting user behaviour from the user profiles. Both global and local information are used for discovering the user background knowledge in a better way. This discovery can be further improved by using ontological user profiles.

The commonly used knowledge bases include generic ontologies e.g. Word net, Thesauruses, digital libraries. Word Net was reported as helpful in capturing user interest. It is used in creating ontological user profiles.

The goal of ontology learning is to semi-automatically extract relevant concepts and relations from a given corpus or other kinds of data sets to form ontology. In this paper, a customized ontology to evaluate this hypothesis is proposed.

The ideas which we have implemented in this paper:

1. Global search produces search results based on the existing global knowledge.
2. Local search produces search results based on the user interest which is analysed using user profiles.
3. Content-based clustering is done which searches not only the query with the document name but also with the content present in it.

All local and global repositories have content-based descriptors referring to the subjects. However, a large volume of documents existing on the web may not have such content-based descriptors. To refer those non-content based descriptors clustering technique is used which also groups the documents which does not have

descriptors. Compared with other benchmark models customized ontology model is successful.

II. OVERALL DESIGN

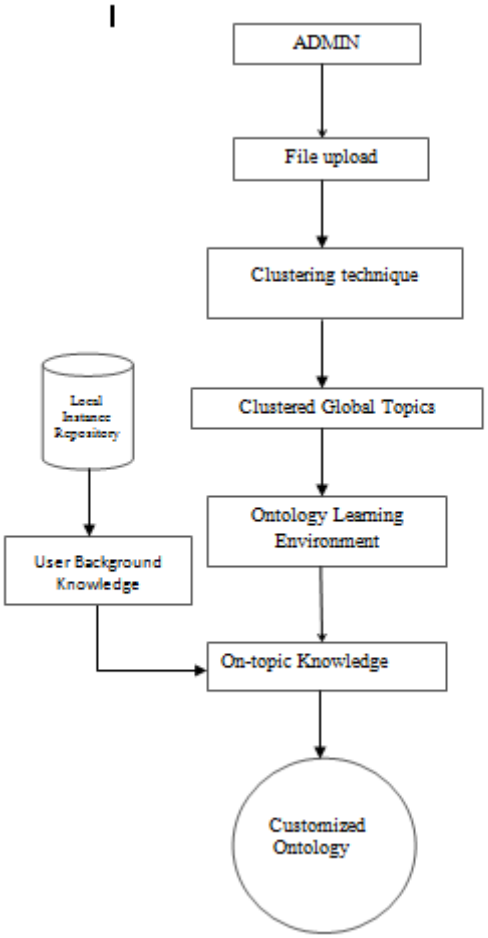

Fig 1. Overall Design for Web Ontology

III. RELATED WORK

A. *Customized Search*

Customized search discovers user background knowledge from the local instance repository which is known as user profiles. User profiles maintains the documents which a user views and downloads based on his/her interest. Customized search produces results based on the user interest so that the user gets the required and efficient results.

B. *User Profile*

For capturing the user information needs User Profiles were used in web Information gathering. A user profile is a collection of personal data associated to a specific user. A profile refers therefore to the explicit digital representation of a person's identity [11]. A user profile can also be considered as the computer representation of a user model. A profile can be used to store the description of the characteristics of person.

User profiles are categorized into three groups: Interviewing, semi-interviewing, and non-interviewing. Interviewing user profiles are considered to be perfect user profiles. They are acquired by using manual techniques, such as questionnaires, interviewing users, and analyzing user classified training sets. One typical example is the TREC Filtering Track training sets, which were generated manually [12]. The users read each document and gave a positive or negative judgment to the document against a given topic.

Semi-interviewing user profiles are acquired by semi-automated techniques with limited user involvement. These techniques usually provide users with a list of categories and ask users for interesting or non-interesting categories. One typical example is the web training set acquisition model introduced by Tao et al. [13], which extracts training sets from the web based on user fed back categories. Non interviewing techniques do not involve users at all, but ascertain user interests instead. They acquire user profiles by observing user activity and behavior and discovering user background knowledge [14].

