A Personalized Ontology Model for Knowledge Representation & Reasoning User Profiles

Kanneganti Avanthi & Satya P Kumar Somayajula
Department of CSE
Avanthi Institute of Engg & Tech
Tamaram, Visakhapatnam, India.

Abstract-As a model for knowledge description and formalization, ontologies are widely used to represent user profiles in personalized web information gathering. However, when representing user profiles, many models have utilized only knowledge from either a global knowledge base or user local information. 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 results show that this ontology model is successful.

Key Terms—Ontology, personalization, semantic relations, world knowledge, local instance repository, user profiles, web information gathering

I. INTRODUCTION

An ontology model is proposed for representing user background knowledge 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. A multidimensional ontology mining method, exhaustivity and specificity, is also introduced for user background knowledge discovery. In evaluation, the standard topics and a large test bed were used for experiments. The model was compared against benchmark models by applying it to a common system for information gathering. The experiment results demonstrate that our proposed model is promising. A sensitivity analysis was also conducted for the ontology model. In this investigation, we found that the combination of global and local knowledge works better than using any one of them. In addition, the ontology model using knowledge with both is-a and part-of semantic relations works better than using only one of them. When using only global knowledge, these two kinds of relations have the same contributions to the performance of the ontology model. While using both global and local knowledge, the knowledge with part-of relations is more important than that with is-a.

The proposed 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. we will investigate the methods that generate user local instance repositories to match the representation of a global knowledge base. The present work assumes that all user local instance repositories have content-based descriptors referring to the subjects; however, a large volume of documents existing on the web may not have such content-based descriptors. For this problem, strategies like ontology mapping and text classification/clustering were suggested. These strategies will be investigated in future work to solve this problem. The investigation will extend the applicability of the ontology model to the majority of the existing web documents and increase the contribution and significance of the present work

II. RETREIVEL SYSTEM EVALUATION

Evaluation of retrieval system performance has been an integral part of the field since its beginning, but can be difficult to do well. Tague catalogs dozens of decisions that are required to design and execute a valid, efficient, and reliable retrieval test. A common way of simplifying the experimental process is to perform laboratory tests using test collections, a tradition started by the Cranfield tests. A test collection consists of a set of documents, a set of topics, and a set of relevance judgments. A topic is a description of the information being sought. The relevance judgments specify the documents that should be retrieved in response to each topic. In this paradigm, the effectiveness of different retrieval mechanisms can be directly compared on the common task defined by the test collection.

At least two questions remain when constraining retrieval experimentation to laboratory tests using test collections: how to build and validate good test collections, and what measure should be used to assess the effectiveness of retrieval output. The first question was addressed by Sparck Jones and van Rijsbergen who listed a set of criteria that an ideal test collection would meet. The test collections created through the TREC workshops have been validated by demonstrating the stability of relative retrieval scores despite incomplete relevance judgments and different opinions as to what constitutes a relevant document. Zobel and Cormack, Palmer, and Clarke proposed methods for efficiently building large test collections.

The second question—what measures should be used to evaluate retrieval effectiveness has received enormous attention in the literature. Van Rijsbergen contains a good summary, while Keen gives a detailed account on how to present retrieval results. Different evaluation measures have different properties with respect to how closely correlated they are with user satisfaction criteria, how easy they are to interpret, how meaningful aggregates such as average values are, and how much power they have to discriminate among retrieval results.

Most retrieval evaluation measures are derived in some way from *recall* and *precision*, where precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents that are retrieved. An exception are measures based on utility-theory for which the quality of retrieval output is measured in terms of its worth to the user. Utility-based measures are frequently used to evaluate set-based retrieval output such as in the TREC filtering task.

While many different evaluation measures have been defined and used, differences among measures have almost always been discussed based on their principles. That is, there has been very little empirical examination of the measures themselves. Correlation studies that build equivalence classes of measures based on how similarly they rank systems are one type of empirical study. In this paper we perform a different empirical study to quantify how stable evaluation measures are.

III. Vector Space Models

The basic assumption behind vector space models is that words that share similar contexts will have similar vector representations. Since texts consist of words, similar words will form similar texts. Therefore, the meaning of a text is represented by the sum of the vectors corresponding to the words that form the text. Furthermore, the similarity of two texts can be measured by the cosine of the angle between two vectors representing the two texts

The four items of Figure 1 can be described as follows. First, the corpus is the collection of words comprising the target texts. Second, word representation is a matrix G used to represent all words. Each word is represented by a row vector g of the matrix G. Each column of G is considered a "feature". However, it is not always clear what these features are. Third, text representation is the vector $v = G^T a$ representing a given text, where each entry of a is the number of occurrences of the corresponding word in the text. Fourth, text similarity is represented by a cosine value between two vectors.

