Personalized Web Information Collection Using Knowledge-based Ontologies

K.V. Narayana Rao, U. Jwalitha, K.D.N.V Rajesh
Adams Engineering College, Paloncha

Abstract -Ontologies are widely used to represent user profiles in personalized web information gathering. 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.

Keywords: Knowledge-based Information Gathering, Ontology, World Knowledge Base, Local Instance Repository, User Information.

I. 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 when one search topic has different syntactic represent a- discovery refer to the same topic of discovering knowledge from raw data. However, by using key word 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. A common example is the query 'apple', which may mean apples (fruit), or iMac (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)' [14,12]. 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 [11]. 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 [24]. These systems apply user background knowledge to information gathering. This mechanism was suggested by Yao [18] 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. The paper is organized the framework of the knowledge-based information gathering model. The implementation of the knowledge-based model is introduced.

II. RELATED WORK

Knowledge-based Information

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], it was also reported as degrading performance in some other works [17]. Another concept representation in Web information gathering systems is pattern-based representation, including [14]. In such representation, concepts 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 [19]. 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. [24], by Navigli [51], and by Velardi et al. [11]. Also used for subject-based concept representation are the library systems, like Dewey Decimal Classification used by [15], 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 Project1 used by [8,12]. 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.

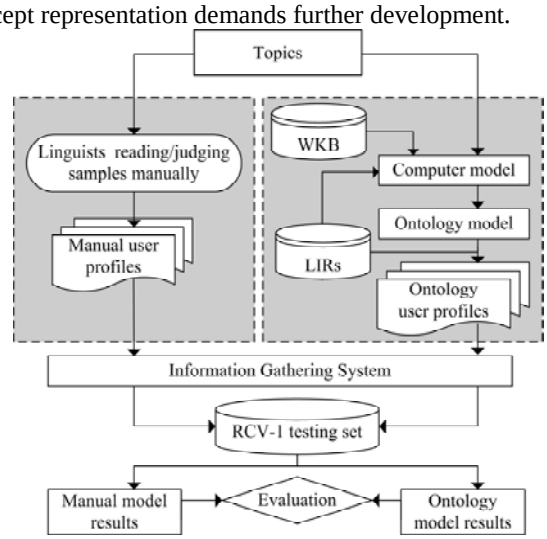

Fig1. Knowledge Based information system Architecture.

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 [19]. The performance of text classification relies upon the accuracy of these classifiers [23]. Existing techniques for learning classifiers include Rocchio [16], Naïve Bayes (NB) [14], Dempster-Shafer [18], Support Vector Machines (SVMs) [27], and the probabilistic approaches [10]. 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]. However, by using text classification techniques, the Web information gathering performance 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 [12]. Li and Zhong [16] 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 [20]. Text clustering techniques were used by [21] 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 [21], these existing Web content mining techniques have some limitations. One of these limitations is the incapability 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].

Ontology

Ontologies are an important technology in the semantic Web and Web information gathering systems. They provide a common understanding of topics for communication between systems and users, and enable Web-based knowledge processing, sharing, and reuse between applications [10]. Ontologies have been widely used by many groups to specify user background knowledge. Li and Zhong [4] used ontologies to describe the user conceptual level model: the so-called 'intelligent' part of the world knowledge model possessed by human beings. They [22] also used pattern recognition and association rule mining techniques to discover knowledge from Web content and learned ontologies for user profiles. Tran et al. [26] introduced an approach to translate keyword and reuse, concept extract, concept prune, and concept refine. The framework extends typical ontology engineering environments by using semi-

automatic ontology learning tools with human intervention, and constructs ontologies adopting the paradigm of balanced cooperative modelling. Typical ontologies learned by using manual mechanisms areWordNet [15] and its extensive models, such as Sensus [8] and HowNet [12]. The manual ontology learning mechanism is effective in terms of knowledge specification but expensive in terms of finance and computation. Automated ontology learning is then completed using the hierarchical collections of documents or thesauruses. One example is the so-called reference ontology used by [19]. This ontology was constructed based on the subject hierarchies and their associated Web pages in Yahoo!, Lycos, and Open Directory Project. King et al. [11] proposed the IntelliOnto, an ontology describing world knowledge by using a three-level taxonomy of subjects constructed on the basis of Dewey Decimal Classification. These learning methods increase the efficiency of ontology learning. However, the effectiveness of ontology learning is limited by the quality of the knowledge bases used in these methods. Many works tried to learn ontologies automatically without using knowledge bases. Web content mining techniques were used by Jiang and Tan [25] to discover knowledge from domain-specific text documents for ontology learning. Abulaish and Dey [1] proposed a framework to extract concepts from Web documents and construct ontologies with fuzzy descriptors. Jin et al. [26] attempted to integrate data mining and information retrieval techniques to further enhance ontology learning techniques. Doan et al. [13] proposed a model called GLUE and used machine learning techniques to extract similar concepts from different taxonomies. Dou et al. [14] proposed a framework to learn domain ontologies using pattern decomposition, clustering and classification, and association rule mining techniques. An ontology learning tool called OntoLearn was developed by Navigli et al. [11] in an attempt to discover semantic relations among the concepts from Web documents. These works have explored a new route to specify knowledge efficiently. The semantic association between concepts in ontologies can be discovered by computing the conceptual similarity (or distance) between them in the space of ontologies [24]. The node-based conceptual similarity methods measure the extent of information shared in common by the measured concept queries to the Description Logics conjunctive queries and to specify user background knowledge in ontologies. Gauch et al. [18] learned personalized ontologies for individual users in order to specify their preferences and interest. Cho and Richards [9] proposed to construct ontologies from user visited Web pages to improve Web document retrieval performance. Ontologies were used in these works to specify user background knowledge for personalized Web information gathering. Ontology learning is the process of constructing ontologies. Zhong and Hayazaki [14] introduced a two-phase ontology learning approach: conceptual relationship analysis and ontology prototype generation. Alternatively, Maedche [19] proposed an ontology learning framework.

