2nd International Conference on Communication, Computing & Security [ICCCS-2012]

Design and development of an ontology based personal web search engine

Sakthi Priya T, Revathy P, Pradeesh T, Rene Robin C.R

Department of Computer Science and Engineering, Jerusalem College of Engineering, Pallikaranai, Chennai-100, India

Abstract

As a model for knowledge description and formalization, ontologies are widely used to represent user profiles in personalized web information gathering. However, when representing user profiles, many models have utilized only knowledge from either a global knowledge base or user local information. A personalized search engine which is a hybrid system has been proposed for personalized search and reasoning over user profiles. The hybrid system contains both the global knowledge base and the user local information. User profiles are created for every users to gather their interest and relevance. This system also learns user profiles based on which the user is provided with candidates in Ontology Learning Environment (OLE) tool. The search results are personalized and have topic specificity. The efficiency of a search application is defined by the accuracy by which the search results match the user interests. The efficiency improvement in our application can be defined by the reduction in the number of pages which a particular user who searches for a string. This search engine is evaluated by getting feedback from three kinds of users. The results show that this ontology model is successful.

* Corresponding author. Tel.: +91-996-295-8785
E-mail address:sakthipriya134 @gmail.com

1. INTRODUCTION:

Over the last decade, we have witnessed an explosive growth in the data accessible on the Web. However, there are two fundamental issues regarding the effectiveness of information gathering from the Web: mismatch and overload. Mismatch means some useful and interesting data has been overlooked, whereas overload means some gathered data is not what users want.

In general, people have two ways to find the data they are looking for: they can search, and they can browse. Search engines index millions of documents on the Internet and allow users to enter keywords to retrieve documents that contain these keywords. Browsing is usually done by clicking through a hierarchy of subjects until the area of interest has been reached. Indeed, in terms of searching, about one half of all retrieved documents have been reported to be irrelevant. The main problem is that there is too much information available, and that keywords are not always an appropriate means of locating the information in which a user is interested. Presumably, information retrieval will be more effective if individual users' idiosyncrasies are taken into account.

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. For example, a historian may enter the query Madonna and child while browsing Web pages about art history, while a music fan may issue the same query to look for updates on the famous pop star.

This paper studies ways to model a user's interests and shows how these models - also called profiles - can be deployed for more effective information retrieval and filtering. This paper shows how the profiles can be used to achieve search performance improvements. The increases in performance are modest, but they are noticeable, and they are a first step.

2. RELATED WORKS:

Global knowledge bases were used by many existing models to learn ontologies for web information gathering. For example, Gauch et al. [2] and Sieg et al. [10] learned personalized ontologies from the Open Directory Project to specify users' preferences and interests in web search. Wikipedia was used by Downey et al. [1] to help understand underlying user interests in queries. These works effectively discovered user background knowledge; however, their performance was limited by the quality of the global knowledge bases. Li and Zhong [6] used pattern recognition and association rule mining techniques to discover knowledge from user local documents for ontology construction Zhong [7] proposed a domain ontology learning approach that employed various data mining and natural-language understanding techniques. Navigli et al. [8] developed OntoLearn to discover semantic concepts and relations from web documents. Web content mining techniques were used by Jiang and Tan [4] to discover semantic knowledge from domain-specific text documents for ontology learning. Finally, Shehata et al. [11] captured user information needs at the sentence level rather than the document level, and represented user profiles by the Conceptual Ontological Graph.

User profiles were used in web information gathering to interpret the semantic meanings of queries and capture user information needs [2], [3], [5], [11], [12]. User profiles were defined by Li and Zhong [5] as the interesting topics of a user's information need. They also categorized user profiles into two diagrams: the data diagram user profiles acquired by analyzing a database or a set of transactions [2], [5], [13], [9], [14]; the

information diagram user profiles acquired by using manual techniques, such as questionnaires and interviews [13], [11] or automatic techniques, such as information retrieval and machine learning. Van der Sluijs and Huben [15] proposed a method called the Generic User Model Component to improve the quality and utilization of user modeling. Wikipedia was also used by [1], [16] to help discover user interests. In order to acquire a user profile, Chirita et al. [8] and Teevan et al. [17] used a collection of user desktop text documents and emails, and cached web pages to explore user interests. Makris et al. [18] acquired user profiles by a ranked local set of categories, and then utilized web pages to personalize search results for a user. World knowledge is important for information gathering. According to the definition provided by[6], world knowledge is commonsense knowledge possessed by people and acquired through their experience and education.

