International Journal of Research Studies in Science, Engineering and Technology
Volume 1, Issue 9, December 2014, PP 1-9
ISSN 2349-4751 (Print) & ISSN 2349-476X (Online)

Challenges in Web Search Engines

Sindhupriya Pemmasani [1], P.Vemana[2]

P.G.scholar, Dept.of CSE, Krishnaveni Engineering College for Women, Narasaraopet, Andhra Pradesh, India [1]
Asst Professor, Krishnaveni Engineering College for Women, Narasaraopet, Andhra Pradesh, India [2]
priyasindhu32@gmail.com, vemana006@gmail.com

Abstract: *In this paper we propose a new type of search engine for web personalization approach. It will capture the interests and preferences of the user in the form of concepts of mining search results and their click troughs. A web search engine consists of three parts: (1) A crawler that retrieves web pages to be put into the engine's collection of web pages; (2) an indexer that builds the inverted index (also called the index), which is the main data structure used by the search engine and represents the crawled web pages; (3) and a query handler that answers user queries using the index. In internet, a wide range of web information increases rapidly, user wants to retrieve the information based upon his preference of using search engines. Our approach is to improve the search accuracy by means of separating the concepts into content based concepts and location based which plays an important role in global search. Moreover, recognizing the fact that different users and queries may have different emphasis on content and location information, we introduce the content and location based concepts and achieves their respective results. Additionally, search engine also provides the facility of local search by entering keywords without using internet. And feature of integrity of the search engines at one location so that user can work with different search engines in parallel.*

Keywords: *Google, Web Ontology Language (OWL), Personalization, SpyNB(NAÏVE BAYESIAN), Ontology based Multi-Facet (OMF),WKB (World Knowledge Base).*

1. INTRODUCTION

Search Engines have grown into by far the most popular way for navigating the web. The evolution of search engines started with the static web and relatively simple tools such as WWWW [McB94]. In 1995 AltaVista launched and created a bigger focus on search engines SRR97].