A typical model is OBIWAN, proposed by Gauch et al. [15], which acquires user profiles based on users' online browsing history. The interviewing, semi-interviewing, and non-interviewing user profiles can also be viewed as manual, semiautomatic, and automatic profiles, respectively.

C. *Ontology Learning Environment*

The subjects of user interest are extracted from the World Knowledge base via user

interaction. Ontology Learning Environment is developed to assist users with such interaction. A topic is divided into two subjects: Positive subjects and Negative subjects.

Positive subjects are the concepts relevant to the information needed by the user. Negative subjects are the concepts which resolve paradoxical or ambiguous interpretation of the information need. The subjects which are not categorized by the users become the neutral subjects to the given topic.

IV.CUSTOMIZED ONTOLOGY CONSTRUCTION USING CLUSTERING TECHNIQUE

Customized ontology is constructed which describes the user background knowledge. For example a user might have different expectations for searching the same query. For example if we are searching for the term "Singapore", business travellers may expect different search from leisure travellers. A user's concept model may change according to different information needs.

Constructing a customized ontology groups the related documents for the given query. But the documents which have contents which is related to the query will be left unsearched. Using clustering technique the content-based clustering is done which searches not only the query with the document name but also with the content present in it.

A. *World Knowledge Representation*

Global Knowledge representation is the analysis of how to accurately and effectively reason and how best to use a set of symbols to represent a set of facts with in a knowledge domain.

In this model user background knowledge is extracted from the set of files, documents and links loaded in the server.

The initial step is the construction of world knowledge base. The user expects various results for searching a single query so the world knowledge base should cover the wide range of topics.

The World Knowledge base is created by the administrator. The administrator uploads files, documents and links which are commonly referred by the users.

B. *Ontology Construction*

An ontology is constructed using the feedback provided for the subjects by the user for the given topic. The structure of the ontology is based on the semantic relations linking those subjects.

Depending on the users interest the subjects are provided ranks and based on the ranks the data are classified and the customized ontology for each user is constructed.

During the global search, after the construction of ontology the data are retrieved based on the information given in the user's profile.

V.PROPOSED MODEL

In the proposed model two types of search operations are performed. The two types of search operations are global search and customized search.

The global search considers the subjects provided in the world knowledge base. The customized search considers only the subjects provided by the individual based on their interests.

Clustering is used in the information retrieval systems to enhance the efficiency and effectiveness of the retrieval process.

Clustering is a division of data into groups of similar objects. Each group consists of objects that are similar between themselves and dissimilar to objects of the group. In our proposed model the concept of clustering is applied at the initial level i.e. global knowledge representation level, which makes the user to search in the respective domain of the given key word. This will results in effective search and the accurate output.

We use relationships between the keywords to cluster the documents. The relationships are retrieved from the Word Net ontology and represented in the form of a graph. The document graphs, which reflect the essence of the documents, are searched in order to find the frequent sub graphs. To discover the frequent sub graphs, we use the Frequent Pattern Growth (FP-growth) approach. The common frequent sub graphs discovered by the

FP-growth approach are later used to cluster the documents.

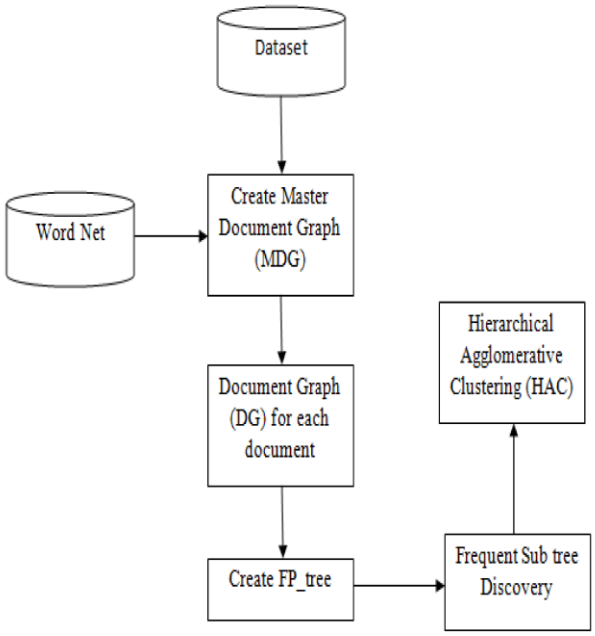

Fig. 2 Overall mechanism of graph-based document clustering using FP-growth

The goal of this model is to cluster text and word documents based on their senses rather than keywords, we use Hierarchical Agglomerative Clustering (HAC) technique.

