A New Web-Mining Technique for Web Information Gathering Based on Personalized Ontology (PO) Model

Andra Vijaya Krishna [#1], A.Mary Sowjanya [*2]

[1#] Integrated M.Tech Scholar
Department of Software Engineering ,
Andhra University College Of Engineering,
Visakhapatnam,AP,India.
[2*] Assistant Professor
Department of Computer Science & Systems Engineering,
Andhra University College Of Engineering,
Visakhapatnam,AP,India.

Abstract

World Wide Web is an interlinked collection of billions of documents formatted using HTML. Ironically the very size of this collection has become an obstacle for information retrieval. The user has to shift through scores of pages to come upon the information he/she desires. Web crawlers are the heart of search engines. Web crawlers continuously keep on crawling the web and find any new web pages that have been added to the web, pages that have been removed from the web. Due to growing and dynamic nature of the web, it has become a challenge to traverse all URLs in the web documents and to handle these URLs. A focused crawler is an agent that targets a particular topic and visits and gathers only relevant web pages. In this paper, a personalized ontology model is proposed for knowledge representation and reasoning over user profiles. This model learns ontological user profiles from both a world knowledge base and user local instance repositories. The ontology model is evaluated by comparing it against benchmark models in web information gathering. The comparison results show that this ontology model is successful.

Keywords: Ontology, Personalization, semantic relations, user profiles, Local Instance Repository (LIR), Web Information gathering, world knowledge.

1. Introduction

World Wide Web has rapidly increased its users from the past decades. In the past decades the information available on World Wide Web has exploded rapidly Web information is available in a great range of topics and different categories. How to collect the required information is a challenging task. Search engines usually return more than 1,500 results per query, yet out of the top twenty results, only one half turn out to be relevant to the user. One reason for this is that Web queries are in general very short and give an incomplete specification of individual users'information needs. User Profiling explores ways of incorporating users'interests into the search process to improve the results.

The user profiles are structured and populated by watching over a user's shoulder'while he is surfing. No explicit feedback is necessary. The profiles are shown to converge and to reflect the actual interests quite well. Web user profiles are widely used by web information systems for user modeling and personalization. User profiles reflect the interests of users [1]. User profiles are used in Web information gathering to capture user information needs in order to get personalized web information for users. When acquiring user profiles, the content, life cycle and applications are taken into consideration since user interests are approximate

and unambiguous it is suggested it can be represented by ontologies [2]. On the last decades, the amount of web-based information available has increased dramatically. How to gather useful information from the web has become a challenging issue for users. 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.

Global analysis uses existing global knowledge bases for user background knowledge representation. Commonly used knowledge bases include generic ontologies (e.g., WordNet [3]), thesauruses (e.g., digital libraries), and online knowledge bases (e.g., online categorizations and Wikipedia). The global analysis techniques produce effective performance for user background knowledge extraction. However, global analysis is limited by the quality of the used knowledge base. For example, WordNet was reported as helpful in capturing user interest in some areas but useless for others [4].Local analysis investigates user local information or observes user behavior in user profiles. For example, Li and Zhong [5] discovered taxonomical patterns from the users' local text documents to learn ontologies for user profiles. Some groups [6], [7] learned personalized ontologies adaptively from user's browsing history.Here, first domain is selected and the seed url is entered and search is done on the basis of local and global databases.

In this paper, an ontology model to evaluate this hypothesis is proposed. This model simulates users' concept models by using personalized ontologies and attempts to improve web information gathering performance by using ontological user profiles. The world knowledge and a user's local instance repository (LIR) are used in the proposed model. World knowledge is commonsense knowledge acquired by people from experience and education [8]; an LIR is a user's personal collection of information items. From a world knowledge base, we construct personalized ontologies by adopting user feedback on interesting knowledge. A multidimensional ontology mining method, Specificity and exhaustively, is also introduced in the proposed model for analyzing concepts specified in ontologies. The users' LIRs are then used to discover background knowledge and to populate the personalized ontologies. The proposed ontology model is evaluated by comparison against some benchmark models through experiments using a large standard data set. The evaluation results show that the proposed ontology model is successful. The research contributes to knowledge engineering, and has the potential to improve the design of personalized web information gathering systems. The contributions are original and increasingly significant, considering the rapid explosion of web information and the growing accessibility of online documents.