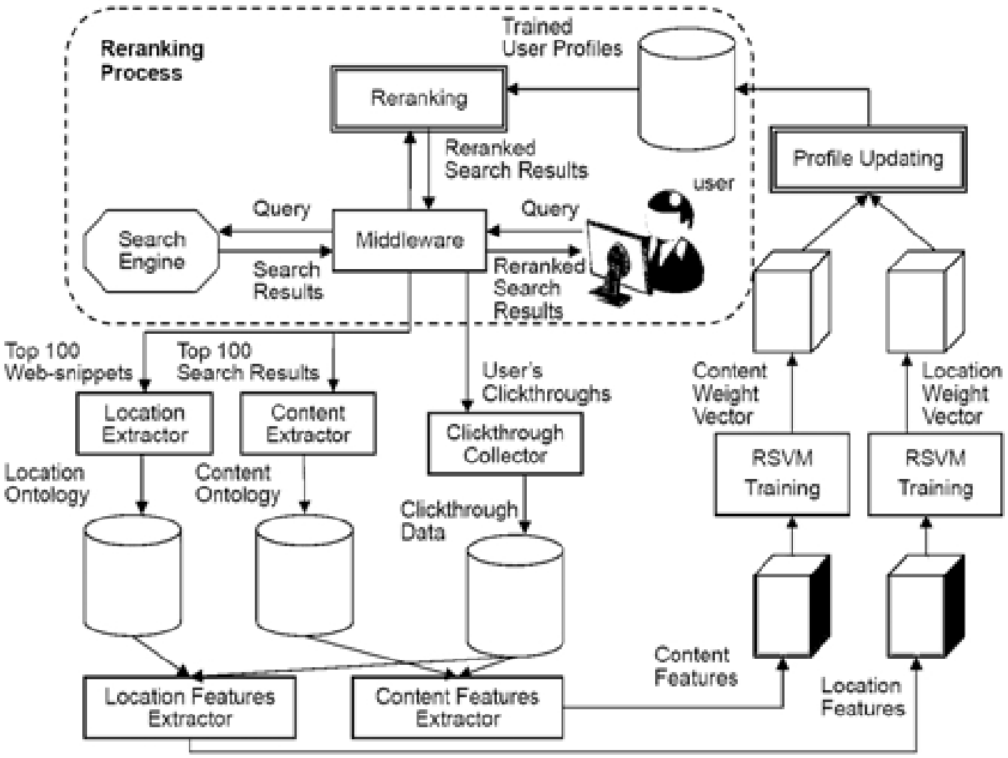

Fig1. *The general process of proposed personalization approach*

The marketplace for search engines is still dynamic, and actors like FAST (www.alltheweb.com), Google, Inktomi and AltaVista are still working on different technical solutions and business models in order to make a viable business, including paid inclusion, paid positioning, advertisements, OEM searching, etc. A large number of analyses have been made on the structure and dynamics of the web itself some information provided is of use to the end users, and others of no use to them. Current web information gathering systems attempt to satisfy user requirements by capturing their information needs. For this purpose, user profiles are created for user background knowledge description. By capturing the users' interests in user profiles, a personalized search middleware is able to adapt the search results obtained from general search engines to the users' preferences through personalized reranking of the search results. The conceptual relationship between the documents has to be represented in order to identify the information that a user wants from those represented concepts. To represent the semantic relation, the ontology is used here. To build a user profile, the Web pages that the user visited are monitored and the system represents the long-term and short-term preference weights as the preference ontology after inferring relevant concepts from the general ontology. At the recommendation stage, the system recommends documents according to user preference concepts and document similarity measure.

We propose an (OMF) user profiling strategy to capture both of the users' content and location preferences (i.e., .multi-facets.) for building a personalized search engine for mobile users. Fig 1 shows the general process of our approach, which consists of two major activities: 1) Reranking and 2) Profile Updating.

1.1 Re Ranking

When a user submits a query, the search results are obtained from the backend search engines (e.g. Google, MSN Search, and Yahoo). The search results are combined and reranked according to the user's profile trained from the user's previous search activities.

1.2 Profile Updating

After the search results are obtained from the backend search engines, the content and location concepts (i.e. important terms and phrases) and their relationships are mined online from the search results and stored, respectively, as content ontology and location ontology. When the user clicks on a search result, the clicked result together with its associated content and location concepts are stored in the user's clickthrough data. The content and location ontologies, along with the clickthrough data, are then employed in RSVM training to obtain a content weight vector and a location weight vector for reranking the search results for the user. There is a number of challenging research issues we need to overcome in order to realize the proposed personalization approach. First, we aim at using concepts to represent and profile the interests of a user. Therefore, we need to build up and maintain a user's possible concept space, which are important concepts extracted from the user's search results. Additionally, we observe that location concepts exhibit different characteristics from content concepts and thus need to be treated differently. Thus, we propose to represent them in separate content and location ontologies. These ontologies not only keep track of the encountered concepts accumulated through past search activities but also capture the relationships among various concepts, which Plays an important role in our personalization process. Second, we recognize that the same content or location concept may have different degrees of importance to different users and different queries. Thus, there is a need to characterize the diversity of the concepts associated with a query and their relevance to the user's need. To address this issue, we introduce the notion of content and location entropies to measure the amount of content and location information a query is associated with. Similarly, we propose click content and location entropies to measure how much the user is interested in the content and/or location information in the results. We can then use these entropies to estimate the personalization effectiveness for a given query, and use the measure to adapt the personalization mechanism to enhance the accuracy of the search results. Finally, the extracted content and location concepts from search results and the feedback obtained from clickthroughs need to be transformed into a form of user profile for future reranking. The ontology-based, multi -facet (OMF) framework is an innovative approach for personalizing web search results by mining content and location concepts for user profiling. To the best knowledge of the authors, there is no existing work in the literature that takes into account both types of concepts. This paper studies their unique characteristics and provides a coherent strategy to integrate them into a uniform solution. A location ontology and content

ontology is proposed here to accommodate the extracted content and location concepts as well as the relationships among the concepts. Based on the proposed ontologies and entropies, an SVM is adapted to learn personalized ranking functions for content and location preferences. The personalization effectiveness is used to integrate the learned ranking functions into a coherent profile for personalized reranking. A working prototype is proposed to validate the proposed ideas. It consists of a middleware for capturing user clickthroughs, performing personalization, and interfacing with commercial search engines at the backend. The rest of the paper is organized as follows. We review the related work in Section II. In Section III, our ontology extraction method is presented for building the upper and lower ontologies. In Section IV, the method to extract user preferences from the clickthrough data to create the user profiles is reviewed. In Section V, the personalized ranking function in discussed to rank the given concepts. The experimental results are displayed in section VI. Section VII concludes the paper.

