Web Information Gathering Using Personalized Ontology Model

Sahaja Kolla[1*], K.Ravi Kumar[2*]

1. *M.Tech (SE) Student, Dept of CSE, Kakinada Institute of Engg & Tech, Korangi, Dist: E.Godavari, AP, India*
2. *Asst.Professor, Dept of CSE, Kakinada Institute of Engg & Tech, Korangi, Dist: E.Godavari, AP, India*

Keywords:
Ontology, Information Gathering, User Profiles and Knowledge

Abstract: *Nowadays, how to gather useful and meaningful information from the Web has become challenging to all users because of the explosion in the amount of Web information. However, the mainstream of Web information gathering techniques has many drawbacks, as they are mostly keyword-based. It is argued that the performance of Web information gathering systems can be significantly improved if user background knowledge is discovered and a knowledge-based methodology is used. In this paper, a knowledge-based model is proposed for Web information gathering. The model uses a world knowledge base and user local instance repositories for user profile acquisition and the capture of user information needs. The knowledge-based model was successfully evaluated by comparing a manually implemented user concept model. The proposed knowledge-based model contributes to better designs of knowledge-based and personalized Web information gathering systems.*

Ontology's are mostly used to represent user profiles in personalized web information gathering as a model for Knowledge description and formalization. However, when representing user profiles, so many models had utilized only knowledge from either a user local information or global knowledge base. In this paper, an ontology model is proposed for representing user background knowledge for personalized web information gathering. This model constructs user personalized ontology's by extracting world knowledge from the system and discovering user background knowledge from user local instance repositories. By comparing it against benchmark models in web information gathering, the ontology model is evaluated.

1. INTRODUCTION

In recent decades, the amount of Web information has exploded rapidly. How to gather useful information from the Web has become a challenging issue to all Web users. Many information retrieval (IR) systems have been developed in an attempt to solve this problem, resulting in great achievements. However, there is still no complete solution to the challenge [11].The current Web information gathering systems cannot completely satisfy Web search users, because they are mostly based on keyword-matching mechanisms and suffer from the problems of information mismatching and overloading [42]. Information mismatching means valuable information is being missed in information gathering. This usually occurs whenone search topic has different syntactic representa-tions. For example, 'data mining' and 'knowledge discovery' refer to the same topic of discovering knowledge from raw data. However, by using keyword matching mechanisms, documents containing 'knowledge discovery' may be missed if using the query 'data mining' in the search. The other problem, information overloading, usually occurs when one query has different semantic meanings.

*** *Sahaja Kolla***
M.Tech (SE) Student, Dept of CSE, Kakinada Institute of Engg & Tech, Korangi, Dist: E.Godavari, AP, India

A common example is the query 'apple', which may mean apples (fruit), ori Mac (computer). By using the query 'apple' to describe the information need 'apple (fruit)', the search results may be mixed with useless information about 'iMac (computer)' [41,42]. From these examples, a hypothesis arises that if user information needs can be captured and interpreted, more useful and meaningful information can be gathered for users. Capturing user information needs via a given query is difficult. In most Web information gathering cases, users provide only short phrases in their queries to express information needs [61]. Also, Web users formulate queries differently because of different personal perspectives, expertise, and terminological habits and vocabularies. These differences cause difficulties in capturing user information needs. Thus, to capture user information needs effectively, understanding user background knowledge is necessary. For this purpose, user profiles are widely used in personalized Web information gathering systems [34]. These systems ap ply user background knowledge to information gathering. This mechanism was suggested by Yao [81] as knowledge retrieval.

In this paper, we introduce a knowledge-based personalized information gathering model, aiming at improving the performance of information gathering systems by utilizing user background knowledge. This knowledge-based model learns personalized ontologies for user profiles and applies user profiles to information gathering. Given a query, the user's background knowledge is discovered from a world knowledge base and the user's local instance repository. Based on these, a personalized ontology is constructed that simulates the user's concept model and captures the user information need. The semantic relations of is-a, part-of, and related-to are specified for the concepts in the constructed ontological user profile. The acquired user profile is then used by Web information gathering systems to gather useful and meaningful information for the user. The knowledge-based model was evaluated by being compared with a model that manually specified user background knowledge, and the evaluation result was promising and encouraging.

The proposed knowledge-based model contributes to better understanding of user information needs and user profile acquisition, as well as better design for personalized Web information gathering systems.

