Research Inventy: International Journal Of Engineering And Science
Issn: 2278-4721, Vol.2, Issue 4 (February 2013), Pp 42-47
Www.Researchinventy.Com

Study of Personalized Ontology Model for Web Information Gathering

[1] Mr.G.S.Deokate, [2]Prof.Mrs.V.M.Deshmukh
[1] Final Year Master of Engineering, PRMIT & R, Badnera (M.S) India
[2]Associate Professor, Department of CSE, PRMIT &R Badnera (M.S.) India

Abstract **-** *Ontology as model for knowledge description and formalization is used to represent user profile in personalized web information gathering. While representing user profiles many models used a global knowledge bases or user local information for representing user profiles. In this paper we study a personalized ontology model for knowledge representation and reasoning over user profiles. World knowledge base and local instance repositories are both are used in this model.*

Keywords *– Local Instance Repository, Ontology, Personalization, Semantic Relations, User Profiles, Web Information Gathering*

I. INTRODUCTION

Today is the world of internet. The amount of the web-base information available on the internet has increased significantly. But gathering the useful information from the internet has become the most challenging job today's scenario. People are interested in the relevant and interested information from the web. The web information gathering systems before this satisfy the user requirements by capturing their information needs. For this reason user profiles are created for user background knowledge description. The user profiles represent the concepts models possessed by user while gathering the web information. A concept model is generated from user background knowledge and possessed implicitly by user. But many ontologists have observed that when user read a document they can easily determined whether or not it is of their interest or relevance to them .If the user concept model can be simulated, and then a better representation of the user profile can be build. To Simulate use concepts model, ontologies are utilized in personalized web information gathering which are called ontological user profiles or personalized ontologies [1] ,[2],[3].In Global analysis, global knowledge bases are used for user background knowledge representation. Local analysis use local user information. Global analysis is limited by quality of knowledge base whereas local analysis is not sufficient for capturing user knowledge. If we integrate global and local analysis within a hybrid model the global knowledge will be constrain the background knowledge discovery form the user local information. Such an ontology model will give the better representation of user profiles. [4]

II. LITERATURE REVIEW

A. Ontology Learning

Ontologies are means of knowledge sharing and reuse. They are semantic containers. The term 'Ontology' has various definitions in various texts, domains and applications. Many existing knowledge bases are used by many models to learn ontologies.Gauch et al. [1] and Sieg et al. [5] learned personalized ontologies from the Open Directory Project to specify users' preferences and interests in web search. King developed *IntelliOnto* based on the basis of the Dewey decimal classification. The Dewey Decimal Classification (DDC) system is a general knowledge organization system that is continuously revised to keep pace with knowledge. The DDC is used around the world in 138 countries; over sixty of these countries also use Dewey to organize their national bibliographies. Over the lifetime of the system, the DDC has been translated into more than thirty languages [6]. Doweney et al. [7] used Wikipedia which helps in understanding user interests in queries. The above work discovered user background knowledge but the performance is limited by quality of the global knowledge base.Much work has been done for discovering user background knowledge from user local information.Pattern reorganization and association rule mining technique to discover knowledge from user local information is used by Li and Zhong [3]. A domain ontology learning approach was proposed by Zhong [3] that uses various data mining and natural language understanding techniques to discover knowledge from user local documents for ontology construction. Semantic relations and concepts are discovered by Navigli et al. [8] for which he developed a system called *Ontolearn. OntoLearn* system is an infrastructure for automated ontology learning from domain text. It is the only system, as far as we know, that uses natural language processing and machine learning techniques. Jiang and Tan [9] use content mining techniques to find semantic knowledge from domain-specific text documents for ontology learning. Much of user background knowledge is discovered using these data mining technique but d In many work ontologies are used for getting better performance in knowledge discovery. Lau et al. [10] in 2009 construct concept maps based on the posts on online discussion forums using a fuzzy domain ontology extraction algorithm. Doan developed a model called GLUE and used machine learning technique to find similar concepts in different ontologies. For given two ontologies, for each concept in one ontology, GLUE finds the most similar concept in the other ontology. GLUE can work with all of them. Another key feature of GLUE is that it uses multiple learning strategies, each of which exploits well a different type of information either in the data instances or in the taxonomic structure of the ontologies. These works explores more efficiently. [11].