2. LITERATURE SURVEY

Most commercial search engines return roughly the same results to all users. However, different users may have different information needs even for the same query. For example, a user who is looking for a laptop may issue a query 'apple'. To find products from Apple Computer, while a housewife may use the same query .apple. to find apple recipes. The objective of personalized search is to disambiguate the queries according to the users' interests and to return relevant results to the users. Click through data is important for tracking user actions on a search engine. Many personalized web search systems are based on analyzing users' clickthroughs. Joachims proposed document preference mining and machine learning to rank search results according to user's preferences. More recently, extended Joachims method by combining a spying technique Together with a novel voting procedure to determine user preferences. Leung et al. introduced an effective approach to predict users' conceptual preferences from clickthrough data for personalized query suggestions. The differences between our work and existing works are: Existing works require the users' to manually define their location preferences explicitly (with latitude-longitude pairs or text form). With the automatically generated content and location user profiles, our method does not require users to explicitly define their location interest manually. Our method automatically profiles both of the user's content and location preferences, which are automatically, learnt from the user's clickthrough data without requiring extra efforts from the ser. Our method uses different formulations of entropies derived from a query's search results and a user's clickthroughs to estimate the query's content and location ambiguities and the user's interest in content or location information. The entropies allow us to classify queries and users into different classes and effectively combine a user's content and location preferences to rerank the search results.

3. PROPOSED METHODS

3.1 Concept Extraction

The personalization approach is based on concepts to profile the interests and preferences of a user. An issue to be addressed is how to *extract* and *represent* concepts from search results of the user. An OMF profiling method is proposed in which concepts can be further classified into different types, such as content concepts (location ontology), location concepts (content ontology), name entities, dates etc. An important first step is to focus on two major types of concepts, namely, content concepts and location concepts. A content concept, like a keyword or key-phrase in a Web page, defines the content of the page, whereas a location concept refers to a physical location related to the page. The interests of a search engine user can be effectively represented by concepts extracted from the user's search results. The extracted concepts indicate a possible concept space arising from a user's queries, which can be maintained along with the click through data for future preference adaptation.

3.2 Location Ontology

If a keyword/phrase exists frequently in the web-snippets arising from the query q, it represents an important concept related to the query, as it co-exists in close proximity with the query in the top documents. Thus, our content concept extraction method first extracts all the keywords and phrases from the web-snippets arising from q.

$$\text{Support } (C_i) = \frac{Sf(C_i)}{n} * |(C_i)|$$

After obtaining a set of keywords/phrases (ci), the following support formula, which is inspired by the well known problem of finding frequent item sets in data mining, is employed to measure the interestingness of a particular keyword/phrase ci with respect to the query q: where sf(ci) is the snippet frequency of the keyword/phrase ci (i.e. the number of web-snippets containing ci), n is the number of web-snippets returned and |ci| is the number of terms in the keyword/phrase ci. If the support of a keyword/phrase ci is higher than the threshold s (s = 0:03 in our experiments), where ci is a concept for the query q. As mentioned, the ontologies are used to maintain concepts and their relationships extracted from search results. The location ontology is built here to represent these content concepts. The location ontology is built based on the following types of relationships for content concepts:

Similarity: Two concepts which coexist a lot on the search results might represent the same topical interest. If coexist (ci, cj) > _1 (_1 is a threshold), then ci and cj are considered as similar.

Parent-Child Relationship: More specific concepts often appear with general terms, while the reverse is not true. Thus, if pr (cj,ci) > _2 (_2 is a threshold), where ci as cj 's child.

Fig 2 shows an example content ontology created for the query ‘apple’. Content concepts linked with a double sided arrow ($) are similar concepts, while concepts linked with a one-sided arrow (!) are parent-child concepts. The ontology shows the possible concept space arising from a user's queries. In general, the ontology covers more than what the user actually wants. For example, when the query ‘apple’ is submitted, the concept space for the query composes of MAC, software, fruit... etc. If the user is indeed interested in apple as a fruit and clicks on pages containing the concept ‘fruit’ the clickthrough is captured and the clicked concept fruit is favored. The content ontology together with the clickthrough serves as the user profile in the personalization process.