3. PROPOSED SYSTEM ARCHITECTURE:

World Knowledge Base (WKB) is the global knowledge base like Wikipedia, WordNet, thesaurus etc. WKB consists of various domains and its related concepts. Figure 1 shows the overall architecture of the system. The initial step of this system involves the extraction of domains, classes, and its subclasses from the WKB. Proceeding to this, is the construction of ontology with the extracted domain, classes and subclasses.

The ontology construction is made using the Ontology editing tool, Protégé. Protégé is a free, open source ontology editor and knowledge-base framework. The protégé platform supports two main ways of modeling ontologies via the Protégé-Frames and Protégé-OWL editors. In this paper Protégé-OWL editor is used for modeling Ontology. In this paper food Ontology is constructed using the concepts gathered from the WKB. Once the ontology is constructed, the visualization tool like OntoViz, OwlViz, TZViz or Ontosphere is used to visualize the ontology construction. This constructed ontology is converted in to an ontology database (MS-Access).

The next step is the Login Authentication where the user enters the username and password. If the user is not already registered a User profile appears where the user enters their interests along with the username and password. This proceeds to the Ontology Learning Environment (OLE) screen where the user can select their positive and negative candidates. The OLE gets its input in two ways: Ontology database and the User profile. The User profile and the OLE does not appear for a registered user. For already registered user, once he provides the login details it directly leads him to the search page. The user profile and OLE appears once in every 15 days to gather his/her interests.
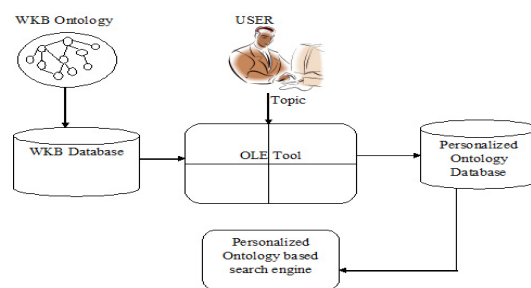

Figure 1 Overall system architecture

Once the user selects the positive and negative candidates, it gets stored in the user's personalized database. Once the user submits after selecting the positive and negative candidates it leads to a search page

where the user can enter their keyword. The search results are based on the users interests i.e., positive and negative candidates. The neutral candidates appear after this.

4. MODULES:

There are four main modules involved in the implementation of the proposed system,

- Protégé and database connectivity
- OLE tool input
- OLE tool creation
- Search engine

4.1 PROTEGE AND DATABASE CONNECTIVITY:

All the subjects regarding food are gathered from the world knowledge base and ontology is constructed using protégé, an Ontology editor tool. This ontology is visualized using ontograph which is shown in Figure 2(a). Then the entire ontology is converted in to the database using the options provided in the protégé. The database used here is MS-Access. Now all the classes and subclasses are saved in the tuples of the database which is illustrated in Figure 2(b).

4.2 OLE TOOL INPUT:

The user is provided with a login page wherein he can enter the username and password as a measure of authentication and this directly leads to search page. In case if the user visits after a period of 15 days. Then automatically the user gets the user profile and the OLE tool where he has to select the positive and negative candidates.

In case, if the user is new to the system then he/she gets a user profile page shown in Figure 2(c) where the user has to enter his/her interests which is asked for based on the food domain because this system is built based on the food domain. This proceeds to the OLE page where the user needs to select the positive and negative candidates. In the search page the user enters the keyword and the search results are generated based on the positive and negative candidates selected by the user.

4.3 OLE TOOL CREATION:

The OLE gets its input from the protégé database. All the candidates are listed in the frame. These candidates are listed based on the user profile of the particular user. For instance if the user selects Vegetarian then only vegetarian related candidates are displayed first in the OLE tool shown in Figure 2(d). The user needs to select the positive and negative candidates based on his/her interest. The positive candidates are generally the parent and some of the children of the parent are selected as negative candidates. Once the user selects the positive and negative, all the details gets saved in the user's personalized database.

4.4 SEARCH ENGINE:

Once the user is done with the OLE tool the search engine gets displayed as shown in Figure 2(e). Now the

user enters the keyword in the textbox and the search results are generated based on his preferences i.e., based on his/her positive and negative candidates selected by him in the OLE screen.