B. User Profiles

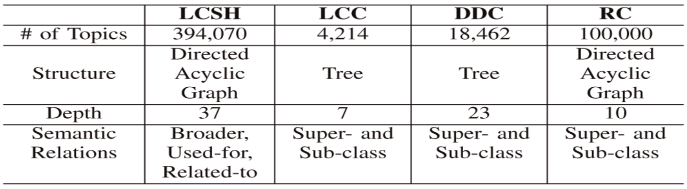
In the web information gathering, user profiles were used to understand the semantic meanings of queries and capture user Information needs. User profiles are used for user modeling and personalization. It is used to reflect the interests of user. Li and Zhong defined user profiles as the interesting topics of a user's information need. The user profiles are categorized into two diagrams: the data diagram and which are acquired by analyzing a database or a set of transaction whereas the information diagram user profiles acquired by using manually such as questionnaires and interviews or automatic techniques such as information retrieval and machine learning. User profiles are categorized into three groups: interviewing, semi-interviewing, and non-interviewing. [1], [3], [12], [13].

III. CONSTRUCTION OF PERSONALIZED ONTOLOGY

Personalized ontologies describe and specify user background knowledge forms a conceptualization model. We know that, web users might have different expectations for the same search query. For example, for the topic "New York", business travelers may have demand for different information from leisure travelers. Same user may have different expectation from same query if applied in the different situation. A user may become a business traveler when planning for a business trip, or a leisure traveler when planning for a family holiday. From this observation an assumption is formed that web users have a personal concepts model for their information needs, a user's concept model may change according to different information needs. [14]

C. World Knowledge Representation

For the information gathering the world knowledge is very important. World knowledge is commonsense knowledge possessed by people and acquired by through the experience and education. User background knowledge is extracted from a world knowledge base encoded from the Library of Congress Subject Heading (LCSH).The Library of congress subject Heading (LCSH) is ideal for world knowledge base. The LCSH system is a thesaurus developed for organizing and retrieving information from a large volume of library collections. LCSH has undergone continuous revising and enriching. The LCSH system is better than other world knowledge taxonomies used. Table 1 shows a comparison of the LCSH with Library of Congress Classification (LCC) used by Frank and Paynter [16], the Dewey Decimal Classification (DDC) used by Wang and Lee [17], and the reference categorization (RC) developed by Gauch et al. [1] using online categorizations.



	LCSH	LCC	DDC	RC
# of Topics	394,070	4,214	18,462	100,000
Structure	Directed Acyclic Graph	Tree	Tree	Directed Acyclic Graph
Depth	37	7	23	10
Semantic Relations	Broader, Used-for, Related-to	Super- and Sub-class	Super- and Sub-class	Super- and Sub-class

Table1 – Comparison of World Taxonomies [14]

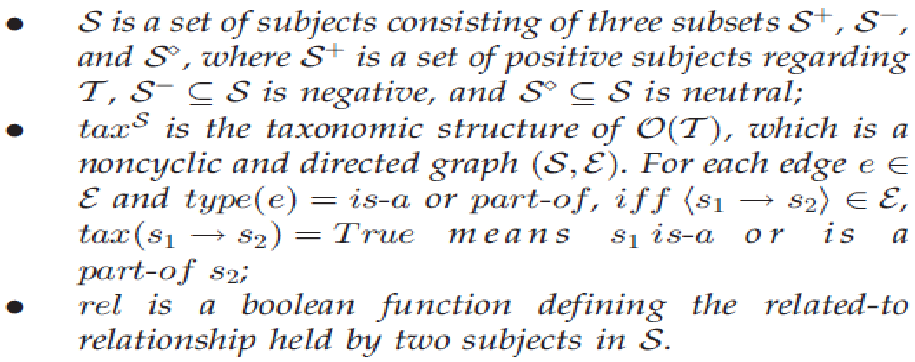
As shown in table1 LCSH has more topics, more specific structure and more semantic relations. [14], [15].