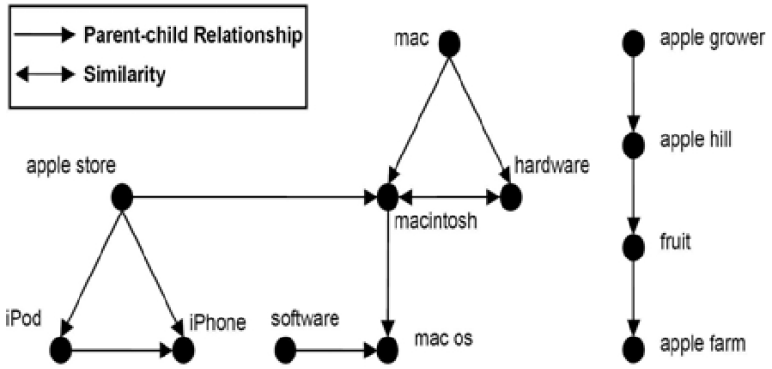

Fig2. *Example Content Ontology Extracted for the Query .apple.*

3.3 Content Ontology

The approach for extracting location concepts is different from that for extracting content concepts. First, a websnippet usually embodies only a few location concepts. As a result, very few of them co-occur with the query terms in web snippets. To alleviate this problem, the location concepts are extracted from the full documents. The content ontology is built to represent these location concepts. Second, due to the small number of location concepts embodied in documents, the similarity and parent-child relationship cannot be accurately derived statistically. Additionally, the content ontology extraction method extracts all of the keywords and key-phrases from the documents returned for q. If a keyword or key-phrase in a retrieved document matches a location name in the predefined location ontology, it will be treated as a Location concept of d. Similar to the content ontology; locations are assigned with different weights according the user’s click through.

4. USER REFERENCE EXTRACTION

Given that the concepts and click through data are collected from past search activities, user's preference can be learned. In this section, two alternative preference mining algorithms, namely, Joachims Method and SpyNB Method are reviewed to adopt in our personalization framework.

4.1 Joachim’s Method

Joachim’s method assumes that a user would scan the search result list from top to bottom. If a user skips a document dj at rank j but clicks on document di at rank I where j < i, he/she must have read

dj's web snippet and decided to skip it. Thus, Joachims method concludes that the user prefers di to document dj (denoted as dj <r′ di, where r′ is the user's preference order of the documents in the search result list).

4.2 SPYNB Method

Similar to Joachim's method, SpyNB learns user behavior models from preferences extracted from clickthrough data. SpyNB assumes that users would only click on documents that are of interest to them. Thus, it is reasonable to treat the clicked documents as positive samples. However, unclicked documents are treated as unlabeled samples because they could be either relevant or irrelevant to the user. Based on this interpretation of clickthroughs, the problem becomes how to predict from the unlabeled set reliable negative documents which are Irrelevant to the user. The details of the SpyNB method can be found to do this; the Spy technique incorporates a novel voting procedure into Naive Bayes classifier. Let P be the positive set, U the unlabeled set and PN the predicted negative set $PN \subset U$ obtained from the SpyNB method. SpyNB assumes that the user would always prefer the positive set rather than the predicted negative follows. di < dj , li ϵ P ; lj ϵ PN Similar to Joachim's method, the ranking SVM algorithm is also employed to learn a linear feature weight vector to rank the search results according to the user's content and location preferences.

5. PERSONALIZED RANKING FUNCTION

Ranking SVM is employed in our personalization approach to learn the user's preferences. For a given query, a set of content concepts and a set of location concepts are extracted from the search result as the document features. Since each document can be represented by a feature vector, it can be treated as a point in the feature space. Using click through data as the input, RSVM aims at finding a linear ranking function, which holds for as many document preference pairs as possible. In these experiments, an adaptive implementation, SVM light is used for the training.

It outputs a content weight vector (wc, q, u) and a location weight vector (wL, q, u) which best describes the user interests based on the user's content and location preferences extracted from the user click through, respectively. The two issues in the RSVM training process: How to extract the feature vectors for a document? How to combine the content and location weight vectors into one integrated weight vector?

5.1 Extracting Features for Training

Two feature vectors, namely, content feature vector (denoted by$\varphi C\ q, d\ \varphi L\ q\ , d\ )$ and location feature vector (denoted by$\varphi L\ q\ , d\ )$ are defined to present documents. The feature vectors are extracted by taking into account the concepts existing in a document and other related concepts in the ontology of the query. The similarity and parent-child relationships of the concepts in the extracted concept ontologies are also incorporated in the training based on the following four different types of relationships: (1) Similarity, (2) Ancestor, (3) Descendant, and (4) Sibling, in our ontologies.

