Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 1, January 2014, pg.204 – 209

RESEARCH ARTICLE

An Optimized Solution to Map Web User Profile Based on Domain Ontology

Prof. Ratheesh Kumar A.M[1], Prof. Velusamy A[2], Prof. Shobana G[3]

[1]Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India

[2]Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India

[3]Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India

Abstract*— Ontology defines concepts, attributes and relations used to describe and represent an area of knowledge. The aim of this paper is to create a personalized ontology for web information gathering using language processing techniques. When representing user profiles, many models have utilized only knowledge from either a global knowledge base or user's local information. To generate user's local instance repositories to match the representation of a global knowledge base and to develop a combined ontology using strategies like ontology mapping technique, text categorization, jakard and cosine similarity methods are used to evaluate the efficiency.

***Keywords*— Ontology; Text categorization; Jakard and Cosine Techniques**

I. INTRODUCTION

The concept of ontology have recently received popularity in the area of knowledge management, knowledge description and formulization model. Web personalization alleviates the burden of information overload by tailoring the information presented based on an individual user's needs. Every user has a specific goal when searching for information through entering keyword queries into a search engine. Keyword queries are inherently ambiguous but often formulated while the user is engaged in some larger task. In recent years, personalized search has attracted interest in the research community as a means to decrease search ambiguity and return results that are more likely to be interesting to a particular user and thus providing more effective and efficient information access.

We present a novel approach for building ontological user profiles by assigning interest scores to existing concepts in domain ontology. Many ontology-based information sharing approaches rely on mapping between ontologies from different sources. Mapping tools use different techniques to suggest matches between ontology elements, and vary in input requirements, output formats, and modes of interaction with the user. Due to their diversity, there has been little work on the comparative evaluation of mapping techniques in the information integration literature. Thus, there is a lack of understanding of their pitfalls with real World data. User background knowledge can be better discovered and represented if we can integrate global and local analysis within a hybrid model. Such a personalized

Ontology model should produce a superior representation of user profiles for web information gathering. The world knowledge and a user's local instance repository (LIR) are used in the proposed model. World knowledge is commonsense knowledge acquired by people from experience and education; an LIR is a user's personal collection of information items. From a world knowledge base, we construct personalized ontologies by adopting user feedback on interesting knowledge.

A multidimensional ontology mining method, Specificity and exhaustively, is also introduced in the proposed model for analyzing concepts specified in ontologies. Two approaches are used to make the search efficient. The one approach is based on text categorization and the other approach is based on ontology mapping technique.

II. RELATED WORKS

Effective personalization (fig 1) of information access involves two important challenges: accurately identifying the user context and organizing the information in such a way that matches the particular context. Since the acquisition of user interests and preferences is an essential element in identifying the user context, most personalized search systems employ a user modelling component. In ontology matching, an ontology species a conceptualization of a domain in terms of concepts, attributes, and relations. The concepts provided model entities of interest in the domain. They are typically organized into a taxonomy tree where each node represents a concept and each concept is a specialization of its parent.

III. ARCHITECTURE OF HYBRID MODEL

Global knowledge bases were used by many existing models to learn ontologies for web information gathering. Aiming at learning personalized ontologies, many works mined user background knowledge from user local information. The use of data mining techniques in these models leads to more user background knowledge being discovered. However, the knowledge discovered in these works contained noise and uncertainties.

Additionally, ontologies were used in many works to improve the performance of knowledge discovery. Using a fuzzy domain ontology extraction algorithm User profiles were used in web information gathering to interpret the semantic meanings of queries and capture user information needs. User profiles can be categorized into three groups: interviewing, semi-interviewing, and non interviewing. Interviewing User profiles can be deemed 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. In this paper, the model is proposed to combine both local instance repository and global knowledge called as the hybrid model. The proposed ontology model aims to discover user background knowledge and learns personalized ontologies to represent user profiles.