D. Ontology Construction

User interested subjects are extracted from the *WKB* via user interaction. Ontology Learning Environment (OLE) tool is developed to assists users with such interaction. Related to the topic, the interesting subjects consist of two sets; positive subjects and negative subjects. The subjects which are relevant to the information need are positive subjects and the subjects which resolve ambiguous interpretation of information need are negative subjects. The OLE provides users with a set of candidates to identity positive and negative subjects. These candidate subjects are extracted from the *WKB*. The ontology contains three types of subject candidates: positive, negative and neutral. The candidates which are not feedback as positive or negative are treated as neutral subjects. The ontology is formalized for a given topic as follows. The structure of an ontology that describes and specifies topic τ is a graph consisting of a set of subject's nodes. The structure can be formalized as a 3-tuple

$\mathcal{O}(\tau) :=< S, tax^{s}, rel >$.

Where



- $\mathcal{S}$ *is a set of subjects consisting of three subsets* $\mathcal{S}^{+}$, $\mathcal{S}^{-}$, *and* $\mathcal{S}^{\diamond}$, *where* $\mathcal{S}^{+}$ *is a set of positive subjects regarding* $\mathcal{T}$, $\mathcal{S}^{-} \subseteq \mathcal{S}$ *is negative, and* $\mathcal{S}^{\diamond} \subseteq \mathcal{S}$ *is neutral;*
- $tax^{\mathcal{S}}$ *is the taxonomic structure of* $\mathcal{O}(\mathcal{T})$, *which is a noncyclic and directed graph* $(\mathcal{S}, \mathcal{E})$. *For each edge* $e \in \mathcal{E}$ *and* $type(e) = is\text{-}a$ *or* $part\text{-}of$, iff $\langle s_1 \longrightarrow s_2 \rangle \in \mathcal{E}$, $tax(s_1 \longrightarrow s_2) = True$ *means* s_1 *is-a or is a part-of* s_2*;*
- *rel is a boolean function defining the related-to relationship held by two subjects in* $\mathcal{S}$.

User selects positive and negative subject for their interests and preferences hence constructed ontology is personalized. [14]