More specifically, Equation 1 can be used to measure the similarity between two texts represented by a and b, respectively. For reasons of clarity, we do not include word weighting in this formula

Latent Semantic Analysis (LSA)

LSA is one type of vector-space model that is used to represent world knowledge (Landauer & Dumais, 1997). LSA extracts quantitative information about the co-occurrences of words in documents (paragraphs and sentences) and translates this into an N-dimensional space. The input of LSA is a large co-occurrence matrix that specifies the frequency of each word in a document. Using singular value decomposition (SVD), LSA maps each document and word into a lower dimensional space. In this way, the extremely large co-occurrence matrix is typically reduced to about 300 dimensions. Each word then becomes a weighted vector on K dimensions. The semantic relationship between words can be estimated by taking the cosine between two vectors. This algorithm can be briefly described as follows.

(1) Find the word-document occurrence matrix A from a corpus[1] .

(2) Apply SVD: TVUAΣ=K

(3) Take the row vectors of the matrix U as the vector representations of words.

Non-Latent Similarity (NLS) Model

NLS relies on a first order, non-latent matrix that represents the non-latent associations between words. The similarity between words (and documents) is calculated based on a second-order matrix. The second order matrix is created from the cosines between the vectors for each word drawn from the FOM. Hence, for NLS, the cosines are calculated based on the non-latent similarities between the words, whereas for LSA, the similarities are based on the cosines between the latent vector representations of the words. The following section describes the components and algorithms used in NLS.

Lin's (1998) Algorithm Our starting point for NLS is Lin's (1998) algorithm for extracting the similarity of words. Similarity is based upon the syntactic roles words play in the corpus. A syntactic role is designated here as a feature. For example, "the Modifier of the NP man" is a feature. A word has this feature if and only if it is used as the modifier of man when man is part of an NP in the corpus. For example, if the corpus contains the phrase the rich man, then rich has the (adjectival) feature of modifying man. Each feature is assigned a weight to indicate the feature's importance. This algorithm is briefly described as follows.

(1) For each word base, form a feature vector.

(2) For each pair of word bases, find the similarity of two word bases from the corresponding two feature vectors.

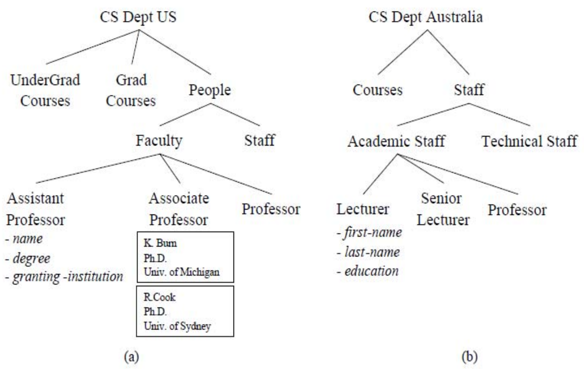

Figure 1: Computer Science Department Ontologies

IV. RELATEDWORK

ERP data consist of time series [9], representing temporal fluctuations in the EEG that are time-locked to events of interest (e.g., word or picture stimuli). In dense-array EEG and ERP research, these time series are measured across multiple locations on the scalp surface. A variety of tools are available for ERP preprocessing and pattern analysis. For example, Net Station is a suite of tools, which includes data cleaning, statistical extraction and visualization techniques. EEGLAB is a Matlab toolbox that provides advanced statistical methods for EEG/MEG and ERP pro- cessing, including independent component analysis (ICA) and joint time-frequency analysis (TFA). APECS is a Mat- lab toolbox that contains tools for data cleaning (ICA and related techniques) and evaluation of data decomposition results. The Dien PCA Toolbox 1includes Princi-pal Component Analysis (PCA) tools that are optimized for ERP data decomposition. Ontology mining is a process for learning an ontology, in- cluding classes, class taxonomy, properties and axioms. In the existing work, researchers mainly focus on mining the ontologies from text documents (e.g., web content) or other web data (web usage, web structure and web user pro- files). Clustering is used to discover the concepts in the ontology. Association rule mining has been adoptedto discover the relationships between different concepts. The NetAffx Gene ontology mining tool is an interactive platform for visualizing and analyzing microarray data. In this paper, we propose a generic framework for developing and mining domain ontologies, with specific applicationto the development of a first-generation ERP ontology. Thetarget data type consists of spatiotemporal data (ERPs), and summary statistics (e.g., the "latent" or principal com- ponents that emerge from statistical analysis of ERP data). In addition to identifying classes, a hierarchy of classes and part-of relations of classes, our approach includes classification methods for mining properties and axioms (rules). This is also an important extension from our previous work, which focuses only on ERP pattern mining. In this paper, we first use the previous ERP pattern mining results (data from Experiment 1-2) to develop ERP classes. Furthermore, we adopt hierarchical clustering methods to generate class taxonomies and association rules to discover the property relations respectively from a new dataset