A personalized ontology is constructed, according to a given topic. Two knowledge resources, the global world knowledge base and the user's local instance repository, are utilized by the model. The world knowledge base provides the taxonomic structure for the personalized ontology. The user background knowledge is discovered from the user local instance repository. Against the given topic, the specificity and exhaustively of subjects are investigated for user background knowledge discovery. The efficiency of the model is evaluated using cosine and jakard similarity. The proposed ontology model aims to discover user background knowledge and learns personalized ontologies to represent user profiles.

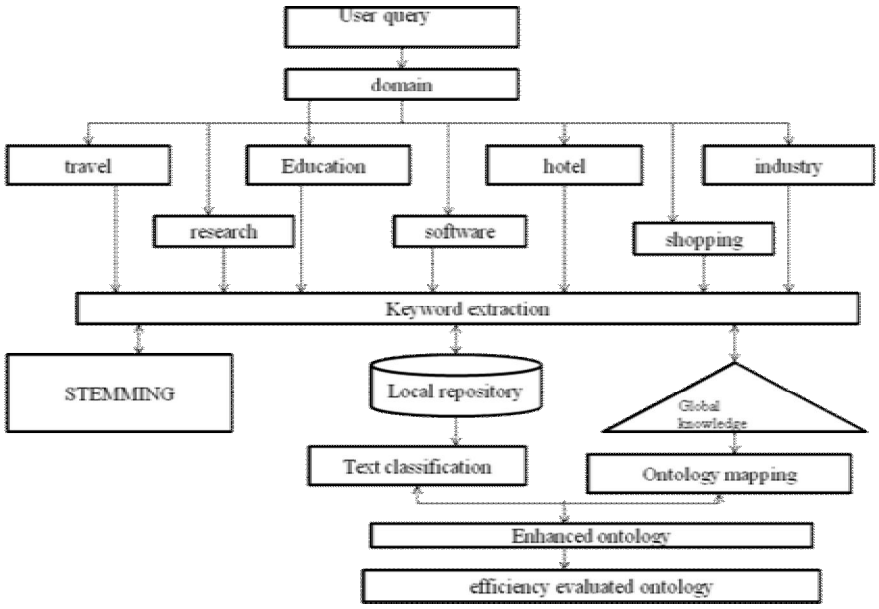

Fig.1 Architecture of Hybrid model

Fig. 1 illustrates the architecture of the hybrid model. A personalized ontology is constructed, according to a given topic. Two knowledge resources, the global world knowledge base and the user's local instance repository, are utilized by the model. The world knowledge base provides the taxonomic structure for the personalized ontology. The user background knowledge is discovered from the user local instance repository. Against the given topic, the specificity and exhaustivity of subjects are investigated for user background knowledge discovery. The efficiency of the model is evaluated using cosine and jakard similarity. Ontology mapping technique and text categorization are the two approaches to increase the efficiency based on various domains.

IV. IMPLEMENTATION

Personalized ontologies generally refer to the conceptual model that refers to the user background knowledge. Since semantic knowledge is an essential part of the user context, we use domain ontology as the fundamental source of semantic knowledge in our framework. Using ontologies as the basis of the profile allows the initial user behaviour to be matched with existing concepts in the domain 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 Web user. Therefore, we define our ontology as a hierarchy of topics, where the topics are utilized for the classification and categorization of Web pages. The hierarchical relationship among the concepts is taken into consideration for building the ontological user profile as we update the annotations for existing concepts using spreading activation.

Identifying semantic correspondences (mappings) between ontologies and database schemas has been the focus of many works from diverse communities. There are two major approaches for discovering mappings between ontologies. If the ontologies share the same upper model, then this common grounding can be used to establish mappings. These are similar to schema matching techniques but sometimes use automated reasoning to identify hierarchies. Some tools also use other external reference ontologies to establish mappings.