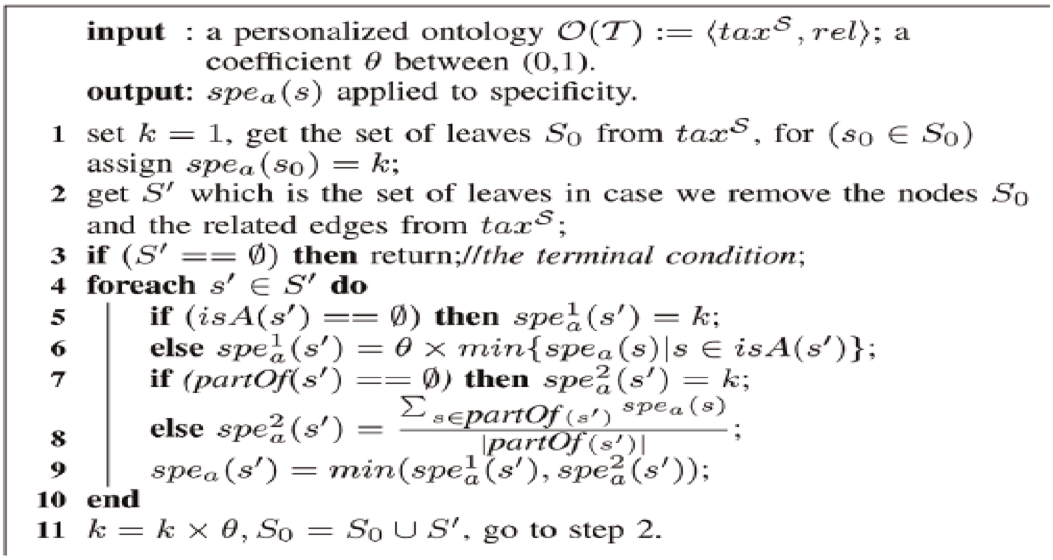
IV. Multidimensional Ontology Mining

Using ontology mining we can discover interesting and on-topic knowledge from the concepts, semantic relations and instances in ontology. Here we discuss 2D ontology mining method: specificity and exhaustivity. Subject's focus on a given topic is described by Specificity and subject's semantic space dealing with the topic is restricted by exhaustivity. Using this method we can investigate subject and the strength of their association in ontology. The subject's specificity has two focuses which are semantic specificity and topic specificity. [14]

A. Semantic Specificity

It is investigated based on the structure of $\mathcal{O}(\tau)$ inherited from the world knowledge base. The lower bound subjects have a stronger focus because it has fewer concepts in its space. Hence, the semantic specificity of a lower bound subjects is greater than that of an upper bound subjects. It is measured based on the hierarchical semantic relations (is-a and part-of) held by a subjects and its neighbors. The subjects have a fixed locality on the tax^{s} of $\mathcal{O}(\tau)$. It is also called as absolute specificity and denoted by $spe_{a}(s)$. The determination of a subject's $spe_{a}(s)$ described in algorithm1 [14]. The $isA(s')$ and $partOf(s')$ are two functions in the algorithm satisfying $isA(s') \cap partOf(s') = \emptyset$.

The $isA(s')$ returns a set of subjects $s \in tax^{s}$ that satisfy $tax(s \rightarrow s') = True$ and $type(s \rightarrow s') = is - a$. The $partOf(s')$ returns a set of subjects $s \in tax^{s}$ that satisfy *tax(s*→s`) =*True* and *type(s*→*s*`) = *part-Of*. The algorithm terminates eventually because tax^s is a directed acyclic graph



```
input  : a personalized ontology $\mathcal{O}(\mathcal{T}) := \langle tax^S, rel \rangle$; a
         coefficient $\theta$ between (0,1).
output: $spe_a(s)$ applied to specificity.
1  set $k = 1$, get the set of leaves $S_0$ from $tax^S$, for $(s_0 \in S_0)$
   assign $spe_a(s_0) = k$;
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 $s' \in S'$ do
5      if $(isA(s') == \emptyset)$ then $spe_a^1(s') = k$;
6      else $spe_a^1(s') = \theta \times min\{spe_a(s) | s \in isA(s')\}$;
7      if $(partOf(s') == \emptyset)$ then $spe_a^2(s') = k$;
8      else $spe_a^2(s') = \frac{\sum_{s \in partOf(s')} spe_a(s)}{|partOf(s')|}$;
9      $spe_a(s') = min(spe_a^1(s'), spe_a^2(s'))$;
10 end
11 $k = k \times \theta$, $S_0 = S_0 \cup S'$, go to step 2.
```

Algorithm 1. Analyzing Semantic Relations for Specificity

As the ta of $\mathcal{O}($ is a graphic taxonomy, the leaf subjects have no descendants. Thus, they have the strongest focus on their referring- to concepts and highest $spe_a($. The leaf subjects have the strongest $spe_a($of 1 in the range of 0 to1.The root subjects have the weakest $spe_a($ and smallest value in (0, 1). [4], [14]

B .Topic Specificity

Topic specificity measures the focus of subjects on the given topic. It is investigated based on the user background knowledge discovered from user local information. User background knowledge can be discovered from user local information collections, such as user's stored documents, browsed web pages, and composed/received emails. Such collections is called Local instance Repository. Catalogs of the QUT library are used as user *LIR* to populate the $\mathcal{O}(\tau)$.The reference strength between an instance and a subject is evaluated. The subjects cited by an instance are indexed by their focus. Many subjects cited by an instance may mean loose specificity of subjects, because each subject deals with only a part of the instance. Hence, denoting an instance by *i*, the strength of *i* to a subject *s* is determined by

$$str(i, s) = \frac{1}{priority(s, i) \times n(i)}$$

Where $n(i)$ the number of subjects on the citing list of *i* and $priority(s, i)$ is the index of *s* on the citing list. The $str(i, s)$ aims to selects the right instances to populated $\mathcal{O}(\tau)$.