2. Background Theory

2.1 Ontology Learning

Global knowledge bases were used by many existing models to learn ontologies for web information gathering. For example, Gauch et al. [7] and Sieg et al. [6] learned personalized ontologies 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.[4] developed IntelliOnto to improve performance in distributed web information retrieval. Wikipedia was used by Downey et al. 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. Learning personalized ontologies, many works mined user background knowledge from user local information.Li and Zhong [5] used pattern recognition and association rule mining techniques to discover knowledge from user local documents for ontology construction. Translated keyword queries to Description Logics' conjunctive queries and used ontologies to represent user background knowledge. Zhong [5] proposed a domain ontology learning approach that employed various data mining and natural-language understanding techniques. Navigli et al. [9] developed OntoLearn to discover semantic concepts and relations from web documents.

Web content mining techniques were used to discover semantic knowledge from domain-specific text documents for ontology learning. Finally, captured user information needs at the sentence level rather than the document level, and represented user profiles by the Conceptual Ontological Graph. The use of data mining

techniques in these models leads to more user background knowledge being discovered. However, the knowledge discovered in these works contained noise and uncertainties. Additionally, ontologies were used in many works to improve the performance of knowledge discovery.Uptil, keyword based search, concept based search is available but URL searching is text based online searching is different than available search.

2.2 Ontology Construction

The term ontology can be defined in many different ways. Ontology as an explicit specification of a set of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them [6]. As implied by the general definition, an ontology is domain dependent and it is designed to be shared and reusable. Usually, ontologies are defined to consist of abstract concepts and relationships (or properties) only. In some rare cases, ontologies are defined to also include instances of concepts and relationships [5, 1]. For this purpose, it is defined as ontology to be a set of concepts C and relationships R. The relationships in R can be either taxonomic or non-taxonomic. For example, **Fig.1** depicts a simple University ontology consisting of a set of concepts
C univ = {Person, Faculty, Staff, Student, Department, Project, Course}, and a set of relationshipsR
univ={Department_Of(Person,Department),Member_Of(Person,Project),Instructor_Of(Course,Person),Superclass_Of(Faculty,Person)Superclass_Of(Staff,Person), Superclass_Of(Student, Person)}.
Superclass_Of represents the taxonomic relationship while the rest are not. With this definition, the instances of ontology refer to the instances of its concepts and relationships. If each
concept instance exists in the form of a Web page, a relationship instance will then exist in the form of a Web page pair. This view has been adopted in most the Web classification research. In practical terms, developing ontology includes:

a) Defining classes in the ontology,

b) Arranging the classes in a taxonomic (subclass–super class) hierarchy,

c) Defining slots and describing allowed values for these slots,

d) Filling in the values for slots for instances.Then create a knowledge base by defining individual instances of these classes filling in specific slot value information and additional slot restrictions.

Then create a knowledge base by defining individual in-stances of these classes filling in specific slot value information and additional slot restrictions.

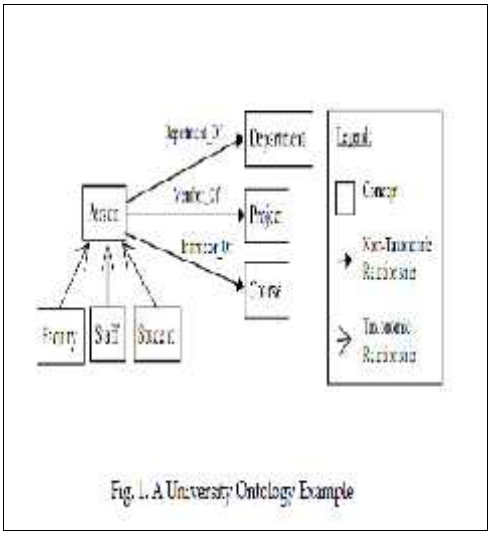

Fig. 1. A University Ontology Example

2.3 Techniques of Generating User Profile

When acquiring user profiles, the content, life cycle and applications are taken into consideration since user interests are approximate and unambiguous it is suggested it can be represented by ontologies [10]. User profile acquisition techniques can be categorized into three groups: 1) Interviewing 2) Non-interviewing 3) Semi-interviewing. The interviewing technique is done manually by asking ques-tions, interviewing and user trained datasets. Users read training sets and then assign positive or negative feedback based on user‘s interests.e.g. TREC model is used to acquire training set manually. The topic coverage of

