

Performance of Web Information Gathering System using TREC Model, Category Model, Web Model and Ontology Model

¹K. Leela Lakshmi Devi, ²P. Radhakrishna, ³K. Parimala

^{1,2,3}Dept. of CSE, NOVA College of Engineering & Technology, Vijayawada, AP, India

Abstract

In this paper, 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, exhaustively and specificity, is also introduced for user background knowledge discovery. In evaluation, the standard topics and a large tested 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. 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.

Keywords

Ontology, personalization, semantic relations, world knowledge, local instance repository, user profiles, web information gathering.

I. Introduction

Over the last decade, the rapid growth and adoption of the World Wide Web have further exacerbated user need for efficient mechanisms for information and knowledge location, selection and retrieval. Web information covers a wide range of topics and serves a broad spectrum of communities [1]. How to gather useful and meaningful information from the Web, however, become challenging to Web users. This challenging issue is referred by many researchers as Web information gathering [3]. Given information needs, Web information gathering aims to acquire use- full and meaningful information for users from the Web. The Web information gathering tasks are usually completed by the systems using keyword-based techniques. The keyword-based mechanism searches the Web by ending the documents with the specific terms matched. This mechanism is used by many existing Web search systems, for example, Google and Yahoo! information gathering. Han and Chang [7] pointed out that by using keyword-based search techniques, the Web information gathering systems can access the information quickly; however, the gathered information may possibly contain much useless and meaningless information. This is particularly referred as the

fundamental issue in Web information gathering: information mismatching and information overloading [2]. In attempting to solve these fundamental problems, many researchers have aimed at gathering personalized Web information for users with better effectiveness and efficiency. These researchers have not only moved information gathering from keyword-based to concept-based, but also take user background knowledge into consideration. In these works, Web user profiles are widely used for user modeling and personalization, because they react the interest and preferences of users. They are used in Web information gathering to describe user background knowledge, to capture user information needs, and to gather personalized Web information for users. This survey paper attempts to review the development of the concept-based, personalized Web information gathering techniques. The review notes the issues in Web personalization, focusing on Web user profiles and user information needs in personalized Web information gathering. The reviewed scholar reports that the concept-based models utilizing user background knowledge are capable of gathering useful and meaningful information for Web users. However, the representation and acquisition of user profiles need to be improved for the effectiveness of Web information gathering. This survey has contributions to better understanding of existing Web information gathering systems. The paper is organized as follows. Section 2 reviews the concept-based Web information gathering techniques, including concept representation and extraction. Section 3 presents the survey of personalized Web information gathering, including user profile representation and acquisition. Finally, Section 4 makes the final remarks for the survey.

II. Existing System

In this section we discuss about the models of web gathering information by using TREC model, Web Model, category Models.

A. TREC MODEL

When selecting the TREC queries to be used in the evaluation benchmark, we had two practical constraints. First, the queries must be able to be formulated in a way that is suitable for ontology-based search systems; for example, queries such as “discuss the financial aspects of retirement planning” cannot be tackled because they are navigational and not research searches. Second, ontologies must be available for the domain of the query. As discussed above, the second point is a serious constraint. Finally, we selected 20 queries.

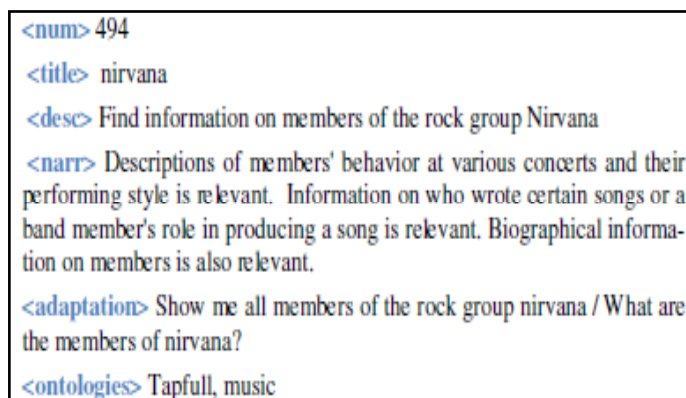


Fig. 1: Example of a TREC Topic

TREC queries are described by:

- A title, which is the original user query extracted from users' logs.
- A description.
- A narrative, which explains in more detail the relevant information that the user is looking for. We added to the queries introduced in the benchmark.
- A detailed request, potentially more suitable for ontology-based search approaches.
- Notes on available ontologies covering that query. The complete list of the selected TREC topics and their adaptation available.