With the $str(i, s)$determined, the relationship between an *LIR* and $\mathcal{O}(\tau)$ can be defined. Let $\Omega = \{i_1, i_2, \dots i_k\}$ be a finite and nonempty set of instances in an *LIR*, and $\min \_str$ be the minimal str value for filtering out the noisy pairs with weak strengths. Given $i \in \Omega$, we can get a set of subjects using the following mappings.

$$\eta : \Omega \rightarrow 2^S, \eta(i) = \{s \in S | str(i, s) \geq \min \_str\}$$

The mapping function η (*i*) describe the subject cited by *i*. In order to classify instances, the reverse mapping η^{-1} of η can also be defined as

$$\eta^{-1} : S \rightarrow 2^S, \eta^{-1} = \{i \in \Omega | str(i, s) \geq \min \_str\}$$

The mapping η and η^{-1} shows the relationship between instances and subjects. Each *i* maps to a set of subjects in S, and each *s* is cited by a set of instances in Ω.[14]

C. Evaluating Topic Specificity

Set of positive, negative and neutral subjects is presents in an $\mathcal{O}(\tau)$. Depending on the mapping of subject and instances, if an instances refers only to positive subjects, the instances fully supports the τ and in case of negative subjects, it is strongly against the τ. Measure of the strength of an instances to the τ is calculated as follows.

$$str(i,\tau) = \sum_{s\in(\eta(i)\cap S^{+})} str(i,s) - \sum_{s\in(\eta(i)\cap S^{-})} str(i,s)$$

If $str(i,\tau) > 0,$ i contains knowledge relevant to the τ. Otherwise, i is against the τ.

The topic specificity of a subject is evaluated based on the instance-topic strength of its citing instances. The topic specificity can also be called relative specificity with respect

to the absolute specificity and denoted by $spe_r(s,\tau,LIR)$. A subject's $spe_r(s,\tau,LIR)$ is calculated by

$$spe_r(s,\tau,LIR) = \sum_{i\in\eta^{-1}} str(i,\tau)$$

. The specificity of subject is composition of semantic specificity and topic specificity and calculated by

$$spe(s,\tau) = spe_a(s) \times spe_r(s,\tau,LI$$

The value of $spe_r(s,\tau,LI$ could be positive or negative. [4], [14]

D. Multidimensional Analysis of Subjects

The exhaustivity of a subject is the extent of its concepts space dealing with a given topic. If subjects has more positive descendants this space extends. Otherwise its exhaustivity decreases. Let $desc(s)$be a given function that returns the descendants of s in $\mathcal{O}(\tau)$.Subject's exhaustivity is calculated by aggregating the semantic specificity of its descendants

$$exh(s,\tau) = \sum_{s'\in desc(s)} \sum_{i\in\eta^{-1}(s')} str(i,\tau) \times spe_a(s',\tau)$$

If specificity and exhaustivity of the subjects are positive then subjects are interesting.

$$\begin{aligned}
\mathcal{S}^{+} &= \{s|spe(s,\mathcal{T}) > 0, exh(s,\mathcal{T}) > 0, s \in \mathcal{S}\};\\
\mathcal{S}^{-} &= \{s|spe(s,\mathcal{T}) < 0, exh(s,\mathcal{T}) < 0, s \in \mathcal{S}\};\\
\mathcal{S}^{\diamond} &= \{s|s \in (\mathcal{S} - (\mathcal{S}^{+} \cup \mathcal{S}^{-}))\}.
\end{aligned}$$

The subject sets of $S^{+}, S^{-},$can be refined after ontology mining for the specificity and exhaustivity of subject [4].