HAC for given *n* elements it creates a hierarchy of clusters such that at the bottom level of the hierarchy every element is considered as a single independent cluster and the top level all the elements are grouped in a single cluster. It does not require more number of clusters as input since the desired number of clusters can be achieved by cutting the hierarchy at a desired level. It has two approaches Agglomerative and Divisive. Agglomerative merges the closest pair of elements into a single cluster whereas Divisive groups all the elements in a single cluster.

Here we have implemented Group Average method to cluster the documents where the distance between two clusters is defined by the average distance between points in both the clusters and Cosine measure to find the similarity between the clusters.

To cluster the documents we have used a dissimilarity matrix which stores the dissimilarity between every pair of document-graphs using the formula *dissimilarity = 1-similarity*. The value ranges from 0 to 1.

User's interest is derived from the analysis of result which he/she searches in the clustered document of the global knowledge repository. The user can perform customized search in which the results for the key word which user inputted is based on both the derived user's interest and the On- topic knowledge. This will result in effective search and produces the accurate output for the user.

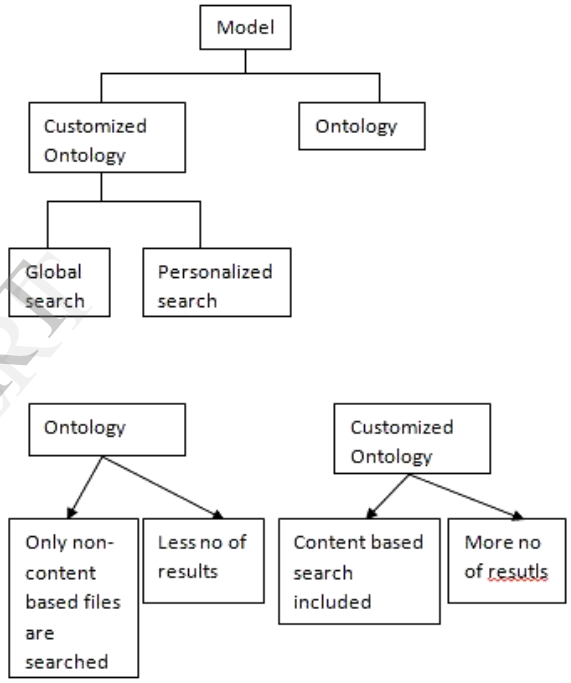

Fig. 3 Comparison between simple Ontology model and Customized ontology model

In this Fig.3, the global information is retrieved based on the local database which is uploaded by the admin. The uploaded files are clustered because of this time consumption for execution is very less and it gives accurate results, cost is also reduced [20]. The search considers both the content and non-content based descriptors for retrieving the data so it fetches result only when the key word exactly matches with the file name or the

content present inside the text document and it produces the absolute results.

VI.PROPOSED FP GROWTH ALGORITHM

Pseudocode:

1. *Create a list called transactionDB for all* $DG_i \in DB$
2. *Create headerTable for all edge* $a_i \in MDG$
3. *FilterDB (transactionDB,headerTable,min_sup)*
4. *FPTreeConstructor()*
5. *FPMining()*
6. *For each sub graph* $subGraph_i$ *include* $SubgraphSupportDocs(subGraph_j)$

Input Documents graphs' database DB
Master Document graph MDG
Minimum support *min_sup*
Output Frequent sub graphs $subGraph_j$

We proposed this algorithm to discover frequent connected sub graphs. We start by creating a hash table called *transactionDB* for all the DG_s which is similar to original FP-growth procedure. Then *headerTable* is created from all the edges appearing in the MDG. After that *FilterDB()* method is called to sort the sub graphs in descending order by frequency based on the *min_sup* provided by the user. *TransactionDB* is then updated by pruning the header table at top and bottom for a second time to reduce too specific and abstract edges. After this refinement, FP tree is created by calling the *FPTreeConstructor()* method. Later, the method *FPMining()* generates the frequent sub graphs by traversing the FP-tree.

