www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume1 Issue 3 Dec 2012 Page No. 121-124

ONTOLOGY MINING FOR PERSONALIZED WEB INFORMATION GATHERING

Miss. Deshmukh Rupali R. Prof. Keole R.R

M.E.Final year CSE H.V.P.M's, C.O.E.T, Amravati

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. User profiles represent the concept models possessed by users when gathering web information. A concept model is implicitly possessed by users and is generated from their background knowledge. While this concept model cannot be proven in laboratories, many web ontologists have observed it in user behaviourThe results show that this ontology model is successful.

Keywords— Ontology, personalization, world knowledge, local instance repository, user profile, semantic relations, web information gathering.

Introduction

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 .User profiles represent the concept models possessed by users when gathering web information. A concept model is implicitly possessed by users and is generated from their background knowledge. While this concept model cannot be proven in laboratories, many web ontologists have observed it in user behavior. When users read through a document, they can easily determine whether or not it is of their interest or relevance to them, a judgment that arises from their implicit concept models. If a user's concept model can be simulated, then a superior representation of user profiles can be built. To simulate user concept models, ontologies—a knowledge description and formalization model—are utilized in personalized web information gathering. Such ontologies are called ontological user profiles or personalized ontologies. To represent user profiles, many researchers have attempted to discover user background knowledge through global or local analysis. Global analysis uses existing global knowledge bases for user background knowledge representation. Commonly used knowledge bases include generic ontologies (e.g.,WordNet), 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, WorldNet was reported as helpful in capturing user interest in some areas but useless for others. 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 ineffectiveness at capturing formal user knowledge. From this, 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.
Ontology mining discovers interesting and on-topic knowledge from the concepts, semantic relations, and instances in an ontology. In this section, a 2D ontology mining method is introduced: Specificity and Exhaustivity. Specificity (denoted spe) describes a subject's focus on a given topic. Exhaustivity restricts a subject's semantic space dealing with the topic. This method aims to investigate the subjects and the strength of their associations in an ontology.

Literature Review & Related work:

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

We have to analysis the **DATA MINING Outline Survey**:

Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides.

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Architecture for Data Mining

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

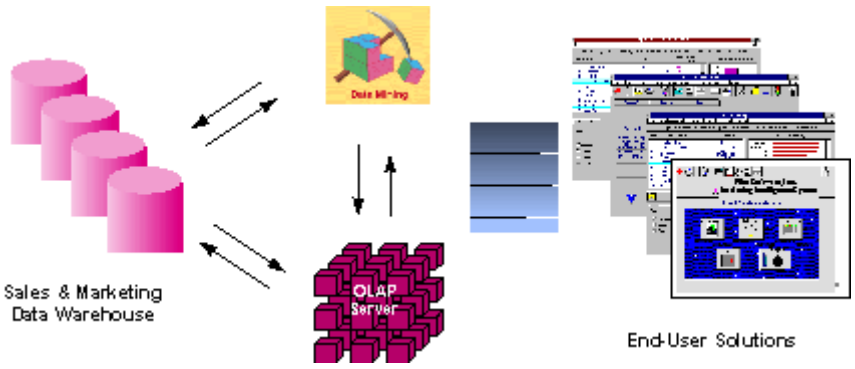

Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

Problem Definition

We 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, in Section 4.2, 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.

Existing System

A. **Golden Model: TREC Model**

The TREC model was used to demonstrate the interviewing user profiles, which reflected user concept models perfectly. For each topic, TREC users were given a set of documents to read and judged each as relevant or nonrelevant to the topic. The TREC user profiles perfectly reflected the users' personal interests, as the relevant judgments were provided by the same people who created the topics as well, following the fact that only users know their interests and preferences perfectly.

B. Baseline Model: Category Model

This model demonstrated the non-interviewing user profiles, a 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 re-ranks the results based on their semantic similarity with the subjects. The similar documents are awarded and re-ranked higher on the result list.

C. Baseline Model: Web Model

The web model was the implementation of typical semi interviewing user profiles. It acquired user profiles 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.

Limitations Of Existing System

The topic coverage of TREC profiles was limited. The TREC user profiles had good precision but relatively poor recall performance. 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.

Proposed System

The world knowledge and a user's local instance repository (LIR) are used in the proposed model.

1) World knowledge is commonsense knowledge acquired by people from experience and education

2) 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 Figure2 shows the proposed System Framework.

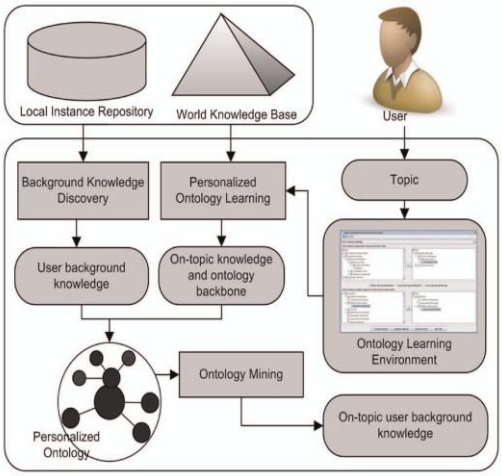

Figure2: Proposed System framework.

Advantages Of Proposed System

A. Compared with the TREC model, the Ontology model had better recall but relatively weaker precision performance. The Ontology model discovered user background knowledge from user local instance repositories, rather than documents read and judged by users. Thus, the Ontology user profiles were not as precise as the TREC user profiles.

B. The Ontology profiles had broad topic coverage. The substantial coverage of possibly-related topics was gained from the use of the WKB and the large number of training documents.

C. Compared to the web data used by the web model, the LIRs used by the Ontology model were controlled and contained less uncertainties. Additionally, a large number of uncertainties were eliminated when user background knowledge was discovered. As a result, the user profiles acquired by the Ontology model performed better than the web model.

Conclusion

This ontology model is proposed for representing user background knowledge for personalized web information gathering. The model constructs user personalized ontologies by extracting world knowledge and discovering user background knowledge from user local instance repositories. 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.

Reference

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.
[2] G.E.P. Box, J.S. Hunter, and W.G. Hunter, Statistics For Experimenters.John Wiley & Sons, 2005.
[3] C. Buckley and E.M. Voorhees, "Evaluating Evaluation Measure Stability," Proc. ACM SIGIR '00, pp. 33-40, 2000.
[4] Z. Cai, D.S. McNamara, M. Louwerse, X. Hu, M. Rowe, and A.C. Graesser, "NLS: A Non-Latent imilarity Algorithm," Proc. 26th Ann. Meeting of the Cognitive Science Soc. (CogSci '04), pp. 180-185, 2004.
[5] 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.
[6] L.M. Chan, Library of Congress Subject Headings: Principle and Application. Libraries Unlimited, 2005.
[7] E. Frank and G.W. Paynter, "Predicting Library of Congress Classifications from Library of Congress Subject Headings," J. Am Soc. Information Science and Technology, vol. 55, no. 3, pp. 214-227,2004.
[8] R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen, "Using Google Distance to Weight Approximate Ontology Matches," Proc. 16th Int'l Conf. World Wide Web (WWW '07), pp. 767-776, 2007.