V. Ontology Model

The architecture of ontology model is shown in fig.1 Two knowledge resources, world knowledge base and local instance repository is are utilized by the model. Taxonomic structure is provided by world knowledge base whereas the background knowledge is discovered by user local instance repository.

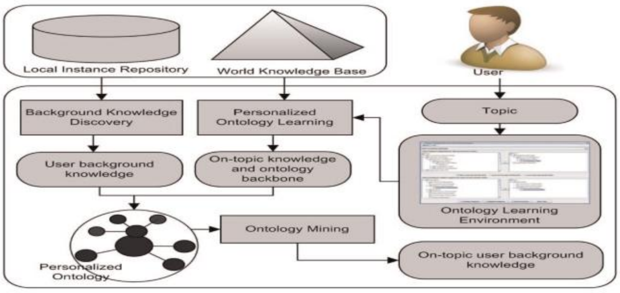

Local Instance Repository
World Knowledge Base
User
Topic
Background Knowledge Discovery
Personalized Ontology Learning
User background knowledge
On-topic knowledge and ontology backbone
Ontology Learning Environment
Ontology Mining
Personalized Ontology
On-topic user background knowledge

Fig.1. Architecture of Ontology Model [14]

VI. CONCLUSION

In this paper, we study an ontology model for personalized web information gathering. The model constructs user personalized ontologies by extracting world knowledge from the LCSH system and discovering user background knowledge from user local instance repositories. The ontology model in this paper provides a solution to emphasizing global and local knowledge in a single computational model. The findings in this paper can be applied to the design of web information gathering systems. The model also has extensive contributions to the fields of Information Retrieval, web Intelligence, Recommendation Systems, and Information Systems

REFERENCES

[1] 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.

[2] Y. Li and N. Zhong, "Web Mining Model and Its Applications for information Gathering" Knowledge-Based Systems, vol. 17, pp. 207-217, 2004.

[3] 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.

[4] X. Tao, Y. Li, N. Zhong, and R. Nayak, "Ontology Mining for Personalized Web Information Gathering," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 351-358, 2007

[5] 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

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

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

[8] R. Navigli, P. Velardi, and A. Gangemi, "Ontology Learning and Intelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.Its Application to Automated Terminology Translation," IEEE Intelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.

[9] X. Jiang and A.-H. Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents," Proc. Fifth IEEE Int'l Conf. DataMining (ICDM '05), pp. 665-668, 2005.

[10] R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, and J.X. Hao, "Towards a Fuzzy Domain Ontology Extraction Method for Adaptive e- Learning," IEEE Trans. Knowledge and Data Eng.,vol.21,no.6,pp.800-813, june 2009

[11] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to Map between Ontologies on the Semantic Web," Proc. 11th Int'l Conf ,World wide Web(WWW'02),pp.662-673,2002

[12] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence, "Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002

[13] A. Sieg, B. Mobasher, and R. Burke, "Web Search Personalization with Ontological User Profiles", Prof.16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 525-534, 2007.

[14] Xiaohui Tao, Yuefeng Li, and Ning Zhong."A Personalized Ontology Model for Web Information Gathering" IEEE Transaction on knowledge and data Engineering,vol-23,no-4,,pp-496-509,2011

[15] L.M. Chan, Library of Congress Subject Headings: Principle and Application, Libraries Unlimited,2005

[16] E. Frank and G.W. Paynter, "Predicting Library of Congress Classification from Library of Congress Subject Headings",J.Am Soc. Information Science and Technology, vol. 55, no. 3, pp. 214-227, vol.55 ,no.3.pp.214-227,2004

[17] J. Wang and M.C. Lee, "Reconstructing DDC for Interactive Classification," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 137-146, 2007.