2. RELATED WORK

2.1 Semantic Concepts Extraction

Knowledge-based information gathering is based on the semantic concepts extracted from documents and queries. The similarity of documents to queries is determined by the matching level of their semantic concepts. Thus, concept representation and knowledge discovery are two typical issues and will be discussed in this section. Semantic concepts have various representations.

In some models, concepts are represented by controlled lexicons defined in terminological ontologies, thesauruses, or dictionaries. A typical example is the synsets in WordNet, a terminological ontology [15]. The models using WordNet for semantic concept representation include [6,17,22] and [33]. The lexiconbased representation defines the semantic concepts in terms and lexicons that are easily understood by users and easily utilized by computational systems. However, though the lexicon-based concept representation was reported to improve information gathering performance in some works [28,33,47], it was also reported as degrading performance in some other works [72,70].Another concept representation in Web information gathering systems is pattern-based representation, including [42,38,77,14,57]. In such representation, con-cepts can be discriminated from others only when the length of patterns representing concepts are adequately long. However, if the length is too long, the patterns extracted from Web documents would be of low frequency. As a result, they cannot substantially support the concept-based information gathering systems [77]. Many Web systems rely upon subject-based representation of semantic concepts for information gathering. Semantic concepts are represented by subjects that are defined in knowledge bases or taxonomies, including domain ontologies, digital library systems, and online categorization systems. Typical information gathering systems utilizing domain ontologies for concept representation include those developed by Lim et al. [10], by Navigli [11], and by Velardi et al. [12].

Also used for subject-based concept representation are the library systems, like Dewey Decimal Classification used by, Library of Congress Classification and Library of Congress Subject Headings by [16].The online categorizations are also widely used by many information gathering systems for concept representation, including the Yahoo! categorization used by [18] and Open Directory Project.

However, the semantic relations associated with the concepts in these existing systems are specified as only super-class and sub-class. They have inadequate details and poor specificity level. Thus, the specification of semantic relations for subject-based concept representation demands further development. Techniques used by Web information gathering systems to discover knowledge from text include text classification and Web mining. Text classification is the process of classifying an incoming stream of documents into categories by using the classifiers learned from training samples [46]. The performance of text classification relies upon the accuracy of these classifiers [37,80]. Existing techniques for learning classifiers include Rocchio [56], Naïve Bayes (NB) [54], Dempster-Shafer [58], Support Vector Machines (SVMs) [27], and the probabilistic approaches [50]. Treating the classifiers as semantic concepts, the process of learning classifiers is then a process of extracting semantic concepts to represent the categories. Text classification techniques are widely used in concept-based Web information gathering systems, like [7,18,48]. However, by using text classifi- cation techniques, the Web information gathering per formance largely relies on the accuracy of predefined categories [20]. Also, the 'cold start' problem occurs when there is an insufficient number of training samples available to learn classifiers. Web mining discovers knowledge from the content of Web documents, and attempts to understand the semantic meaning of Web data . Li andZhong represented semantic concepts by maximal patterns, sequential patterns, and closed sequential patterns, and extracted semantic concepts from Web documents. Association rule mining was also used by many systems for knowledge discovery from web documents, including . Text clustering techniques were used by to discover user interest for personalized Web information gathering. Some works, such as Dou et al. [14], used hybrid Web content mining techniques for concept extraction. However, as pointed out by Li and Zhong ,these existing Web content mining techniques have some limitations. One of these limitations is the in capability of specific semantic relation (e.g. is-a and part-of) specification for concepts. Therefore, the current concept extraction techniques need to be improved for better specific semantic relation specification, especially given the fact that the current Web is becoming the semantic Web [3].

2.2 Ontology Learning

The information gathering is for Global knowledge web information gathering is very default job. For example, learned personalized ontology's from the Open Directory Project to specify users' preferences and interests in web search. On the basis of the Dewey decimal classification, King et al. [7] developed IntelliOnto to improve performance in distributed web information retrieval. Wikipedia was used by Downey et al. [10] to help understand underlying user interests in queries. These works effectively discovered user background knowledge and limits to global knowledge bases.

To reduce this problem at learning personalization background knowledge to form user local knowledge ,construction of ontology based on keyword queries to description logics it employs natural language understanding techniques. Improved fuzzy expert knowledge[16] will help discover knowledge performance will high, if you combine the data mining and information retrieval techniques improve the performance present world knowledge .background knowledge; however, their performance was limited by the quality of the global knowledge bases.