5.2 Combining Weight Vectors

The content feature vector $\varphi\ C\ q, d$ together with the document preferences obtained from Joachims or SpyNB methods are served as input to RSVM training to obtain the content weight vector (wc, q, u) . The location weight vector (wL, q, u) is obtained similarly using the location feature vector $\varphi\ L\ q\ , d$ and the document preferences. The two weights vectors (wc, q, u) and (wL, q, u) represent the content and location user profiles for a user on a query q in our OMF user profiling method.

6. EXPERIMENTAL RESULTS

A metasearch engine is developed which comprises Google, MSN Search and Yahoo as the backend search engines to ensure a broad topical coverage of the search results. The metasearch engine collects clickthrough data from the users and performs personalized ranking of the search results based on the learnt profiles of the users. The users are invited to submit totally test queries to our metasearch engine. For each query submitted, the top search results are returned to the users. The topical categories of the test queries. Each of the 50 users is assigned 8 test queries randomly selected from the 15 different categories in chart to avoid any bias. The users are given the tasks to find results that are relevant to their interests. The clicked results are stored in the click through database and are

treated as positive samples in RSVM training. The clickthrough data, the extracted content concepts, and the extracted location concepts are used to create OMF profiles.

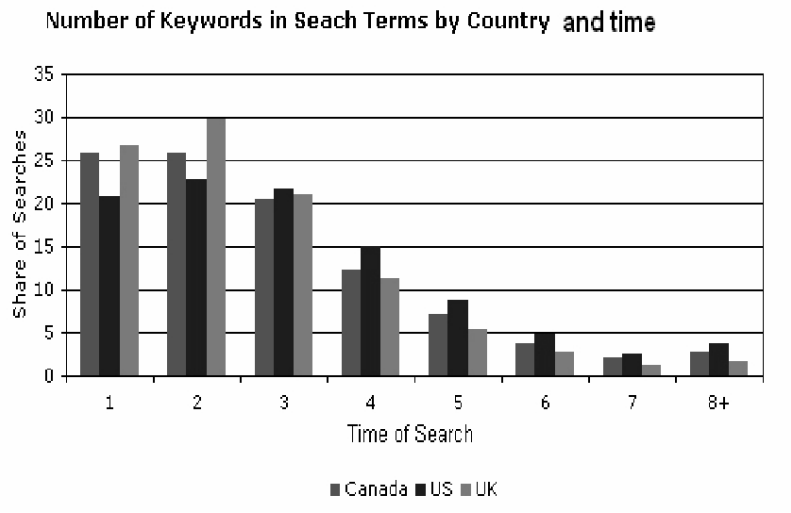

Fig3. *Statistics of click through data*

The threshold for content concept is set to 0.03. A small mining threshold is chosen because we want as many content concepts as possible that can be included in the user profiles. As discussed, the location concepts are prepared. They consist of 3 countries and 8 hours. Fig 3shows the statistics of the clickthrough data collected. In addition to the clickthrough data, the users are asked to perform relevance judgment on the top results for each query by filling in a score for each search result to reflect the relevance of the search result to the query.

Subject	Jan-08	Jan-09	% Change
1 word	20.96%	20.29%	**-3%**
2 words	24.91%	23.65%	**-5%**
3 words	22.03%	21.92%	**0%**
4 words	14.54%	14.89%	**2%**
5 words	8.20%	8.68%	**6%**
6 words	4.32%	4.65%	**8%**
7 words	2.23%	2.49%	**12%**

Table1. *Relevance Score*

The table1 relevance score indicates three levels of relevancy (.Zero, Positive, negative). Documents rated as 'Good' are considered relevant (positive samples), while those rated as 'Poor' are considered irrelevant (negative samples) to the user's needs. The documents rated as 'Fair' are treated as unlabeled. Documents rated as 'Good' (relevant documents) are used to compute the average relevant rank improvements (i.e., the difference between the average ranks of the relevant documents in the search results before and after personalization) and *top* N *precisions*, the two primary metrics for our evaluation.

