ISSN No. 0976-5697

Volume 3, No. 4, July- August 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Ontology Mining for Personalized Web Information Gathering

Rupa R*
Department of CSE
Sri Sivani College of Engineering
Srikakulam, India
rupa.reyyi@gmail.com

M M M K Varma
Department of CSE
Sri Sivani College of Engineering
Srikakulam, India
varmassce@gmail.com

*Abstract**:* It is well accepted that ontology is useful for personalized Web information gathering. However, it is challenging to use semantic relations of “kind-of”, “part-of”, and “related-to” and synthesize commonsense and expert knowledge in a single computational model. In this paper, a personalized ontology model is proposed attempting to answer this challenge. A two-dimensional (Exhaustively and Specificity) method is also presented to quantitatively analyze these semantic relations in a single framework. The proposals are successfully evaluated by applying the model to a Web information gathering system. The model is a significant contribution to personalized ontology engineering and concept-based Web information gathering in Web Intelligence.

Keywords: kind-of, mining, ontology and clustering

I. INTRODUCTION

Ontology is a formal description and specification of knowledge. It provides a common understanding of top-ics to be communicated between users and systems [1,2]. By using an ontology, information systems are expected to be able to understand the semantic meaning of words and phrases, and be able to compare information items by con-cepts instead of keywords [3]. Ontology is deemed by the Web Intelligence community as one of the most useful tech-niques for Web information gathering.Over the last decade, many attempts have been suggested to learn ontology in order to describe and specify the knowl-edge possessed by humans. Li & Zhong [4] proposed to discover the backbone of an ontology based on the patterns found in documents. Gauch et al. [3] proposed to learn personalized ontology based on the online portals. King et al. [5] proposed to learn ontology based on the Dewey Decimal Classification (DDC)1. However, these existing works only specify “super-” and “sub-class” relations in the ontology, and do not extend beyond the ontology learning framework proposed by Maedche [6]. Maedche’s ontol-ogy learning framework consists of four phases: Import, Extract, Prune, and Refine. The Maedche’s framework, however, has a pitfall of relying on the manpower of on-tology engineers heavily, e.g. for an incoming lexical entry, the engineer needs to manually determine either assigning it to an existing concept or defining a new concept for it. Consequently, these ontology methods are either incompre-hensive or expensive in knowledge acquisition.Web users possess a concept model in the process of in-formation gathering. Usually, users can easily determine if a Web page interests them or not while they read through the content. The rationale behind this is that users implic-itly possess a concept model based on their knowledge, al-though they may not be able to express it [7].

There ex-ists a potential that by describing and specifying this con-cept model, the semantic meaning of a user’s information need can be well interpreted. In this paper, a personal-ized ontology model is proposed, which extract the com-monsense knowledge possessed by the user in her con-cept model and the expert knowledge revising the concept model. The model synthesizes these two kinds of knowl-edge and formally specifies the semantic relations of “kind-of”, “part-of”, and “related-to” in a single computational model, instead of simple “super-” and “sub-class” in the ex-isting models [7,8,9]. In this paper, a two dimensional method, Specificity and Exhaustivity, is also presented to analyze these semantic relations in order to discover knowl-edge from the learnt personalized ontology and to use the ontology for personalized Web information gathering. The proposed model is evaluated by assessing its applications to a system that gathers information from a large corpus.

The model is a significant contribution to personalized ontology engineering and concept-based personalized Web informa-tion gathering in Web Intelligence.The paper is organized as follows. Section 2 is the prob-lem statement. Section 3 introduces the personalized on-tology learning method attempting to formally ontologize a user’s concept model. Section 4 presents the two dimensional method for ontology mining. Section 5 describes the related user profiling for personalized Web information gathering, and Section 6 discusses the evaluation. Finally, Section 7 presents the related work, and Section 8 makes the conclusions

II. BACKGROUND AND RELATED WORK

The personalized Web information gathering is a difficult task. A great challenge is that the semantic meaning of a user’s information need (called a topic in this paper) is difficult to interpret. One example is “Economic espionage”, which is a topic generated by the linguists in TREC2. It comes with a description of “What is being done to counter economic espionage internationally?” and a narrative of “Documents which identify economic espionage cases and provide action(s) taken to reprimand offenders or terminate their behavior are relevant. Economic espionage would encompass commercial, technical, industrial or corporate types of espionage. Documents about military or political espionage would be irrelevant". A concept model for the topic may be manually constructed, as illustrated in Fig. 1, consisting of various relevant or non-relevant subjects, ac-cording to these linguist generated specifications. However, it is hard for general users to specify such adequate description and narrative. Even this manually generated concept model could still be incomplete, because some important subjects may be missed out, and some semantic relations between the subjects may be overlooked, e.g. the semantic relation existing between "technical espionage" and "Indus-trial espionage" in Fig. 1.