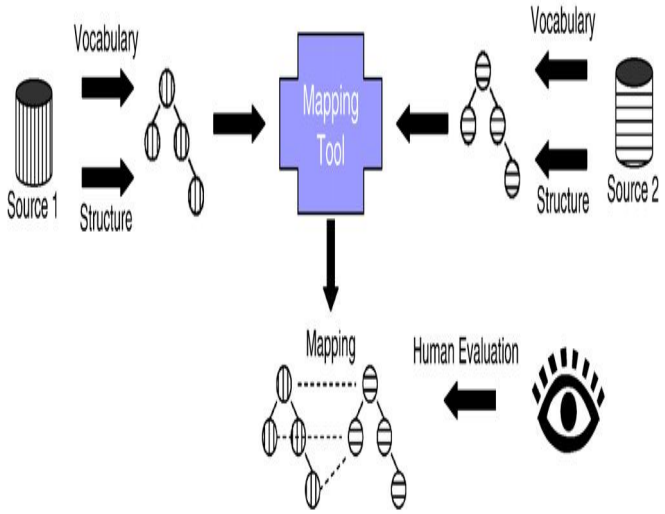

Fig. 2 Research design: Evaluating ontology mapping tools

A. Training the classifier

In order to train the classifier for each concept, all the Web pages available as training data for a particular concept are merged together to create a super document. This creates a collection of super documents, one per concept, that are preprocessed to remove stop words and stemmed using the Porter stemmer (Frakes & Baeza, 1992) to remove common suffixes. After preprocessing, the super documents go through an indexing process to calculate and save vectors for each concept that store the weight of each vocabulary term in that concept. Thus, each concept is treated as *n* dimensional vectors in which *n* represent the number of unique terms in the vocabulary. Each term weights in the concept vectors are calculated using *tf*idf* and normalized by their length. In more detail, *uwij*, the un-normalized weight of term *i* in concept *j*, is calculated as follows:

Where *tfij*= number of occurrences of *ti* in *sdj*

$$\text{idf i} = \frac{\text{log \# of documents in the collection}}{\text{log\# of documents in the collection that contain term t1}}$$

B. Classifying the web pages

Web page and concepts are represented as in the vector space model and the similarity is calculated as cosine between the vectors. Similar to the preprocessing done on the concept upper documents, each of the Web pages collected for a user are preprocessed to remove stop words and then stemmed. The weights for all remaining words in the Web page are calculated using formula (1) and then the words are sorted by weight. Since the words are all selected from the current Web page, the length of the document is a constant and normalization is not done. Based on earlier experiments (Gauch *et al*, 2003), the highest weighted 20 words are used to represent the content of the Web page. Classification thus consists of comparing the vector created for the Web page with each concept's vector (created and stored during training) using the cosine similarity measure.

C. Approaches

1. Exhaustivity and specificity

Ontology mining discovers interesting and on-topic knowledge from the concepts, semantic relations, and instances in ontology. In this section, a 2D ontology mining method is introduced: Specificity and Exhaustivity. Specificity (denoted spe) describes a subject's focus on a given topic. Exhaustivity (denoted exh) restricts a subject's semantic space dealing with the topic. This method aims to investigate the subjects and the strength of their associations in ontology.

2. POS Mapping

The semantic specificity is investigated based on the structure of O(T) inherited from the world knowledge base. The strength of such a focus is influenced by the subject's locality in the taxonomic structure taxS of O(T).The taxS of O(T) is a graph linked by semantic relations. The determination of a subject's spea is described in Algorithm 1. The isA(s') and part of(s') are two functions in the algorithm satisfying is A(s')npartOf(s')=$;. Algorithm 1 is efficient with the complexity of only O(n), where n=|Sj|. The algorithm terminates eventually because taxS is a directed acyclic graph, as defined in Definition 4.

As the taxS of O(T) is a graphic taxonomy, the leaf subjects have no descendants. Thus, they have the strongest focus on their referring-to concepts and the highest spea(s) By setting the spea range as (0, 1] (greater than 0, less than or equal to 1), the leaf subjects have the strongest spea(s) of 1, and the root subject of taxS has the weakest spea(s) and the smallest value in (0, 1]. Toward the root of taxS, the spea(s) decreases for each level up. A coefficient _ is applied to the spea(s) analysis, defining the decreasing rate of semantic specificity from lower bound toward upper bound levels. (_ ¼ 0:9 was used in the related experiments presented in this paper.) The "part-of" relationship is an important relationship between classes in which objects representing the components of something are associated with an object representing the entire assembly. The most significant property of "part-of" is "transitivity", that is, if *A* is part of *B* and *B* is part of *C*, then *A* is part of *C*. "part of" is also "anti-symmetric", that is, if *A* is part of *B* and *A≠B*, then *B* is not part of *A*. Apart from the "part-of" relationship, some classes may have common properties (i.e., they have common base class).