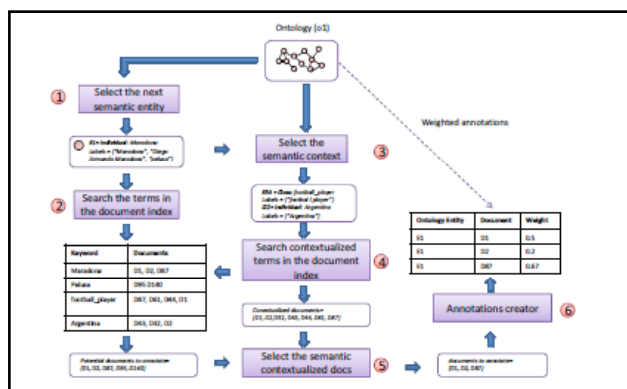


Fig. 2: The Annotation Process

1. Load the information of a semantic entity. Extract the textual representation of the selected semantic entity. Each semantic entity has one or more textual representations in the ontology. E.g., the individual entity describing the football player Maradona can be named as "Maradona", "Diego Armando Maradona", "Pelusa", etc. Here we assume that such lexical variants are present in the ontology as multiple values of the local name or rdfs: label property of the entity.
2. Find the set of potential documents to annotate. The textual representations of the semantic entity are then searched for in the document index using standard search and ranking processes, in order to find the documents that may be associated with it. These documents simply contain a textual representation of the semantic entity, which does not necessarily imply that they contain its meaning; they are candidates for annotation, to be considered by the following steps.
3. Extract the semantic context of the entity. The meaning of a concept is determined by the set of concepts it is linked or related to in the domain ontology. To ensure that a semantic

entity annotates the appropriate set of documents, we exploit the ontological relations to extract its context, that is, the set of entities directly linked in the ontology by an explicit relation. E.g., the semantic entity Maradona is related to the concepts Football player, Argentina, etc.

4. Find the set of contextualized documents. The textual representations of entities in the set of semantically related concepts, or semantic context, produced in the previous step, are then searched for in the document index to extract the set of contextualized documents.
5. Select the final list of documents to annotate. We compute the intersection between the documents having textual representations of the semantic entity (extracted in step 2) and the set of documents having textual representations of the entities on its semantic context (extracted in step 4). Documents in this set Ontology (o1) meaning of the concept in the ontology.
6. Create the annotations. A new entry or annotation is created for every document in the previously obtained set. The annotation will have a weight indicating the degree of relevance of the entity within the document. These weights are computed in the following way: the fusion methodology, described in (Fernandez, 2006), is used on the ranked lists of documents obtained at steps 2 and 4 to produce a ranked list of documents that are candidates for annotations and a ranked list of contextualized documents for semantically related entities, respectively. A document _occurring in both, and hence selected for annotation by step 5, will be given a weight where is a constant used control the influence of the semantic contextualization. We empirically found that a value seems to work well in practice. In the annotation mechanism reported here, the semantic entities are analyzed and searched in a standard keyword-base document index. This annotation process provides two important advantages: on the one hand, the semantic information stored in the ontologies and KBs can be used as background knowledge to improve the accuracy of the annotations; on the other hand, this annotation model constitutes a more scalable and widely applicable approach because it can potentially rely on any keyword-based document index, including the ones generated by big search engines

1. Concerns About using the TREC Web Track Test Collection

Several concerns about using the TREC Web track for the evaluation of ontology-based search approaches should be considered. The judgments: the judgments for each query of TREC 9 and TREC 2001 competitions are obtained using the pooling method described in Section II. In this methodology, retrieval systems that did not contribute to the pools might retrieve nudged documents that are assumed to be non-relevant, which, as described in later Studies (Voorhees E., 2001) leads to their evaluation scores being deflated relatively to the methods that did contribute. The queries: the queries selected for TREC 9 and TREC 2001 are extracted from real Web search engine logs. This means that, the queries are generated in a suitable way for traditional keyword based search engines and therefore, ontology-based search model are not exploiting their capabilities of addressing more complex types of queries. Consequently, the benchmark might be biased and be giving advantage to keyword-based search approaches. The query construction: in TREC 9 and TREC 2001 different evaluation categories are considered depending on how the input queries are

formulated: a) short runs, using just the title or b) no short runs, automatically or manually constructing the query from the title, the description and the narrative. A better performance is expected from approaches which manually construct the queries (not short runs) than from those that use just the title (short runs) because they add a significant amount of additional information to the query. Some semantic search systems could require manually modifying the provided set of queries to fulfill their corresponding input format. However, this does not necessarily mean that additional information is added to the query. In these cases, the comparison of ontology-based search models against traditional IR manual approaches cannot be considered fair.