The personalized Web information gathering is a challenge in Web Intelligence. The research work presented in this paper attempts to answer the challenge by proposing a personalized ontology learning and mining framework. The framework consists of five phases: (i) Building a taxonomic world knowledge base; (ii) Constructing the personalized ontology backbone by interacting with a user; (iii) Extracting expert knowledge to revise the ontology automatically; (iv) Mining the ontology by analyzing the semantic relations; and (v) Generating the personalized user profile

III. PROBLEM FORMULATION

The personalized ontology can describe different con-cept models for different users, although they may have the same topic. In order to do so, we argue that two kinds of knowledge are required: world knowledge covering large number of topics so that the user's individual information need can best match, and expert knowledge revising the concept model. World knowledge is the commonsense knowledge possessed by humans and "the kind of knowledge that humans acquire through experience and education" [10,11]. Expert knowledge is the kind of knowledge classified by the people who hold expertise in that domain. The difficulty in world knowledge extraction is the topic cover-age and semantic relation specification, whereas in expert knowledge extraction is the efficiency, since by traditional means expert knowledge is extracted by experts reading a set of documents manually. In this section, we are going to propose a method to extract the world knowledge and expert knowledge automatically.

A. World Knowledge Representation:

Taxonomic world knowledge base with great coverage of topics is superior of backbone learning for an ontology. The Library of Congress Subject Headings[3] (LCSH) classification is a system developed for organizing large volume of information stored in a library. It comprises a thesaurus of subject headings exhaustively covering a large number of topics in the world (contains 299,000 records according to the retrospective of 1986-2006). The LCSH system specifies the semantic relations existing in the subject headings, and facilitates the user's perspectives in accessing the information items in a library catalogue. Based on the LCSH system, a taxonomic world knowledge base can be constructed by forming each subject heading a class node and using the specified semantic relations as the links between the nodes.

The taxonomic knowledge base is formalized as follows.

Definition 1. Let Onto BASE be a taxonomic ontology base. An ontology base is formally defined as a 2-tuple

Onto BASE :=< S; R >, where

S is a set of subjects S := fs$_1$; s$_2$; ; s$_m$g; R is a set of relations R := fr$_1$; r$_2$; ; r$_n$g.

Definition 2. A subject s 2 S is formalized as a 3-tuple s :=< label; instance Set; >, where label is a label assigned by linguists to a subject s in the LCSH system. The label of s is denoted by label(s); instance Set is a set of objects associated to a subject s, in which each element specifies a semantic meaning referring by s and is called an instance (see Definition 5 for more details); is a signature mapping (: s ! 2^s) that defines a set of relevant subjects to a given s.

Kind Of is a directed relation in which one subject is in different form of another subject. The property of kind Of is Transitivity and Asymmetry. Transitivity means if s$_1$ is a kind of s$_2$ and s$_2$ is a kind of s$_3$, then s$_1$ is a kind of s$_3$ as well. Asymmetry means if s$_1$ is a kind of s$_2$ and s$_1$ 6= s$_2$, s$_2$ may not be a kind of s$_1$ necessarily. One example is that "Business ethics" is a Kind Of "Professional ethic", and "Professional ethic" is a Kind Of "Ethics". Then "Business ethics" is also a Kind Of "Ethics" as well. However, these relations can not be inverse.

Part Of is a directed relation used to describe the relationship held by a compound subject class and its component classes, e.g. subject s$_1$ forms a part of s$_2$. The part Of relationship also holds the properties of transitivity and asymmetry. If s$_1$ is a part of s$_2$ and s$_2$ is a part of s$_3$, then s$_1$ is also a part of s$_3$. If s$_1$ is a part of s$_2$ and s$_1$ 6= s$_2$, s$_2$ is definitely not a part of s$_1$. One example in the knowledge based is that "Economic espionage" is a part Of of "Business intelligence". The latter can not be a part Of the former.

Related To is a non-taxonomic relation describing the relationship held by two subjects that overlap in their semantic spaces. Related To holds the property of symmetry. If s$_1$ is related to s$_2$, s$_2$ is also related to s$_1$. One example in the knowledge base is "Business intelligence" and "Confidential business information".