6.1 Ontology Construction

The ontology is created for the concept as location ontology. Ontology is created to share the Understanding of structure of information among group of people. The subjects of user interest are

extracted from the WKB via user interaction. A tool called Ontology Learning Environment (OLE) is developed to assist users with such interaction. Regarding a topic, the interesting subjects consist of two sets: positive subjects are the concepts relevant to the information need, and negative subjects are the concepts resolving paradoxical or ambiguous interpretation of the information need. Thus, for a given topic, the OLE provides users with a set of candidates to identify positive and negative subjects. These candidate subjects are extracted from the WKB. Fig. 4 is a screen-shot of the OLE for the sample topic "Economic espionage." The subjects listed on the top-left panel of the OLE are the candidate subjects presented in hierarchical form. For each s ϵ S, the s and its ancestors are retrieved if the label of s contains any one of the query Terms in the given topic (e.g., "economic" and "espionage"). From these candidates, the user selects positive subjects for the topic. The user-selected positive subjects are presented on the top-right panel in hierarchical form. The candidate negative subjects are the descendants of the user-selected positive subjects. They are shown on the bottom-left panel. From these negative candidates, the user selects the negative subjects. These user-selected negative subjects are listed on the bottom right panel (e.g., "Political ethics" and "Student ethics"). Note that for the completion of the structure, some positive subjects (e.g., "Ethics," "Crime," "Commercial crimes," and "Competition Unfair") are also included on the bottom-right panel with the negative subjects. These positive subjects will not be included in the negative set. The remaining candidates, who are not fed, back as either positive or negative from the user, become the neutral subjects to the given topic.

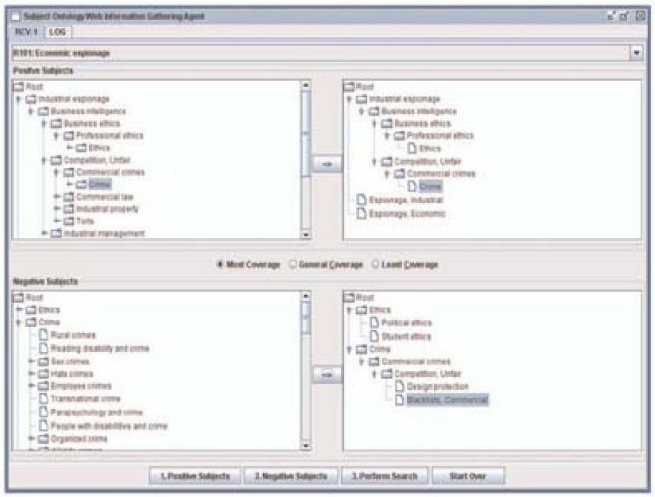

Fig4. *Ontology learning environment*

Ontology is then constructed for the given topic using these users fed back subjects. The structure of the ontology is based on the semantic relations linking these subjects in the WKB. The ontology contains three types of knowledge: Positive subjects, negative subjects, and neutral subjects.

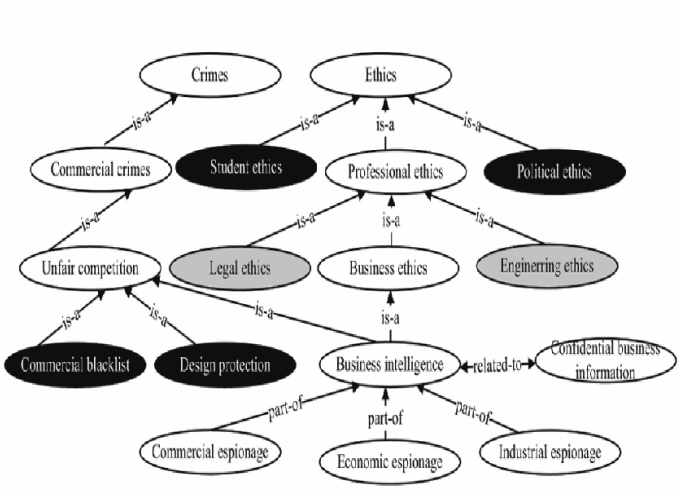

Fig5. *Ontology (partial) constructed for topic "Economic Espionage."*

Fig.5 illustrates the ontology (partially) constructed for the sample topic "Economic espionage," where the white nodes are positive, the dark nodes are negative, and the gray nodes are neutral subjects. The constructed ontology is personalized because the user selects positive and negative subjects for personal preferences and interests.