III. PERSONALIZED ONTOLOGY MINING

Ontology mining is a process of discovering knowledge from the ontology backbone and the associated instances. A two dimensional method is introduced here for mining an ontology. Exhaustivity (exh for short) describes the semantic extent covered by a subject referring to a topic; and Specificity (spe for short) describes the semantic focus of a subject referring to a topic. The two dimensional method aims to analyze the semantic relations held by the subjects existing in the ontology referring to a topic. A subject in the ontology may be deemed highly exhaustive, although it may be not specific to the topic. In contrast, a subject may be highly specific, although it may deal with only a few aspects of the topic. A subject's exhaustively is affected by the number of subjects that are covered in its volume and the belief of these subjects to the topic: exh(s, T) =Xs02vol(s) bel(s0 , T) The semantic extent spreads if more subjects appear in its volume and more details these subjects hold. A subject with the positive exhaustively value makes the semantic meaning of the topic clearer, and a subject with the negative exhaustively value makes it more confusing. Exhaustively can be used to refine the process of expert knowledge extraction for a topic, e.g. the positive exhaustive subjects for the extraction of positive training set, and the negative exhaustive subjects for the negative training set. The specificity of a subject is affected by some factors. Firstly, the specificity increases if more instances refer to the subject, and if greater belief of these instances are to the topic. Secondly, the specificity decreases if a subject locates at a higher level in the taxonomy, since its description becomes more abstractive, e.g. from "Economic espionage" to "Business intelligence" in Fig. 1. Thirdly, a subject's semantic relations with its peers may impact the specificity. If a subject s is combined by a number of n subjects (each one holds the semantic relation partOf(si, s) with s, i = 1 . . . n), it holds only one nth of focus held by si, e.g. "Business intelligence" holds less focus than "Economic espionage".

IV. RESULT ANALYSIS

The TREC user profiles have weaknesses. Every document in the training sets was read and judged by the users. This ensured the accuracy of the judgments. However, the topic coverage of TREC profiles was limited. A user could afford to read only a small set of documents (54 on average in each topic). As a result, only a limited number of topics were covered by the documents. Hence, the TREC user profiles had good precision but relatively poor recall performance. Compared with the TREC model, the Ontology model had better recall but relatively weaker precision performance. The Ontology model discovered user background knowledge from user local instance repositories, rather than documents read and judged by users. Thus, the Ontology user profiles were not as precise as the TREC user profiles. However, the Ontology profiles had a broad topic coverage. The substantial coverage of possibly-related topics was gained from the use of the WKB and the large number of training documents (1,111 on average in each LIR). As a result, when taking into

account only precision results, the TREC model's MAP performance was better than that of the Ontology model. However, when considering recall results together, the Ontology model's F1 Measure results outperformed that of the TREC model, as shown in Table 1. Also, as shown on Fig. 8, when counting only top indexed results (with low recall values), the TREC model outperformed the Ontology model. When the recall values increased, the TREC model's performance dropped quickly, and was eventually outperformed by the Ontology model. The web model acquired user profiles from web documents. Web information covers a wide range of topics and serves a broad spectrum of communities [7]. Thus, the acquired user profiles had satisfactory topic coverage. However, using web documents for training sets has one severe drawback: web information has much noise and uncertainties. As a result, the web user profiles were satisfactory in terms of recall, but weak in terms of precision. Compared to the web data used by the web model, the LIRs used by the Ontology model were controlled and contained less uncertainties. Additionally, a large number of uncertainties was eliminated when user background knowledge was discovered. As a result, the user profiles acquired by the Ontology model performed better than the web model, as shown in Table 1.