TREC profiles was limited. But it provides more accuracy. Non-interviewing is based on observation at user‘s behavior, user‘s interests and preferences are described by a set of weighted subjects learned from the user‘s browsing history. These subjects are specified with the semantic relations of super class and subclass in ontology. When an OBIWAN agent receives the search results for a given topic, it filters and reranks the results based on their semantic similarity with the subjects.

The similar documents are awarded and reranked higher on the result list. E.g. Category model. Semi interviewing in which user profiles are acquired from the web by employing a web search engine. The feature terms referred to the interesting concepts of the topic. The noisy terms referred to the paradoxical or ambiguous concepts. E.g. Web mining model. 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. There was no negative training set generated by this model. In ontology model, semi interviewing is used.

3. Personalized Ontology Construction

Personalized ontologies are a conceptualization model that formally describes and specifies user background knowledge. From observations in daily life, we found that web users might have different expectations for the same search query. For example, for the domain ‚Health‘ a person may demand different information of various medical aids. Sometimes even the same user may have different expectations for the same search query if applied in a different situation. Based on this observation, an assumption is formed that web users have a personal concept model for their information needs. Therefore, domain wise searching of urls is suggested.

3.1 World Knowledge Representation

World knowledge is important for information gathering. World knowledge is commonsense knowledge possessed by people and acquired through their experience and education. Also, “world knowledge is necessary for lexical and referential disambiguation, including establishing co reference relations and resolving ellipsis as well as for establishing and maintaining connectivity of the discourse and adherence of the text to the text producer’s goal and plans.” In this proposed model, user background knowledge is extracted from web.

3.2 Ontology Creation

The subjects of user interest in the form of URLs are extracted from the web via user interaction. A ontology model is developed to assist users with such interaction. Regarding a topic, the interesting URLs consist of two sets: positive urls are the concepts relevant to the information need, and negative urls are the concepts resolving paradoxical or ambiguous interpretation of the information need. Thus, for a given topic, it provides users a set of positive urls. We are concentrating on focused crawler which search for the relevant web pages based on the URL we give. Actually it forms a hierarchy of links.

The crawler on the particular web page for a particular keyword, which we give as, input. It will search for the link on that seed URL and after that switch to that link and find another link on that web page but it should match with the keyword, it will do like that until it reach the limit that we set. But it may be possible that it will not found the number of links that we set before. Then it shows that the web page is not having any further link for that particular keyword. While fetching the links the crawler also make sure that it should fetch only the unique links, means that it should not revisit the same link again and again. Finally, when we finished with the links, we will give one txt file as input and run the pattern matching algorithm. Pattern matching is used for syntax analysis. When we compare pattern matching with regular expressions then we will find that patterns are more powerful, but slower in matching. A pattern is a character string. All keywords can be written in both the upper and lower cases. It is used to extract hidden information from not-structured or semi-structured data. This aspect is fundamental because much of the web information is semi-

structured due to the nested structure of HTML code, much of the web information is linked, and much of the web information is redundant. It should not include images, tags, and buttons. The extracted content should be stored in some file. But it should not include any HTML tags. The constructed ontology is personalized because the user selects positive subjects for personal preferences and interests as by selecting do-main names. This model is developed for four domains as– **1.General 2.Health 3.Education 4.Entertainment**.

It also allows entering 4 URL addresses at a time which pro-vides parallel processing for finding relative URL's.It also avoids time delay since providing parallel processing of input. It also counts every time how many URLs are searched at once, type of protocol, Hash code, web page content ,time of download. It also maintains local database which is used when user is offline and world knowledge base is searched when user is online.

3.3 Algorithm Details

Here we have combined both semantic and KMP searching algorithm for retrieving webpage content. Semantic search technique is used to retrieve a WebPages by finding relations between the texts given. KMP algorithm is used to find a partial match for given input. Knuth-Morris-Pratt algorithm is for pattern recognition. Semantic search is used to identify specificity.

