ONTOLOGY BASED INFORMATION GATHERING FROM WEB PAGES

M.KAMARUNISHA[1],S.GOWRI[2],A.SIVASANKARI[3]

Assistant Professor,Department of Computer Applications,Dhanalakshmi Srinivasan College of Arts and Science For Women(Autonomous),Perambalur.

Abstract

Information retrieval has a well-established tradition of performing laboratory experiments on test collections to compare the relative effectiveness of different retrieval approaches. The experimental design specifies the evaluation criterion to be used to determine if one approach is better than another. Retrieval behavior is sufficiently complex to be difficult to summarize in one number, many different effectiveness measures have been proposed. A concept model is implicitly possessed by users and is generated from their background knowledge. 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.

Keywords: Personalization, Ontology, semantic relations, world knowledge, local instance repository, user profiles.

1. 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.

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 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.

This paper presents the extensive work of, but significantly beyond, an earlier paper published in WI '07. The authors thank the Library of Congress and QUT Library for the use of the LCSH and library catalogs. The authors also thank the anonymous reviewers for their valuable comments. Thanks also go to M. Carey-Smith, P. Delaney, and J. Beale, for their assistance in proofreading and editing the paper.

2.BACKGROUND

This paper examines these three rules-of-thumb and shows how they interact with each other. We present a novel approach for experimentally quantifying the likely error associated with the conclusion "method A is better than method B" given a number of requests, an evaluation measure, and a notion of difference.

One increasingly popular way to structure information is through the use of ontologies, or graphs of concepts. One such system is *Onto Seek,* which is designed for content-based information retrieval from online yellow pages and product catalogs. The system uses the *Sensus* ontology, which comprises a simple taxonomic structure of approximately 70,000 nodes. The system presented in uses *Yahoo!* as an ontology. The system semantically annotates Web pages via the use of Yahoo! categories as descriptors of their content.

The system uses *Telltale* as Its classifier. Telltale computes the similarity between documents using *n*- grams as index terms. The ontologies used in the above examples use simple structured links between concepts. A richer and more powerful representation is provided by *SHOE* is a set of Simple HTML Ontology Extensions that allow WWW authors to annotate their pages with semantic content expressed in terms of ontology.

An ontology is then constructed for the given topic using these user fed back subjects. The structure of the ontology is based on the semantic relations linking these subjects in the WKB. The ontology contains three types of knowledge: positive subjects, negative subjects, and neutral subjects. Fig. 1 illustrates the ontology (partially) constructed for the sample topic "Economic espionage," where the white nodes are positive, the dark nodes are negative, and the gray nodes are neutral subjects. Here, we formalize the ontology constructed for a given topic

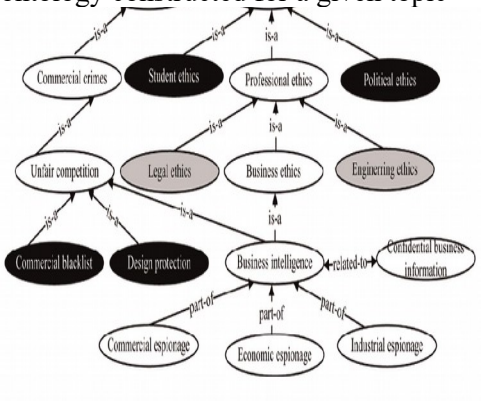


Figure 1 An ontology (partial) constructed for topic "Economic Espionage."

3.PROPOSED MODEL

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization's operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following: -

- Is there sufficient support for the management from the users?
- Will the system be used and work properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration.

So there is no question of resistance from the users that can undermine the possible application benefits. The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

Technical Feasibility

The technical issue usually raised during the feasibility stage of the investigation includes the following:

- Does the necessary technology exist to do what is suggested?
- Do the proposed equipments have the technical capacity to hold the data required to use the new system?
- Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
- Can the system be upgraded if developed?
- Are there technical guarantees of accuracy, reliability, ease of access and data security?

Earlier no system existed to cater to the needs of 'Secure Infrastructure Implementation System'. The current system developed is technically feasible. It is a web based user interface for audit workflow at NIC- CSD. Thus it provides an easy access to the users.