V. EXISTING SYSTEM

Local analysis investigates user local information or observes user behavior in user profiles. For example, Li and Zhong discovered taxonomical patterns from the users' local text documents to learn ontologies for user profiles. Some groups learned personalized ontologies adaptively from user's **browsing history**. Alternatively, Sekine and Suzuki analyzed query logs to discover user background knowledge. In some works, such as, users were provided with a set of documents and asked for **relevance feedback**. User background knowledge was then discovered from this feedback for user profiles. However, because local analysis techniques rely on data mining or classification techniques for knowledge discovery, occasionally the discovered results contain **noisy and uncertain information**. As a result, local analysis suffers from **in-effectiveness** at capturing formal user knowledge.

VI. PROPOSED SYSTEM

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 constrain the background knowledge discovery from the user local information. Such a personalized ontology model should produce a superior representation of user profiles for web information gathering.

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; 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 Exhaustivity, 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**.

Advantages

Which one is more important: the WKB or LIRs? The Loc model using only user LIRs had substantially low performance, compared with the GP, GI, and GIP models using only the WKB. Thus, The WKB is more important than user LIRs. In addition, the GP, GI, and GIP models using the

WKB also have the knowledge with is-a and/or part-of semantic relations. The Loc model, however, has no such relations specified. Hence, it is reasonable to conclude that a part of the improvement achieved by the GP, GI, and GIP models is due to the is-a and/or part-of knowledge. We then have an extensive finding: the knowledge with is-a and/or part of relations is an important component of the ontology model.

VII. CONCLUSION

An ontology model is proposed for representing user background knowledge 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. A multidimensional ontology mining method, exhaustivity and specificity, is also introduced for user background knowledge discovery. In evaluation, the standard topics and a large test bed were used for experiments. The model was compared against benchmark models by applying it to a common system for information gathering. The experiment results demonstrate that our proposed model is promising. A sensitivity analysis was also conducted for the ontology model. In this investigation, we found that the combination of global and local knowledge works better than using any one of them. In addition, the ontology model using knowledge with both is-a and part-of semantic relations works better than using only one of them. When using only global knowledge, these two kinds of relations have the same contributions to the performance of the ontology model. While using both global and local knowledge, the knowledge with part-of relations is more important than that with is-a.

VIII. FUTURE ENCHANCEMENTS

The proposed 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. We will investigate the methods that generate user local instance repositories to match the representation of a global knowledge base. The present work assumes that all user local instance repositories have content-based descriptors referring to the subjects; however, a large volume of documents existing on the web may not have such content-based descriptors. For this problem, strategies like ontology mapping and text classification/clustering were suggested. These strategies will be investigated in future work to solve this problem. The investigation will extend the applicability of the ontology model to the majority of the existing web documents and increase the contribution and significance of the present work.

REFERENCES

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

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

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

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

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

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

[7] 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.

[8] D. Dou, G. Frishkoff, J. Rong, R. Frank, A. Malony, and D. Tucker, "Development of Neuroelectromagnetic Ontologies(NEMO): A Framework for Mining Brainwave Ontologies," Proc. ACM SIGKDD ('07), pp. 270-279, 2007.

[9] 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.

[10] 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.

About the Authors:

Mrs. Kanneganti Avanthi received the B.Tech degree in Prakasam Engineering College in 2005 under JNT University Hyderabad and She is currently pursuing M.Tech in Software Engineering at Avanthi Institute of Engineering and Technology, Vishakhapatnam .Her research interests include Data Mining and Software Engineering.

Mr. Satya P Kumar Somayajula is working as an Asst. Professor, in CSE Department,Avanthi Institute of Engineering and Technology,Tamaram, Vishakapatnam,A.P.,India.
He has received his M.Sc (Physics) from Andhra University, Visakhapatnam and M.Tech (CST) from Gandhi Institute of Technology And Management University (GITAM University), Visakhapatnam, A.P., INDIA. He published 18 papers in reputed International journals & 5 National journals. His research interests include Image Processing, Networks security, Web security, Information security, Data Mining and Software Engineering.