3.3.1 Multidimensional Ontology Mining

Ontology mining discovers interesting and on-topic knowledge from the concepts, semantic relations, and instances in ontology. In this section, a 2D ontology mining method is introduced: Specificity and Exhaustively. Specificity (denoted spe) describes a subject's focus on a given topic. Exhaustively (denoted exh) restricts a subject's semantic space dealing with the topic. This method aims to investigate the subjects and the strength of their associations in ontology. Subject's specificity has two focuses: 1) on the referring-to concepts (called semantic specificity), and 2) on the given topic (called topic specificity), is done on the encrypted data. The output of the processing is deobfuscated by the privacy manager to reveal the correct result. However, the privacy manager provides only limited features in that it does not guarantee protection once the data are being disclosed. In [4], the authors present a layered architecture for addressing the end-to-end trust management and accountability problem in federated systems. The authors' focus is very different from ours, in that they mainly leverage trust relationships for account-ability, along with authentication and anomaly detection. Further, their solution requires third-party services to complete the monitoring and focuses on lower level monitoring of system resources.

input : a personalized ontology $O(T) = (tax^S, rel)$; a coefficient θ between (0,1).
output: $spe_\alpha(s)$ applied to specificity.

1 set $k = 1$, get the set of leaves S_0 from tax^S, for $(s_0 \in S_0)$ assign $spe_\alpha(s_0) = k$;
2 get S' which is the set of leaves in case we remove the nodes S_0 and the related edges from tax^S;
3 **if** $(S' == \emptyset)$ **then** return;//the terminal condition;
4 **foreach** $s' \subset S'$ **do**
5 **if** $(isA(s') == \emptyset)$ **then** $spe_\alpha^1(s') = k$;
6 **else** $spe_\alpha^1(s') = \theta \times min\{spe_\alpha(s) | s \in isA(s')\}$;
7 **if** $(partOf(s') == \emptyset)$ **then** $spe_\alpha^2(s') = k$;
8 **else** $spe_\alpha^2(s') = \frac{\sum_{s \in partOf(s')} spe_\alpha(s)}{|partOf(s')|}$;
9 $spe_\alpha(s') = min\{spe_\alpha^1(s'), spe_\alpha^2(s')\}$;
10 **end**
11 $k = k \times \theta$, $S_0 = S_0 \cup S'$, go to step 2.

Algorithm 1. Analyzing Semantic Relations for Specificity

3.3.2 Topic Specificity

The topic specificity of a subject is investigated, based on the user background knowledge discovered from user local information.

3.3.2.1 User Local Instance Repository

User background knowledge can be discovered from user local information collections,

such as a user's stored documents, browsed web pages, and composed/received emails [5]. The ontology constructed in Section 3 has only subject labels and semantic relations specified. In this section, we populate the ontology with the instances generated from user local information collections. Such a collection the user's local instance repository (LIR). The topic specificity of a subject is evaluated based on the instance-topic strength of its citing URLS. With respect to the absolute specificity, the topic specificity can also be called relative specificity and denoted by

$$spe_r(s, \mathcal{T}, \mathcal{LIR}) = \sum_{i \in s^{-1}(s)} str(i, \mathcal{T}).$$

A subject's specificity has two focuses: semantic specificity and topic specificity. Therefore, the final specificity of a subject is a composition of them and calculated by

$$spe(s, \mathcal{T}) = spe_n(s) \times spe_r(s, \mathcal{T}, \mathcal{LIR}).$$

The lower bound subjects in the ontology would receive greater specificity values, as well as those cited by more positive instances.

3.3.2 KMP (KNUTH MORRIS PRATT)

i. Knuth-Morris-Pratt method takes advantage of the partial-match.
ii. Identify the bad URL in a website.
iii. No. of character present in a web page.
iv. Identify type of protocol used for the web page.
v. Retrieve the web pages we apply pattern recognition over text.
vi. Pattern symbolizes check text only.
vii. Check how much text is available on web page.