A personalized ontology facilitating a user's concept model needs to be dynamically constructed in response to the change of information need. For this purpose, a tool called ontology learning environment interacting with the user is developed to help study a specific information need. The tool analyzes a specific topic, retrieves the possible relevant subjects from the knowledge base and presents them to the user. The user interacts with the tool and identifies the positive and negative (ambiguous) subjects according to the topic and the possessed concept model. The subject based personalized ontology is then built based on the user feed-back and the taxonomic knowledge base.

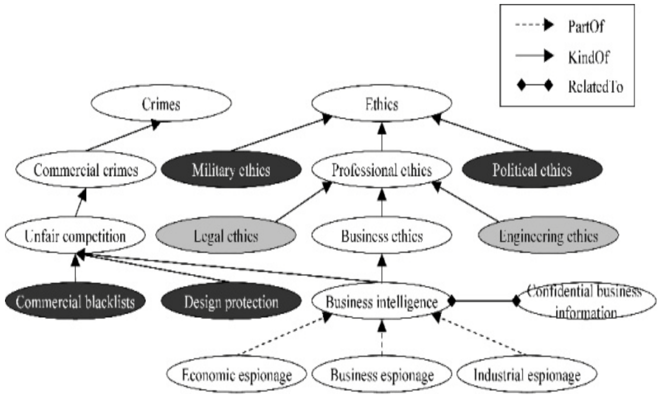

Figure 1. A Constructed Ontology

Fig. 1 shows an incomplete ontology constructed for the topic "Economic espionage", where the white nodes are the positive subjects, the dark gray are the negative, and the light gray are the unlabelled subjects. The unlabelled subjects are those in the volume of a positive subject but not being identified by the user as either positive or negative. The semantic relations existing between the subjects are addressed by different type of lines. The personalized ontology is formalized by the following definition.

Definition 4. The structure of an ontology that formally describes and specifies topic T is a 5-tuple O(T) := fS; R; tax^S; rel; A^{O}g, where

S is a set of subjects and S S. S has three subsets, where S$^+$ S is a set of positive subjects to T , S S is a set of negative subject to T , and S S is a set of unlabelled subjects to T ;

R is a set of relations and R R;

tax^S: tax^S S S is a taxonomic backbone of the ontology, which consists of two directed relations kindOf and partOf;

rel is a function defining non-taxonomic relations; A^O is a set of rules mined from O.

Given a pair of subjects (s$_1$; s$_2$), its dom(s$_1$; s$_2$) refers to their least common ancestor subject in tax^S. Given a subject s, its vol(s) refers to the union of all subjects in its volume. For partOf(tax^S) = (s$_1$; s$_2$) one may also write partOf(s$_1$; s$_2$), which means that s$_1$ is a part of s$_2$. For kindOf(tax^S) = (s$_1$; s$_2$) one may also write kindOf(s$_1$; s$_2$), which means that s$_1$ is a kind of s$_2$.

IV. EXPERIMENTAL EVALUATION

A user profile is the descriptions of the concept model possessed by the user [7]. In terms of Web information gathering, a user profile is the semantic interpretation of a topic based on the user possessed concept model. The subjectbased personalized ontology provides a basis for the user profile generating. The user profile is represented by a set of training documents in this paper, instead of a set of keywords or patterns by traditional means [3, 7]. Training sets are commonly used in Web data mining and text classification to represent knowledge [16]. A training set usually consists of a set of positive documents, a set of negative documents, and sometimes a set of unlabelled documents. Traditionally, the experts are needed to read a set of text documents and provide feedbacks of either positiveness or negativeness of each document according to the given topic. This technique is expensive because of manual effort involved. In this paper, a training set is generated to represent the user profile by using the proposed personalized ontology model .The proposed model is evaluated by assessing the success of its application to a Web information gathering system. In response to a given topic, the user profiles (training sets) are generated by the proposed model and the state of-the-art baselines. The profiles are input into a common system and used to train the system for information gathering. The performance of the system is determined by the quality of input training sets, where the information gathering method remains the same. By comparing the gathering results, the proposed model can then be evaluated quantitatively.