5. IMPLEMENTATION AND RESULTS:

We have implemented the proposed system using JSP. First we have created a authentication form wherein the user sign in or if he is the new user his details are collected through the registration form and the data collected is stored in the back end which is an MS Access DB.
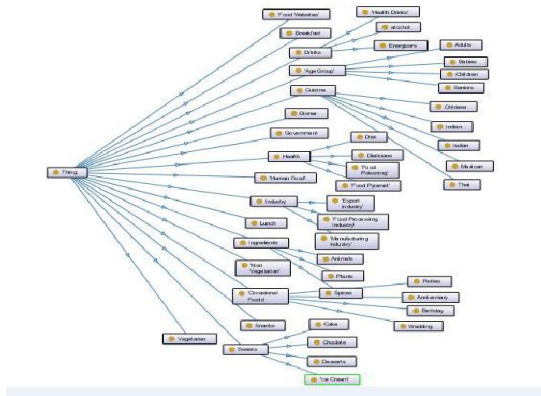

Figure 2 (a) Ontology creation

Figure 2(b) Protégé database

Figure 2 (c) User Profile

Figure 2(d) OLE tool

The user preferences are collected through OLE tool, the positive, negative and neutral candidates are categorized and stored in the backend. The user enters the keyword in the textbox and the search results are generated based on his preferences.
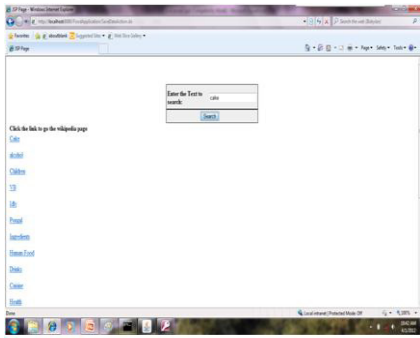
Figure 2(e): Search engine

6. EVALUATION:

The working of the proposed system was demonstrated to a set of users and they were asked to fill a feedback form. The users were categorized into three levels as: primary level users – users who doesn't have any knowledge in this field, secondary level users – users who have minimal knowledge in this field and expert level users – users who are expert in this field. Totally 30 users gave the feedback. All the users were given a detailed explanation about the working of the system before they were asked to assess. There were totally 20 questions in the feedback form and they were categorized based on four parameters, they are: Accessibility, Navigation, Security and Efficiency.

The results of the evaluation may be visualized through the following graph. The questions were categorized into four parameters and the results of the evaluation are as follows:
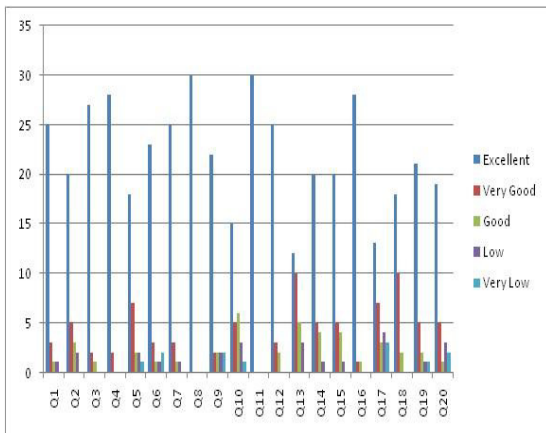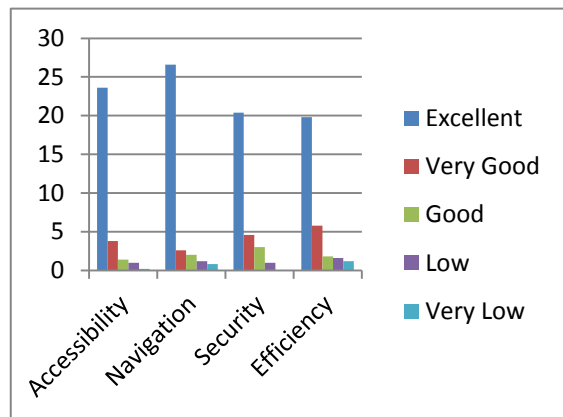

Figure 3(a): Qualitative Evaluation

Figure 3(b): Parameterized Evaluation

Figure 3(c) shows the comparison of the system with already existing systems. Around 20 features were identified and compared. Thus the graph result in Figure 3(a), Figure 3(c) and Figure 3(b) show the system is very efficient in all the four parameters such as accessibility, navigation, security and efficiency.