B. 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. For a given topic, a set of feature terms $f_{tj} 2 T_p$ and a set of noisy terms $f_{tj} 2 T_g$ were first manually identified. The feature terms referred to the interesting concepts of the topic. The noisy terms referred to the paradoxical or ambiguous concepts. Also identified were the certainty factors CF_{dt} of the terms that determined their supporting rates $([-1, 1])$ to the topic. By using the feature and noisy terms, the Google4 API was employed to perform two searches for the given topic. The first search used a query generated by adding “p” symbols in front of the feature terms and “-” symbols in front of the noisy terms. By using this query, about 100 URLs were retrieved for the positive training set. The second search used a query generated by adding “_” symbols in front of feature terms and “p” symbols in front of noisy terms. Also, about 100 URLs were retrieved for the negative set.

These positive and negative documents were filtered by their certainty degrees CD . The CD_{dd} was determined by the document's index ind_{dd} on the returned list from Google and Google's precision rate. The δ was set as 0.9, based on the preliminary test results using experimental topics. If a document d was in the cutoff δ and $\delta \geq 0.9$, the Google's confidence on d would be 0.9. Together with the CF_{dt} values of the feature terms and noisy terms, we had a CD_{dd} calculated by

$$CD(d) = \delta \times \frac{K - ind(d)(mod \kappa) + 1}{K} \times \sum_{t \in (T \cap d)} |CF(t)|,$$

$\delta 14p$ where, K is a constant number of 10 for the number of documents in each cutoff δ , T refers to T_p or T_g , depending on the positive or negative set that d is in. The support value of a document was finally determined by supported δCD_{dd} δCD_{dd} . The positive training set was then generated by the documents with supported > 0 , and the negative set by the documents with supported $= 0$.

C. Category Model

This model demonstrated the no interviewing user profiles, in particular Gauche et al.'s OBIWAN [12] model. In the OBIWAN model, 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 an 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. In this Category model, the sets of positive subjects were manually fed back by the user via the OLE and from the WKB, using

the same process as that in the Ontology model. The Category model differed from the Ontology model in that there were no is-a, part-of, and related-to knowledge considered and no ontology mining performed in the model. The positive subjects were equally weighted as one, because there was no evidence to show that a user might prefer some positive subjects more than others. The training sets in this model were extracted through searching the subject catalog of the QUT library, using the same process as in the Ontology model for user LIRs. However, in this model, a document's supported value was determined by the number of positive subjects cited by d . Thus, more positive subjects cited by d would give the document a stronger supported value. There was no negative training set generated by this model, as it was not required by the OBIWAN model.

III. Proposed Work on Ontology Model

This model was the implementation of the proposed ontology model. As shown in Fig. 7, the input to this model was a topic and the output was a user profile consisting of positive documents (D_p) and negative documents (D_g). Each document d was associated with a supported value indicating its support level to the topic. The WKB was constructed based on the LCSH system, as introduced. The LCSH authority records distributed by the Library of Congress were a single file of 130 MB compiled in Machine-Readable Cataloging (MARC) 21 format. After data preprocessing using expression techniques, these records were translated to human-readable form and organized in an SQL database, approximately 750 MB in size. Theoretically, the LCSH authority records consisted of subjects for personal names, corporate names, meeting names, uniform titles, bibliographic titles, topical terms, and geographic names. In order to make the Ontology model run more efficiently, only the topical, corporate, and geographic subjects were kept in the WKB, as they covered most topics in daily life. The BT, UF, and RT references (referred to by “450 jw j a”, “450,” and “550” in the records, respectively) linking the subjects in the LCSH thesaurus were also extracted and encoded as the semantic relations of is-a, part-of, and related-to in the WKB, respectively. Eventually, the constructed WKB contained 394,070 subjects covering a wide range of topics linked by semantic relations. The user personalized ontologies were constructed as described in Section 3.2 via user interaction. The authors played the user role to select positive and negative subjects for ontology construction, following the descriptions and narratives associated with the topics. On average, each personalized ontology contained about 16 positive and 23 negative subjects.