The database's purpose is to create, establish and maintain a workflow among various entities in order to facilitate all concerned users in their various capacities or roles. Permission to the users would be granted based on the roles specified. Therefore, it provides the technical guarantee of accuracy, reliability and security.

The software and hard requirements for the development of this project are not many and are already available in-house at NIC or are available as free as open source. The work for the project is done with the current equipment and existing software technology. Necessary bandwidth exists for providing a fast feedback to the users irrespective of the number of users using the system.

Analyzing semantic relations for specificity: the Algorithm

Terms Expansion:

Tax: Taxonomic structure can be provide the edge communication

Rel: can be provides the Boolean operations

Aim:

1.It can be identifies the specificity of information like leaves identification process
2. Using leaves maintain the two types of relationship operation like is-a and part-of.
3. In between of two types of relationship to maintain the union operation.
4.All the subjects to related objects to arrange in the form of tree format of structure.

input : a personalized ontology $\mathcal{O}(\mathcal{T}) := \langle tax^{\mathcal{S}}, rel \rangle$; a coefficient θ between (0,1).
output: $spe_a(s)$ applied to specificity.

1 set $k = 1$, get the set of leaves S_0 from $tax^{\mathcal{S}}$, 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 $tax^{\mathcal{S}}$;
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.

Figure 2 Algorithm

Web mining can be useful to add semantic annotations to Web documents and to populate these ontological structures. As stated below, Web content and Web usage mining should be combined to extract ontologies and to adapt them to the usage. Ontology creation and evolution require the extraction of knowledge from heterogeneous sources.

In the case of the Semantic Web, the knowledge extraction is done from the content of a set of Web pages dedicated to a particular domain. Web pages are semi-

structured information. Web usage mining extracts navigation patterns from Web log files and can also extract information about the Web site structure and user profiles

Today it is common to specify systems on higher levels using some natural language (e.g. English). For large systems, where large amounts of information must be handled, problems arise with ambiguities and inconsistencies with such specifications. Errors that are introduced are often detected late in the design cycle - in the simulation of the design after much design work has already been carried out - if detected at all.

By making the initial system specifications in a formal language at a high abstraction level, functionality can be verified/simulated earlier in the development process. Ambiguities and inconsistencies can be avoided, errors can be discovered earlier, and the design iteration cycles can be shortened, thereby reducing development times.

It is of critical importance that the specification language provides modeling concepts at a high abstraction level to allow the representation of system functions at a conceptual level without introducing unnecessary details.

Further, most of the languages that are used for implementation of HW / SW designs (e.g. VHDL, C++) do not lend themselves well to formal verification. This is because they lack a formally defined semantics or because the semantics is complex.

A lack of formal semantics sometimes causes ambiguities in the interpretation of the designs. Our goal is to develop functional system specification method for telecom systems, to demonstrate its efficacy on an industrially relevant example and to develop a tool to support the mapping of such specifications to synthesizable VHDL/C++. The specification language in which the system level functions will be developed will have a formal semantics in order to support formal verification of specifications.

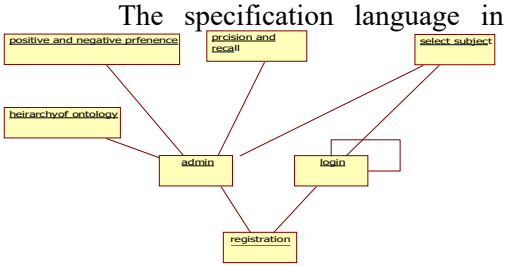


Figure 3 Architecture

System Components:

The fig 3 shows a general view of how desktop and workstation computers are organized. Different systems have different details, but in general all computers consist of components (processor, memory, controllers, video) connected together with a *bus*. Physically, a **bus** consists of many parallel wires, usually printed (in copper) on the main circuit board of the computer.

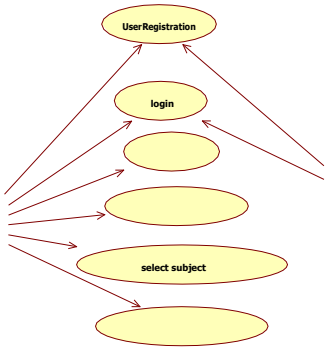


Figure 4 Use Case Diagram

Data signals, clock signals, and control signals are sent on the bus back and forth between components. A particular type of bus follows a carefully written standard that describes the signals that are carried on the wires and what the signals mean.