VII.COMPARISON AND ANALYSIS OF RESULT

The results produced in a search can be measured based on the recall and precision parameter.

Recall is the no of accurate results produced for a search and precision is the accuracy of the result for a search which is based on the user interest.

A Experimental analysis

The experiments were designed to compare the information retrieval performance achieved by using the proposed customized ontology model, to that achieved by using the ontology model.

The comparison is modeled as an graph in Fig.1, 2 and 3.In Customized Ontology model user profiles are used as an aid to search and the search is made efficient by also considering the documents which does not have content based descriptors.

Recall parameter value is increased in our proposed system as the search includes both content based and non-content based descriptor documents. Number of accurate results produced is increased as the search results also displays documents which are searched based on the contents present in it.

In the Fig 4 we consider only the content based descriptors documents. In that the no of results produced will be same for both the existing ontology model and our proposed customized ontology model.

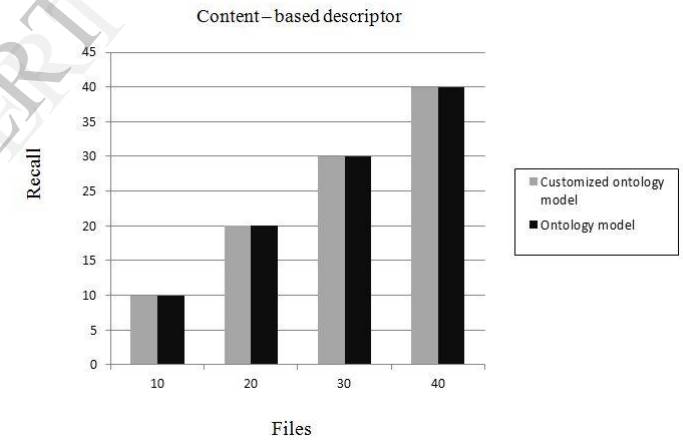

Fig 4.Comparsion of results for content based descriptor documents

But most of the documents do not have content based descriptors. In that case the performance of the existing ontology model degrades by producing less no of results as the search will be carried out only for the documents having content based descriptors.

Fig 5 shows improvement of recall value over the ontology model as the customized ontology model also includes the documents which does not have content based descriptors in the search.

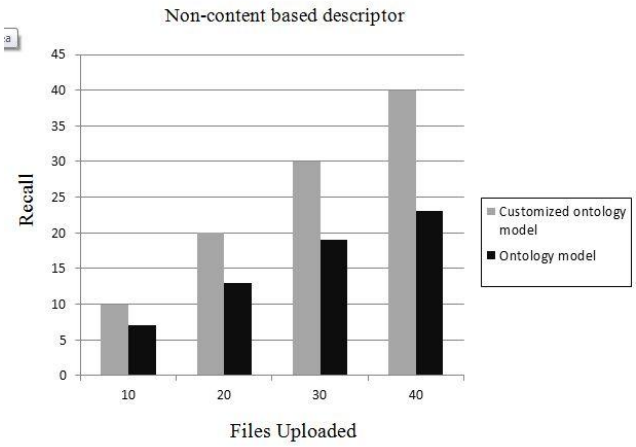

Fig 5.Comparing the results with the inclusion of non-content based descriptor documents

Fig 3 the precision value is mapped with the no of files uploaded. The precision value will be increased as more number of relevant documents are displayed for the user. The results will be more precise as the results will be based on the user's on topic background knowledge.