The experiment design is as follows. The Web information gathering system is implemented based on Li & Zhong's model (see [7] for technical details), including the basic text processing (e.g. stopword removal, word stemming and grouping). For generating the training sets, three models are implemented: TREC model The training sets are manually generated by the TREC linguists who read each document and mark it either positiveness or negativeness according to a topic [13]. These training sets reflect a user's concept model perfectly, and may be deemed as the "perfect" sets;Web model The training sets are automatically generated from the Web (see [16] for technical details). The model analyzes a given topic and identifies the relevant subjects, then uses the subjects to gather a set of Web documents by using a selected Web search engine (Google is chosen for the experiments as it has become the most popular search engine nowadays4). The model then measures the certainty of each document supporting/against the topic and assigns a float type of positive (or negative) judgment to the document. These documents then become the input training set to the Web information gathering system; Ontology model the training sets are generated as described in Section 5, by using the personalized ontology model proposed in this paper. A large volume (138MB) of information stored in the catalogue of a library5 is used, which contains 448,590 documents and 162,751 unique terms. The Reuters Corpus Volume 1 (RCV1) is used as the testbed, which is the official testbed used in TREC-11 2002 and an archive of 806,791 documents. TREC-11 has topics designed by linguists and associated with the training sets and testing sets. These topics (R101-115) are used in the experiments. The performances achieved by the Web information gathering system by applying the three models are compared and analyzed quantitatively. Two schemes are applied in the evaluation: the precision averages at 11 standard recall levels [18] and F1 Measure [5]. The former is used by TREC and computes each recall-precision point by:

$$\frac{\sum_{i=1}^{N} Precision\lambda}{N}$$

Where λ= {0:0; 0:1; 0:2; : : : ; 1:0} and N denotes the number of topics. Fig. 4 illustrates the recall-precision average results of the three models .The perfect TRFC model slightly outperforms others before reaching recall cut off 0.3, and then the Ontology model becomes the best since that on. This may indicate that the perfect TREC training sets are more precise than others, but does not cover as much relevant semantic space as the Ontology model. As a result, the Ontology model's precision catches it up while the recall value increases. The other evaluation scheme, F1 Measure, is calculated by

$$F_1=\frac{2* precison* recall}{precision+recall}$$

Precision and recall are evenly weighted in F1 Measure. The macro-F1 averages each topic's precision and recall values then calculates the F1 Measure, whereas the micro- F1 calculates the F1 Measure for each returned result for a topic and then averages the F1 values. The greater F1 values indicate the better performance. The detailed F1 Measure results are presented in Table. 1. In average, the Ontology model performs best. The highlighted rows are the topics that the Ontology model outperforms the perfect TREC model. This is because the TREC model employs the manpower of linguists to read every single document in the training set, which is perfect but expensive. As a result, the number of documents included in a TREC training set is limited (about 70 documents per topic in average), and some semantic meanings contained by the topic are not fully covered by the TREC training set. In contrast, the knowledge in the Ontology model is extracted from the LCSH and a large volume of expert classified information in library catalogue. The broad semantic coverage is the Ontology model's strength. As a result, the Ontology model has about 1730 documents per topic in average covering much broader semantic extent than the TREC training set. Based on the experiments, the proposed ontology learning and mining model is evaluated and its success is confirmed

Table 1: Result Analysis

	Macro-F1 Measure			Micro-F1 Measuer		
Topic	TREC	Web	Onto	TREC	Web	Onto
R101	0.7333	0.6555	0.5978	0.666	0.5982	0.5428
R102	0.7285	0.5588	0.5754	0.6712	0.5179	0.5327
R103	0.36	0.3347	0.3859	0.3242	0.3059	0.3445
R104	0.6441	0.6162	0.628	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	0.227	0.2586
R107	0.2297	0.2028	0.2057	0.2061	0.1866	0.1936
R108	0.1794	0.152	0.1388	0.1676	0.1424	0.1295
R109	0.4508	0.6564	0.6659	0.4205	0.6026	0.6119
R110	0.2176	0.156	0.2801	0.2019	0.1466	0.2568
R111	0.1082	0.0905	0.1267	0.1017	0.0863	0.1218
R112	0.194	0.1745	0.1987	0.18	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	0.384
R115	0.5063	0.5395	0.5079	0.4523	0.4831	0.4551