2.3 WIG challenges

Last decade huge growth and adoption of the WWW have further exacerbated user need for efficient mechanisms for information and knowledge location, selection and retrieval.web covers large range of topics and spectrum of different communities[17] ,how to gather useful information and meaningful information form web. however ,it becomes challenging to web users this challenging issues is referred by many researchers as web information gathering.

The challenges are gathering information may possibly contain much useless and meaningless information, information mismatching and information overloading, useful information may missed out in the formation gathering[18] and effectiveness of WIG (web information gathering) is a difficult task for all web information gathering systems.

2.4. Concept Model

Web information gathering tasks are triggered by users were in need of some information and conducted an information gathering task, they usually fell into one of the following cases:

1. they knew nothing about that information;
2. they had tried but failed to infer that information from what they already knew;
3. they might know something about that information

but were not sure, so they needed confirmation.user information needs. From observations, when Profile is the tells about the needed information of users on user's interest topics, it can divided into three groups :interviewing ,semi-interviewing, and non-interviewing. User profiles have various representations, user profiles are represented by a previously prepared collection of data reflecting user interests, in many approaches collection of data refers to a set of terms that can be directly used to expand the queries submitted by users, many causes get poor interpretation of user interest to users, sometimes keyword match techniques because many terms are usually ambiguous.

Profiles can represented in the form of personalized ontologies, it can be represented by in the form of sub-taxmony of a predefined hierarchy of concepts[8]. The concepts existing in the these concepts .This kind of user profiles describes user interests explicitly. however ,clearly specifying user interests in ontologies is a difficult task especially for their semantic relations .profiles can also represented by training set of documents ,add used in text classification ,it consist of positive documents that contain user interest topics ,and negative documents that contain ambiguous or paradoxical topics.

3. USER PROFILE GAINING

When acquiring user profiles ,the content ,life cycle, and applications need to be considered ,content of profiles is the description if user interests and life cycle of user profiles refers to the period that the user profiles are valuable for web information gathering[10]. profiles can be classified into long term or short term. Profiles are extracted user interests from the collection of user desktop information such as text documents ,emails, and cached web pages.

Dictionaries and thesaurus are also common global knowledge bases used by information gathering systems for information need analysis, web systems have achieved remarkable .

Web knowledge bases nowadays are being used more and more frequently to analyze the semantic meaning[11] of user information needs.

4..ONTOLOGY CONSTRUCTION MODEL

Ontologies provide common understanding of topics for communication between systems and users and enable web–based knowledge processing ,sharing and reuse between applications, and it help to define and interpret the semantic meaning of web content, and enable intelligenent agents to gather web information for users in knowledge based web gathering .

Type of stored knowledge are divided into two types: domain ontologies and terminological ontologies, domain ontologies specify expert classified concepts and form the core knowledge in particular domains thus ,the content of domain ontologies needs to be updated regularly with updated knowledge .

Generic and terminologies store the lexical relations of concepts in natural languages. The knowledge specified in generic ontologies[12] is usually in large size and does not require regular updates.

Ontologies generally consist of set of concepts sometimes called classes ,a set of vocabularies(instances),semantic relations, and some inference and logic rules(axioms)[14] for general purpose or a particular domain. The semantic relations typically include hierarchical and non-hierarchical relations[15].

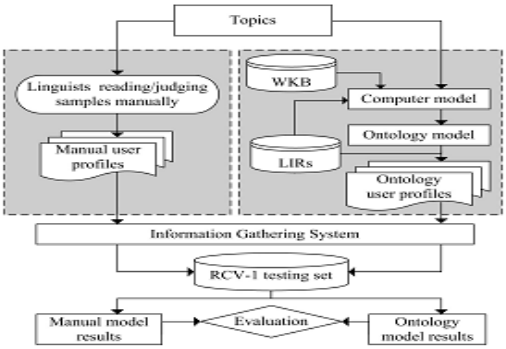

Figure 1 Experimental design .

5. CONCEPT-BASED WEB INFORMATION GATHERING

When web information gathering task starts form a user information need, if you need of some information and begins an information gathering task ,they usually fell into one of the following cases.

1. they knew nothing about that information
2. they had tried but failed to infer information from what they already knew and
3. might know something but were not sure ,so they needed to confirm.