7. CONCLUSION

In this paper, an OMF personalization framework is proposed for automatically extracting and learning a user's content and location preferences based on the user's clickthrough. In the OMF framework, different methods are developed for extracting content and location concepts, which are maintained along with their relationships in the content and location ontologies. The notion of content and location entropies is introduced to measure the diversity of content and location information associated with a query and click content and location entropies to capture the breadth of the user's interests in these two types of information. Based on the weight vectors the personalization effectiveness is derived and showed with a case study that personalization effectiveness differs for different classes of users and queries. Experimental results confirmed that OMF can provide more accurate personalized results comparing to the existing methods. As for the future work, we plan to study the effectiveness of other kinds of concepts such as people Names and time for personalization. We will also investigate methods to exploit a user's content and location preference history to determine regular user patterns or behaviors for enhancing future search.

REFERENCES

[1] Michael Chau, Hsinchun Chen, A machine learning approach to web page filtering using content and structure analysis. Decision Support Systems,Elsevier, Vol 44, pp. 482-494, February 2008.

[2] Seikyung Jung, Jonathan L. Herlocker and Janet Webster, Click data as implicit relevance feedback in web search. Information Processing & Management, Elsevier,Vol 43, pp. 791-807, March 2007.

[3] Liu Shuchao, Li Yongchen, Wu Hongping, Research and Discussion of Web Data Mining. Manufacturing Automation, Vol 32, pp. 163-166, September 2010 (In Chinese).

[4] Du Yajun, Qiu Xiaoping, Xu Yang, Inquiry Intellectual Capacity into Chinese Search Engine. Application Research of Computers, Vol 4, pp. 29- 31, 35, April 2004 (In Chinese).

[5] Wu Yu, Status and Development of Chinese Search Engine. Modern Information, Vol 3, pp. 40-43, March 2003 (In Chinese).

[6] Chen Jihong, Qing Xiao, A Comparative Study of Four Search Engines. Information Science, Vol 21, pp. 1084-1087, October 2003 (In Chinese).

[7] Xu Jiakun, Studying by Comparison the Four Searching Engines in Common Use in the Research of Network Information Measurement. New Technology of Library and Information Service, Vol11, pp. 46-48, November 2004 (In Chinese).

[8] Huang Chen, Advantages and Disadvantages of the Ten Famous Chinese Search Engines. Modern Information, Vol 1, pp. 69-71, January 2006 (In Chinese).

[9] Fang Zhijian, Zhang Ruilin, Tong Xiaosu, Recently research and future development of search engine. Computer Engineering and Design, Vol 28,pp. 4038- 4041, August 2007 (In Chinese).

[10] Zhang Fan, Lin Jian, Research on Filtering Mechanism in Intelligent Search Engine. Library and Information, Vol 4, pp. 52-56, April 2007(In Chinese).

[11] Lai Yonghao, Xie Zanfu, Research on Anti-jamming Bad Web Filter Algorithm. Computer Engineering, Vol 33, pp. 98-99, November 2007 (In Chinese).

[12] Tan Hansong, Li Hong, Web Mining on Information Filtering. Computer Engineering and Applications, Vol 30, pp. 186-187, October 2003 (In Chinese).

[13] Liao Kaiji, Yi Cong, The Study of Web Business Information Extraction Based on Regular Expressions. Journal of Intelligence, Vol 29, pp 159- 62, May 2010 (In Chinese).

[14] Qin Hua, Su Yidan, Li Taoshen, A Data Cleaning Method Based on Genetic Algorithm and Neural Network. Computer Engineering and Applications, Vol 3, pp. 45-46, January 2004 (In Chinese).

[15] Wang Weiling, Liu Peiyu, Liu Kefei, A Feature Selection Algorithm for Web Documents Clustering. Computer Applications and Software, Vol 24, pp. 154-156, January 2007(In Chinese).

[16] Zhu Zhiguo, Deng Guishi, Analysis and research on Web usage mining. Application Research of Computers, Vol 25, pp. 29-32, 36, January 2008 (In Chinese).

[17] Zhu Zhiguo, Design of Architecture of Web Usage Pattern Mining System.

AUTHORS' BIOGRAPHY

Sindhupriya Pemmasani received the MCA degree in the year 2011 and pursuing M.Tech degree in Computer Science and Engineering from Krishnaveni Engineering College for Women.

P.Vemana received his MCA degree. He is currently working as an Asst Professor in Krishnaveni Engineering College for Women.