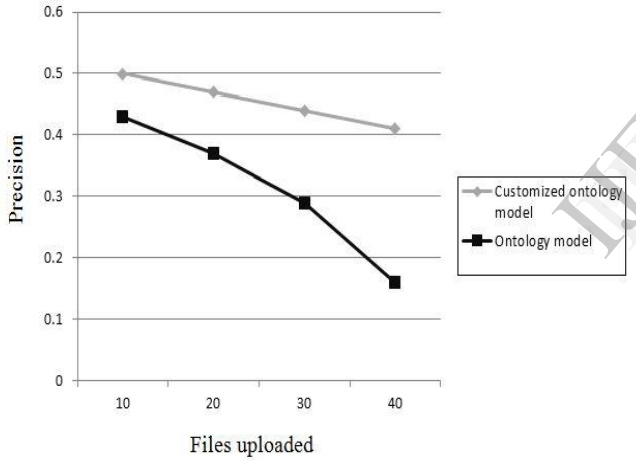

Fig6.Comparision of precision parameter

B.Methodology

The LGSM (Local Global search methodology) it is used to calculate the hit/miss rate. For calculating hit ratio,

$$\text{Hit Ratio} = \frac{\text{No of hits}}{(\text{No of hits} + \text{No of miss})}$$

The performance of memory is frequency measured in terms of quantity is called hit ratio. When cpu needs to find the word in cache, if word is found in cache then it produces a hit. If the word is not found in the cache, it is in main memory it is counted as miss. If it retrieves information from the local repository it is considered as hit. If it retrieves data directly from global it is considered as miss[13].

VIII.CONCLUSION AND FUTURE WORK

A. Conclusion

The customized ontology model for information retrieval performs better in producing the accurate results by clustering the text documents based on its content. Clustering of documents improves the recall parameter by 80%. This in-turn increases the precision parameter value. Since the correctness of the results is more, the user can find documents relevant to his interest in a single search.

B. Future direction

Future work will experiment the algorithm in which search can be extended for all kinds of documents by varying parameters. Multilingual concepts can be introduced. Since the ontologies are constructed in the language that the developers are used to, the search query and the result will be of the same language. For a person who does not know that language will not be able to do the search. So before the customized ontology module a dictionary/wordnet module can be introduced to retrieve all semantic words related to the given keyword and then a multilingual terms module in order to get those words in the language that the user specified. This extends the system for different languages which allows people of different languages to make use of the system.

ACKNOWLEDGMENT

We wish to thank Dr. Santhi Baskaran, Associate Professor, Pondicherry Engineering College for her time and guidance.

REFERENCES

[1] Xiaohui Tao, Yuefeng Li and Ning Zhong, "A Personalized Ontology Model for Web Information Gathering", IEEE Transactions on Knowledge and Data Engineering, vol.23, no.4, April 2011.
[2] C. Buckley and E.M. Voorhees, "Evaluating Evaluation Measure Stability," Proc. ACM SIGIR '00, pp. 33-40, 2000.
[3] L.M. Chan, Library of Congress Subject Headings: Principle and Application. Libraries Unlimited, 2005.
[4] D. Downey, S. Dumais, D. Liebling, and E. Horvitz, "Understanding the Relationship between Searchers' Queries and Information Goals," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp.449-458, 2008.
[5] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003
[6] S.E. Robertson and I. Soboroff, "The TREC 2002 Filtering Track Report," Proc. Text REtrieval Conf., 2002.
[7] Sieg, B. Mobasher, and R. Burke, "Web Search Personalization with Ontological User Profiles," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 525-534, 2007.
[8] M.D. Smucker, J. Allan, and B. Carterette, "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation,"Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 623-632, 2007.
[9] Y. Li and N. Zhong, "Web Mining Models and Its Applications for Information Gathering," Knowledge-Based Systems, vol. 17, pp.207-217, 2004.
[10] Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.
[11] Claudia Marinica and Fabrice Guillet,"Knowledge based interactive postmining of association rules using Ontology" IEEE Transactions on Knowledge and data engineering, Vol. 22, NO. 6, June 2010.
[12] S.E. Robertson and I. Soboroff, "The TREC 2002 Filtering Track Report," Proc. Text Retrieval Conf., 2002.
[13] X. Tao, Y. Li, N. Zhong, and R. Nayak, "Automatic Acquiring Training Sets for Web Information Gathering," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 532-535, 2006.
[14] J. Trajkova and S. Gauch, "Improving Ontology-Based User Profiles," Proc. Conf. Research 'Information Assistee par Ordinateur (RIAO '04), pp. 380-389, 2004.
[15] S. Gauch, J. Chaffee, and A. Pretschner,"Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp.219-234, 2003.