For each topic T , the ontology mining method was performed on the constructed OOT P and the user LIR to discover interesting concepts. The user LIRs were collected through searching the subject catalog of the QUT library by using the given topics. The catalog was distributed by QUT library as a 138 MB text file containing information for 448,590 items. The information was preprocessed by removing the stop words, and stemming and grouping the terms. Librarians and authors have assigned title, table of content, summary, and a list of subjects to each information item in the catalog. These were used to represent the instances in LIRs. For each one of the 50 experimental topics, and thus, each one of the 50 corresponding users, the user's LIR was extracted from this catalog data set. As a result, there were about 1,111 instances existing in one LIR on average. The semantic relations of is-a and part-of were also analyzed in the ontology mining phase for interesting knowledge discovery. For the coefficient δ in Algorithm 1, some preliminary tests had been conducted

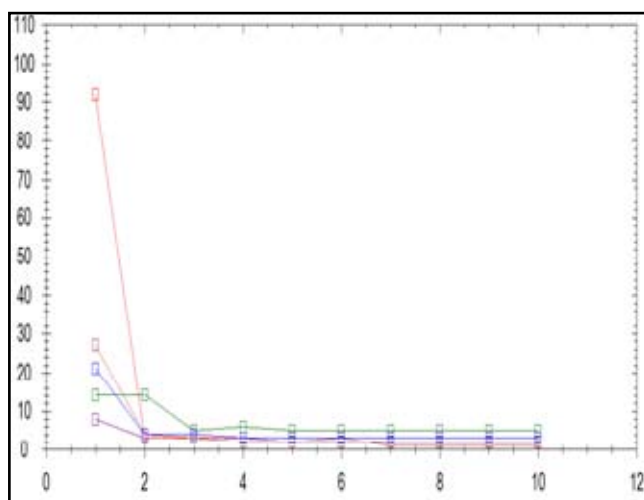
for various values (0.5, 0.7, 0.8, and 0.9). As a result, $\frac{1}{4}$ 0.9 gave the testing model the best performance and was chosen in the experiments. Finally, a document d in the user profile was generated from an instance i in the LIR. The d held a support value supported to the T , which was measured by

$$support(d_i) = str(i, T) \times \sum_{s \in \eta(i)} spe(s, T),$$

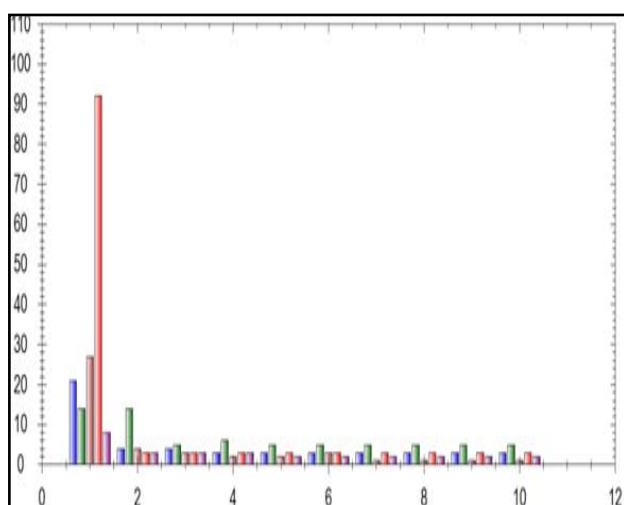
Where supported was defined by, supported by. When conducting the experiments, we tested various thresholds of supported to classify positive and negative documents. However, because the constructed ontologies were personalized and focused on various topics, we could not find a universal threshold that worked for all topics. Therefore, we set the threshold as supported $\frac{1}{4}$ 0, following the nature of positive and negative defined in this paper. The documents with supported > 0 formed D_p , and those with negative supported < 0 formed D_n eventually.