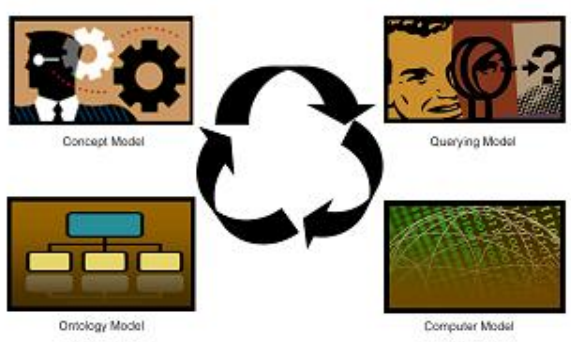

Figure 2.Concept-based Web Information Gathering .

6. CONCLUSION

In this paper, a knowledge-based model is proposed, aimed at discovering user background knowledge for personalized Web information gathering. The framework of knowledge-based information gathering consists of four models: user concept model, user querying model, computer model, and ontology model. Given a topic, the computer model uses a world knowledge base to learn an ontology for user concept model simulation. The ontology is then personalized by using the user's local instance repository. Aiming at describing user background knowledge more clearly, the semantic relations of is-a, part-of, and related-to are specified in the ontology model. The knowledge-based model was successfully evaluated in comparison with a manually implemented user concept model. The proposed knowledge-based model is a novel contribution to better understanding Web personalization using ontologies and user profiles, and to better designs of personalized Web information gathering systems.

REFERENCES:

1. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

2. G.E.P. Box, J.S. Hunter, and W.G. Hunter, Statistics For Experimenters. John Wiley & Sons, 2005.

3. C. Buckley and E.M. Voorhees, "Evaluating Evaluation Measure Stability," Proc. ACM SIGIR '00, pp. 33-40, 2000.

4. Z. Cai, D.S. McNamara, M. Louwerse, X. Hu, M. Rowe, and A.C Graesser, "NLS: A Non-Latent Similarity Algorithm," Proc. 26thAnn. Meeting of the Cognitive Science Soc. (CogSci '04), pp. 180-185,2004.

5. K.S. Lee, W.B. Croft, and J. Allan, "A Cluster-Based ResamplingMethod for Pseudo-Relevance Feedback," Proc. ACM SIGIR '08,pp. 235-242, 2008.

6. D.D. Lewis, Y. Yang, T.G. Rose, and F. Li, "RCV1: A NewBenchmark Collection for Text Categorization Research," J.Machine Learning Research, vol. 5, pp. 361-397, 2004.

7. Y. Li and N. Zhong, "Web Mining Model and Its Applications forInformation Gathering," Knowledge-Based Systems, vol. 17, pp. 207-217, 2004.

8. Y. Li and N. Zhong, "Mining Ontology for AutomaticallyAcquiring Web User Information Needs," IEEE Trans. Knowledgeand Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.

9. C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis,"Category Ranking for Personalized Search," Data and KnowledgeEng., vol. 60, no. 1, pp. 109-125, 2007.

10. S.E. Middleton, N.R. Shadbolt, and D.C. De Roure, "OntologicalUser Profiling in Recommender Systems," ACM Trans. InformationSystems (TOIS), vol. 22, no. 1, pp. 54-88, 2004.

11. G.A. Miller and F. Hristea, "WordNet Nouns: Classes andInstances," Computational Linguistics, vol. 32, no. 1, pp. 1-3, 2006.

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

13. L.M. Chan, Library of Congress Subject Headings: Principle andApplication. Libraries Unlimited, 2005.

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

15. R.M. Colomb, Information Spaces: The Architecture of Cyberspace.Springer, 2002.

16. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning toMap between Ontologies on the Semantic

Web," Proc. 11th Int'lConf. World Wide Web (WWW '02), pp. 662-673, 2002.

17. D. Dou, G. Frishkoff, J. Rong, R. Frank, A. Malony, and D. Tucker,"Development of Neuroelectromagnetic Ontologies(NEMO): AFramework for Mining Brainwave Ontologies," Proc. ACMSIGKDD ('07), pp. 270-279, 2007.

18. D. Downey, S. Dumais, D. Liebling, and E. Horvitz, "Under-standing the Relationship between Searchers' Queries andInformation Goals," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 449-458, 2008.

19. K.Suresh,R.MadanaMohana and Dr.A.RamaMohan Reddy "Improved FCM algorithm for Clustering on Web Usage Mining", IJCSI International Journal of Computer Sciencec Issues, Vol.8 Issue 1, January 2011,ISSN(Online):1694-0814. http://www.ijcsi.org/papers/IJCSI-8-1-42-45.pdf.