The graphical representation is as follows

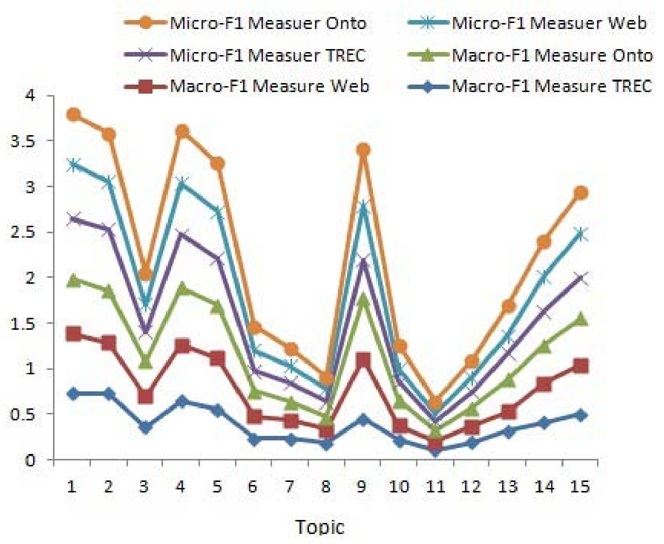

Figure 2: result analysis.

V. CONCLUSION

In this paper, a personalized ontology model is proposed aiming to synthesize world knowledge and expert knowledge for specific topics. The model extracts world knowledge from the LCSH system and discovers expert knowledge from a large volume of specified information in the library catalogue. The proposed model attempts to facilitate the user possessed concept model and to generate the personalized user profile for Web information gathering. It is a challenge to use semantic relations of "kind-of", part-of", and "related-to" in a single computational model. During literature review, we did not find any mathematic model that can well formalize these three relations together. In this paper, the proposed ontology model is an attempt to specify these semantic relations in a single framework. A two-dimensional method (Exhaustively and Specificity) is also presented in the paper to quantitatively analyze these three semantic relations. The proposals are successfully evaluated by comparing knowledge extracted by the personalized ontology model, against knowledge generated manually by linguists. The proposed model is a significant contribution to personalized ontology engineering and to concept based Web information gathering in Web Intelligence.

VI. REFERENCES

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

[2.] K. Curran, C. Murphy, and S. Annesley. Web intelligence in information retrieval. In Proc. of WI' 03, pages 409 – 412,2003.

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

[4.] J. D. King, Y. Li, X. Tao, and R. Nayak. Mining World Knowledge for Analysis of Search Engine Content. Web Intelligence and Agent Systems, 5(3):233–253, 2007.

[5.] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In Proc. of the 18th intl. ACM SIGIR conf. on Res. and development in inf. retr., pages 246–254.ACM Press, 1995.

[6.] Y. Li and N. Zhong. Web Mining Model and its Applications for Information Gathering. Knowledge-Based Systems,17:207–217, 2004.

[7.] Y. Li and N. Zhong. Mining Ontology for Automatically Acquiring Web User Information Needs. IEEE Transactions on Knowledge and Data Engineering, 18(4):554–568, 2006.

[8.] H. Liu and P. Singh. ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, 22(4):211–226,2004. Kluwer Academic Publishers.

[9.] J. Liu. New Challenges in the World Wide Wisdom Web (W4) Research. Lecture Notes in Computer Science, 2871:1–6, Jan 2003.

[10.] S. E. Middleton, N. R. Shadbolt, and D. C. D. Rour. Ontological user profiling in recommender systems. ACM Trans.Inf. Syst., 22(1):54–88, 2004.

[11.] R. Navigli, P. Velardi, and A. Gangemi. Ontology learning and its application to automated terminology translation. Intelligent Systems, IEEE, 18:22–31, 2003.

[12.] S. E. Robertson and I. Soboroff. The TREC 2001 filtering track report. In Text REtrieval Conference, 2001.

[13.] S. Staab and S. R., editors. Handbook on Ontologies.Springer-Verlag Berlin Heidelberg, 2004.

[14.] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In Proc. of the 13th intl. conf. on World Wide Web, pages 675–684, USA, 2004.

[15.] X. Tao, Y. Li, N. Zhong, and R. Nayak. Automatic Acquiring Training Sets for Web Information Gathering. In Proc.of the IEEE/WIC/ACM Intl. Conf. onWeb Intelligence, pages 532–535, HK, China, 2006.

[16.] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In Proc. of RIAO 2004, pages 380–389, France,2004.

[17.] L. Zadeh. Web intelligence and world knowledge - the concept of Web IQ (WIQ). In Processing of NAFIPS ’04., volume 1, pages 1–3, 27-30 June 2004.

[18.] N. Zhong. Representation and construction of ontologies for Web intelligence. International Journal of Foundation of Computer Science, 13(4):555–570, 2002.

[19.] N. Zhong and N. Hayazaki. Roles of ontologies for web intelligence.In Proceedings of Foundations of Intelligent Systems: 13th International Symposium, ISMIS 2002,, volume 2366, page 55, Lyon, France, June 27-29 2002.