V. EVALUATION

A. Experiment Design

The proposed ontology model was evaluated by objective experiments. Because it is difficult to compare two sets of knowledge in different representations, the principal design of the evaluation was to compare the effectiveness of an information gathering system (IGS) that used different sets of user background knowledge for information gathering. The knowledge discovered by the ontology model was first used for a run of information gathering, and then the knowledge manually specified by users was used for another run. Under the same experimental conditions, if the IGS could achieve the same (or similar) performance in two different runs, we could prove that the discovered knowledge has the same quality as the user specified knowledge. The proposed ontology model could then be proven promising to the domain of web information gathering.

User profiles can be categorized into three groups:

Interviewing, semi-interviewing, and non interviewing profiles, as previously discussed in Section 2. In an attempt to compare the proposed ontology model to the typical models representing these three group user profiles, four models were implemented in the experiments:

1. The Ontology model that implemented the proposed ontology model. User background knowledge was computationally discovered in this model.

2. The TREC model that represented the perfect interviewing user profiles. User background knowledge was manually specified by users in this model.

3. The Category model that represented the non interviewing user profiles.

4. The Web model that represented the semi-interviewing user profiles.

Most retrieval evaluation measures are derived in some way from *recall* and *precision*, where precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents that are retrieved. An exception are measures based on utility-theory for which the quality of retrieval output is measured in terms of its worth to the user. Utility-based measures are frequently used to evaluate set-based retrieval output such as in the TREC filtering task.

VI. DOCUMENT AND QUERY REPRESENTATION

The document representation is important in determining both what terms (*i*) are included and how often they occur (*tfi*). Using the full text of documents in the results set is a natural starting place. However, accessing the full text of each document takes considerable time. Thus, we also experimented with using only the title and the snippet of the document returned by the Web search engine. We note that because the Web search engine we used derived its snippets based on the query terms, the snippet is inherently query focused. In the absence of any information other than the user's query, a document's score is calculated by summing over the query terms, the product of the query term weight (*wi*) and the query term occurrence in the document (*tfi*). However, when relevance feedback is used, it is very common to use some form of query expansion. A straightforward approach to query expansion that we experimented with is the inclusion of all of the terms

occurring in the relevant documents. This is a kind of blind or *pseudo relevance* feedback in which the top-*k* documents are considered relevant

VII. CONCLUSION

In this paper, in order to provide each user with more relevant information, we proposed several approaches to adapting search results according to each user's information need. Our approach is novel in that it allows each user to perform a fine-grained search, which is not performed in typical search engines, by capturing changes in each user's preferences. We have investigated the feasibility of personalizing Web search by using an automatically constructed user profile as relevance feedback in our ranking algorithm. Research suggests that the most successful text-based personalization algorithms perform significantly better than explicit relevance feedback where the user has fully specified the relevant documents. The hybrid model does provide a enhanced view of the personalized ontology over the domain. The technique proposed in this paper can be applied to situations where users require more relevant information to satisfy their information needs. A hybrid model is created using ontology mapping technique based on two approaches and based on the above approaches the efficiency is evaluated.

REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1(1):5–32, 1999.

[2] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.

[3] B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web," ACM SIGIR Forum, vol. 32, no. 1, pp. 5-17, 1998.

[4] X. Jiang and A.-H. Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents," Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM '05), pp. 665-668, 2005.

[5] W. Jin, R.K. Srihari, H.H. Ho, and X. Wu, "Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques," Proc. Seventh IEEE Int'Conf. Data Mining (ICDM '07), pp. 193-202, 2007.

[6] J.D. King, Y. Li, X. Tao, and R. Nayak, "Mining World Knowledge for Analysis of Search Engine Content," Web Intelligence and Agent Systems, vol. 5, no. 3, pp. 233-253, 2007.