Figure 5 Collaboration Diagram

Precision and Recall graph

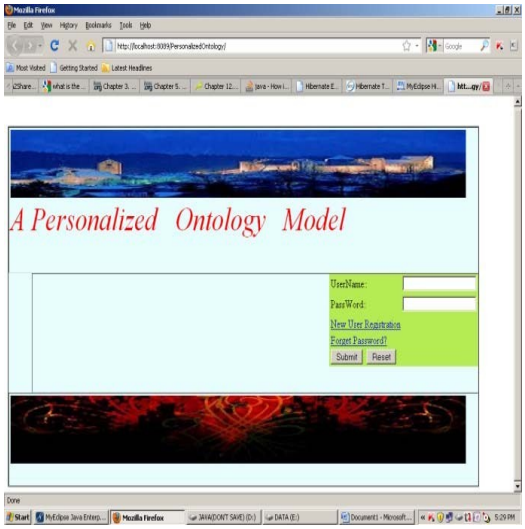


Figure 6 Login Screen

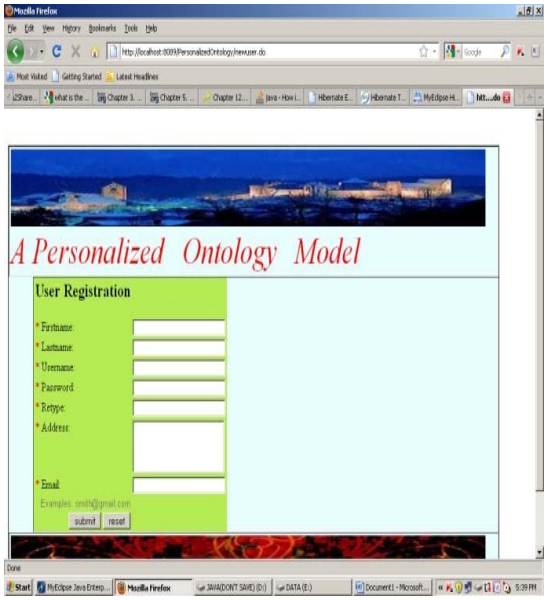


Figure 7 Registration Screen

Figure 8 Home Page

4. CONCLUSIONS & ENHANCEMENTS

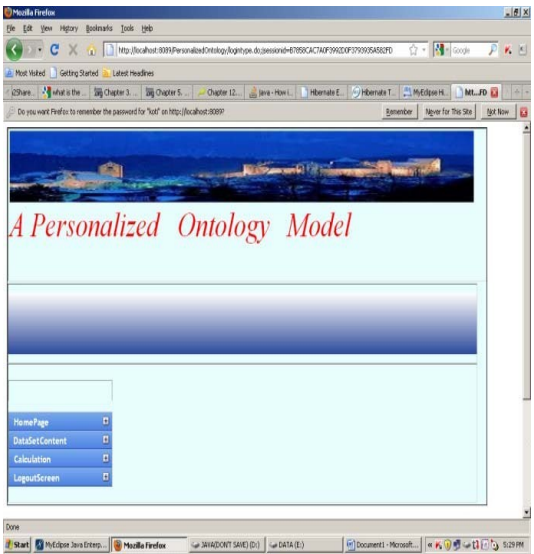
The Above Figures shows the developed model. The UML and use case diagrams are also shown in the above figures. An ontology model is proposed for representing user background knowledge for personalized web information gathering. The model constructs user personalized ontologies by extracting world knowledge from the LCSH system and discovering user background knowledge from user local instance repositories. 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.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their comments which were very helpful in improving the quality and presentation of this paper.

REFERENCES:

[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 Similarity Algorithm," Proc. 26th Ann. Meeting of the Cognitive Science Soc. (CogSci '04), pp. 180-185, 2004.

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

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

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

[8] 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.

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

[10] 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.

[11] 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.