4. Architecture of Ontology Model

The proposed ontology model aims to discover user back-ground knowledge and learns personalized ontologies to rep-resent user profiles. Fig. 2 illustrates the architecture of the ontology model. A personalized ontology is constructed, according to a given topic. Two knowledge resources, the global world knowledge base and the user's local instance repository, are utilized by the model. The world knowledge base provides the taxonomic structure for the personalized ontology. The user background knowledge is discovered from the user local instance repository. Against the given topic, the specificity and exhaustivity of subjects are investigated for user background knowledge discovery.

From the diagram, we can hypothesize that user background knowledge can be better discovered and represented if we can integrate global and local analysis within a hybrid model. The knowledge formalized in a global knowledge base will con-strain the background knowledge discovery from the user local information. Such a personalized ontology model should produce a superior representation of user profiles for web in-formation gathering.

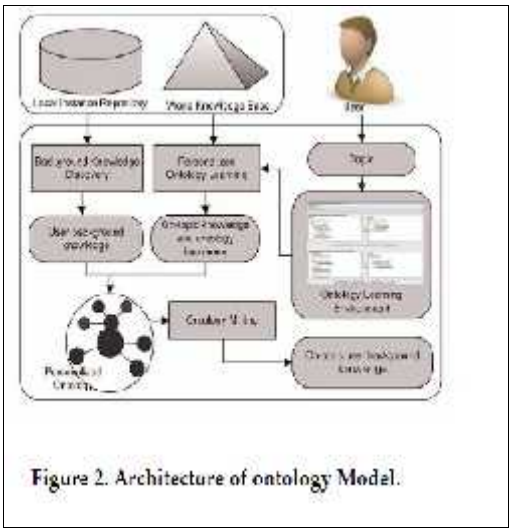

Figure 2. Architecture of ontology Model.

4.1 Comparison of various Models

Sr N.	Trec Model	Category Model	Web Model	Ontology Model
1)	Manual user profile acquiring methods	Non-interviewing user profiles acquiring techniques .	The typical semi – interviewing	Semi-automatic method and automatic method.
2)	Positive documents sup-port(c)=1/| D+|; Negative documents sup-port(c)=0;	Weighted positive subjects in the ontology with super-class and subclass relations .No negative training set.	The positive and negative subjects were identified by users manually.	Positive documents (D+) and negative documents (D-)
3)	Topic coverage of TREC profiles was limited.	Topic coverage is depending upon categorical classification.	User profiles were satisfactory.	Contained less uncertainties. As a result, the user profiles acquired by the Ontology model is better.
4)	Good precision of data.	Good precision of data.	Weak in terms of precision.	Moderate in precision.

5. Conclusion

Focused crawler is developed to extract only the relevant web pages of interested topic from the Internet. Semantic search technique is used to retrieve web pages from search engine and KMP algorithm is used to find a webpage content. Here multidimensional mining, parallel processing is supported. Speed and Query Processing time is high, better Efficiency, good Accuracy, and less Time Delay. The proposed ontology model provides a solution to emphasizing global and local knowledge in a single computational model.

6. References

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

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

[3] G. M. Voorhees and Y. Hou, ,Vector Expansion in a LargeCollection,? Proc. First Text Retrieval Conf., pp. 343-351.

[4] 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.

[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] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, ,Ontology-Based Interpretation of Keywords for Semantic Search,? Proc. Sixth Int'l Semantic Web and Second Asian Semantic Web Conf. (ISWC '07/ASWC '07), pp. 523-536, 2007.

[7] 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.

[8] L.A. Zadeh, "Web Intelligence and World Knowledge—The Concept of Web IQ (WIQ)," Proc. IEEE Ann. Meeting of the North American Fuzzy Information Soc. (NAFIPS '04), vol. 1, pp. 1-3, 2004.

[9] R 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.

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

7. About the Authors

Andra Vijaya Krishna is currently pursuing his 5 Years Integrated M.Tech in Computer Science and Systems Engineering at Andhra University College of Engineering, Visakhapatnam. His area of interests includes Web Mining.

A.Mary Sowjanya is currently working as an Assistant Professor in Computer Science and Systems Engineering department, Andhra University College of Engineering, Visakhapatnam. She is now registered for Ph.D. (CSSE) in Andhra University. Her research interests include Data Mining.