	Macro-*F1* Measure			Micro-*F1* Measure		
Topic	TREC	Web	Onto	TREC	Web	Onto
R101	0.7333	0.6555	0.5978	0.6660	0.5982	0.5428
R102	0.7285	0.5588	0.5754	0.6712	0.5179	0.5327
R103	0.3600	0.3347	0.3859	0.3242	0.3059	0.3445
R104	0.6441	0.6162	0.6280	0.5851	0.5662	0.5786
R105	0.5548	0.5662	0.5782	0.5092	0.5163	0.5293
R106	0.2324	0.2433	0.2794	0.2223	0.2270	0.2586
R107	0.2297	0.2028	0.2057	0.2061	0.1866	0.1936
R108	0.1794	0.1520	0.1388	0.1676	0.1424	0.1295
R109	0.4508	0.6564	0.6659	0.4205	0.6026	0.6119
R110	0.2176	0.1560	0.2801	0.2019	0.1466	0.2568
R111	0.1082	0.0905	0.1267	0.1017	0.0863	0.1218
R112	0.1940	0.1745	0.1987	0.1800	0.1631	0.1813
R113	0.3152	0.2126	0.3519	0.2867	0.1975	0.3252
R114	0.4128	0.4247	0.4192	0.3732	0.3892	0.3840
R115	0.5063	0.5395	0.5079	0.4523	0.4831	0.4551
Avg.	0.3911	0.3722	0.3960	0.3579	0.3419	0.3630

The Category model specified only the knowledge with a relation of super-class and subclass. In contrast, the Ontology model moved beyond the Category model and had more comprehensive knowledge with is-a and part-of relations. Furthermore, specificity and exhaustively took into account subject localities, and performed knowledge discovery tasks in deeper technical level compared to the Category model. Thus, the Ontology model discovered user background knowledge more effectively than the Category model. As a result, the Ontology model outperformed the Category model in the experiments.

V. 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] Muhammad Abulaish and Lipika Dey. Information extraction and imprecise query answering from web documents. Web Intelligence and Agent Systems, 4(4):407–429, January 2006.

[2] G. Antoniou and F. van Harmelen. A Semantic Web Primer. The MIT Press, 2004.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic Web. Scientific American, 5:29–37, 2001.

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

[5] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 33–40, 2000.

[6] A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1):13–47, 2006.

[7] J. Chaffee and S. Gauch. Personal ontologies for Web navigation. In Proceedings of the ninth international conference on Information and knowledge management, pages 227–234, 2000.

[8] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter. Using ODP metadata to personalize search. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 178– 185. ACM Press, 2005.

[9] W. C. Cho and D. Richards. Ontology construction and concept reuse with formal concept analysis for improved web document retrieval. Web Intelligence and Agent Systems, 5(1):109– 126, January 2007.

[10] K.-S. Choi, C.-H. Lee, and P.-K. Rhee. Document ontology based personalized filtering system (poster session). In MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia, pages 362–364, New York, NY, USA, 2000. ACM Press.

[11] R. M. Colomb. Information Spaces: The Architecture of Cyberspace. Springer, 2002.

[12] K. Curran, C. Murphy, and S. Annesley. Web intelligence in information retrieval. In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, pages 409 – 412, 2003.

[13] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the semantic web. The International Journal on Very Large Data Bases, 12(4):303–319, 2003.

[14] D. Dou, G. Frishkoff, J. Rong, R. Frank, A. Malony, and D. Tucker. Development of neuroelectromagnetic ontologies(NEMO): a framework for mining brainwave ontologies. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 270– 279, 2007.

[15] C. Fellbaum, editor. WordNet: An Electronic Lexical Database. ISBN: 0-262-06197-X. MIT Press, Cambridge, MA, 1998.

[16] E. Frank and G.W. Paynter. Predicting library of congress classifications from library of congress subject headings. Journal of the American Society for Information Science and Technology, 55(3):214–227, 2004.

[17] A. Gangemi, N. Guarino, and A. Oltramari. Conceptual analysis of lexical taxonomies: the case of wordnet top-level. In FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems, pages 285–296, New York, NY, USA, 2001. ACM Press.

[18] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems, 1(3-4):219–234, 2003.

[19] S. Gauch, J. M. Madrid, and S. Induri. Keyconcept: A conceptual search engine. Technical report, EECS Department, University of Kansas, 2004.

[20] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web structure for classifying and describing Web pages. In WWW '02: Proceedings of the 11th international conference on World Wide Web, pages 562–569, New York, NY, USA, 2002. ACM Press.

[21] D. Godoy and A. Amandi. Modeling user interests by conceptual clustering. Information Systems, 31(4):247–265, 2006.

[22] C. Hung, S. Wermter, and P. Smith. Hybrid neural document clustering using guided self-organization and wordnet. Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications], 19(2):68–77, 2004.

[23] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the web. SIGIR Forum, 32(1):5–17, 1998.

[24] J. J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference Research on Computational Linguistics (ROCLING X), 1997, Taiwan, pages 19–33, Taiwan, 1997.

[25] X. Jiang and A-H. Tan. Mining ontological knowledge from domain-specific text documents. In Proceedings of the Fifth IEEE International Conference on Data Mining, pages 665– 668, 2005.

[26] 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. In Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pages 193–202, 2007.

[27] T. Joachims. Transductive inference for text classification usingsupport vector machines. In Ivan Bratko and Saso Dzeroski, editors, Proceedings of ICML-99, 16th International