S.NO	FEATURES	BING	ASK.COM	AOL	YUMMLY	RECEPIE BRIDGE	Ontology based personal search engine
1	Personalized search	No	No	No	No	No	Yes
2	Topic specificity	No	No	No	No	No	Yes
3	Ontology Based	No	No	No	No	No	Yes
4	Authenticated	No	Yes	No	No	Yes	Yes
5	User Profile	No	No	No	No	No	Yes
6	Accuracy	Yes	No	No	Yes	Yes	Yes
7	Clarity	Yes	Yes	Yes	Yes	Yes	Yes
8	Ambience	No	No	No	Yes	Yes	Yes
9	Security	No	No	No	No	No	Yes
10	Min. no. of clicks	No	No	No	Yes	Yes	Yes
11	Comprehensiveness	No	No	No	No	No	Yes
12	Authority	Yes	No	No	Yes	No	Yes
13	User interest based	No	No	No	No	No	Yes
14	Operability	Yes	No	No	Yes	No	Yes
15	Genuineness	Yes	Yes	No	No	Yes	Yes
16	Maintenance	Yes	No	Yes	No	No	Yes
17	Noisy	Yes	Yes	Yes	Yes	Yes	No
18	Superior	No	No	No	No	No	Yes
19	Richness	Yes	Yes	Yes	No	Yes	Yes
20	Time gain	No	No	No	No	No	Yes

Figure 3 (c) Evaluation of features of the proposed system

7. CONCLUSION:

The implementation of personalized ontology model for web information gathering has been successfully implemented. The search engine must be considered much more efficient than the existing search engines for the following reasons.

- Better alignment with the user interests which are captured by the user profile and subjects selected by the user.

- Better response times due to the reduction in the number of candidates searched to display the result.

- Reliable - Displays the most relevant topics for the user on the priority list of the results Portable - Any number of domains can be added to the existing food industry to cater various industries.

8. FUTURE ENHANCEMENT:

The following can be considered as future enhancement for this application .

1) Usage of Protégé tool to save the user profile in an ontological format – Instead of saving the user profile in the form of a database , the user profile can be saved in the form of ontology. This will help the end users of the system to understand the personalized ontology which has been created. Users who are not familiar with the usage of database can view this diagram to understand the ontology which has been created for the user. Although the protégé tool has a plug in, which can read the database created to save the profile, the actual details for implementing the same is not available. If the specific detail to enable the ontology creation is given , this implementation could be done.

2) Integration of the LCSH world knowledge base - The LCSH world knowledge base can be linked to the application to create an exhaustive list of all the classes present in the world. This will enable the user to enter any string in the search application.

Acknowledgement:

References:

[1] 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.

[2] 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.

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

[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] 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.

[6] N. Zhong, "Representation and Construction of Ontologies for Web Intelligence," Int'l J. Foundation of Computer Science, vol. 13, no. 4, pp. 555-570, 2002.

[7] R. Navigli, P. Velardi, and A. Gangemi, "Ontology Learning and Its Application to Automated Terminology Translation," IEEE Intelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.

[8] P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," Proc. ACM SIGIR ('07), pp. 7-14, 2007.

[9] A. 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.

[10] S. Shehata, F. Karray, and M. Kamel, "Enhancing Search Engine Quality Using Concept-Based Text Retrieval," Proc. IEEE/WIC/ ACM Int'l Conf. Web Intelligence (WI '07), pp. 26-32, 2007.

[11] J. Trajkova and S. Gauch, "Improving Ontology-Based User Profiles," Proc. Conf. Recherche d'Information Assistee par Ordinateur (RIAO '04), pp. 380-389, 2004.

[12] N. Zhong, "Toward Web Intelligence," Proc. First Int'l Atlantic Web Intelligence Conf., pp. 1-14, 2003.

[13] S.E. Middleton, N.R. Shadbolt, and D.C. De Roure, "Ontological User Profiling in Recommender Systems," ACM Trans. Information Systems (TOIS), vol. 22, no. 1, pp. 54-88, 2004.

[14] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW '04), pp. 675- 684, 2004.

[15] K. van der Sluijs and G.J. Huben, "Towards a Generic User Model Component," Proc. Workshop Personalization on the Semantic Web (PerSWeb '05), 10th Int'l Conf. User Modeling (UM '05), pp. 43-52, 2005.

[16] D.N. Milne, I.H. Witten, and D.M. Nichols, "A Knowledge-Based Search Engine Powered by Wikipedia," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 445-454, 2007.

[17] J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities," Proc. ACM SIGIR'05, pp. 449-456, 2005.

[18] C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis, "Category Ranking for Personalized Search," Data and Knowledge Eng., vol. 60, no. 1, pp. 109-125, 2007.