IV. Results

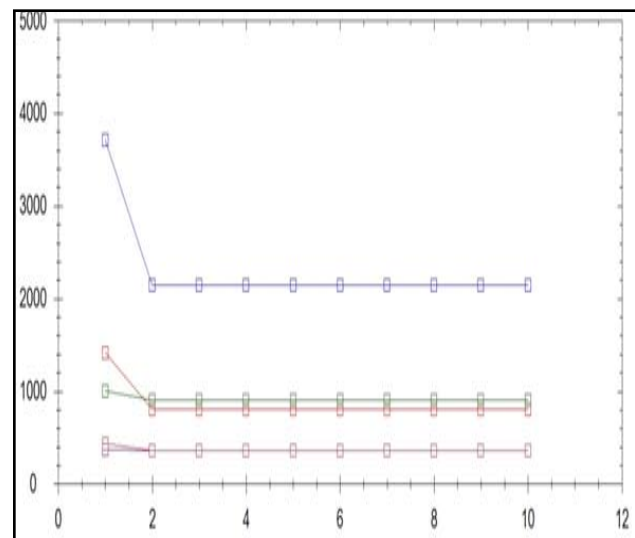
In order to implement the required TREC, Web, Category and Ontology Models to Simulate the graphs and performance Of this models shown by the visually by Graphs.



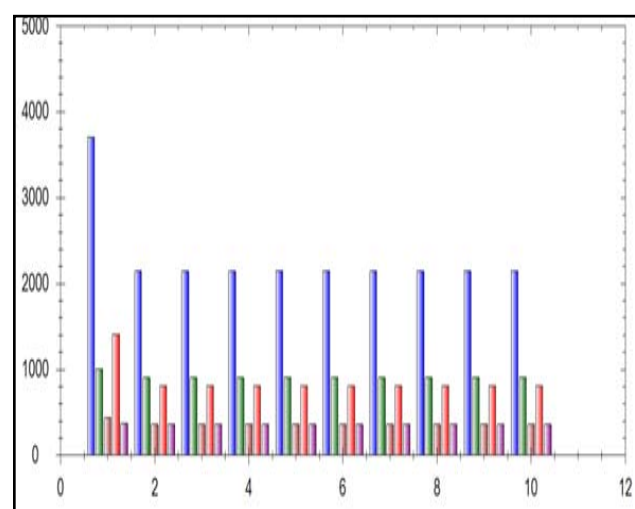
(a)



(b)



(c)



(d)

Fig. 3: Results

V. Conclusions and Future Work

In this paper 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. And also we simulate these models by using graphs and evaluate performance of the models, so finally we proposed the web information gathering by using ontology model its very fastest to investigated the information and we proved.

References

- [1] R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", Addison Wesley, 1999.
- [2] G.E.P. Box, J.S. Hunter, W.G. Hunter, "Statistics For Experimenters", John Wiley & Sons, 2005.
- [3] C. Buckley, 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, A.C.Graesser, "NLS: A Non-Latent Similarity Algorithm", Proc. 26th Ann. Meeting of the Cognitive Science Soc. (CogSci '04), pp. 180-185, 2004. on Mobile Ad-Hoc Networking: Research, Trends and Applications, Vol. 2, pp. 483-502, 2002.
- [5] L.M. Chan, "Library of Congress Subject Headings: Principle and Application", Libraries Unlimited, 2005.
- [6] P.A. Chirita, C.S. Firan, 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, A. Halevy, "Learning to Map between Ontologies on the Semantic".



K. Leela Lakshmi Devi has completed M.Sc in CSE from Annabathuni Satya Narayana P.G & Degree College in 2010 and Pursuing final M.Tech in NOVA College of Engineering Technology for Women. Her research work includes concepts of Data Mining and Data Warehousing.



Mr. P. Radha Krishna did his Master in Computer Application at Lakkireddy Bali eddy college of engineering. He started his teaching career at Anurag Engineering College, Kodad. Where He had guided be so many batches in their projects. He guided students in doing IBM projects at Anurag. He played an active role in the NBA Accreditation process at Anurag Engineering College in to 2007, He had acted as In charge Hod of MCA at Anurag. Mr. P. Radha Krishna is a Members of ISTE. He had participated many workshops and paper presentations, which include sunmicrosystem and open systems (LAMP) workshops. Mr. Radha Krishna is the secretary of Alumni Association of LBRCE. He is now working as the Head of the Department of CSE of Nova college of Engineering & Technology, Jupudi, Ibrahimpatnm.



N. Parimala has completed B.Tech from Potluri VaraPrasad Siddhartha Institute & Technology. in 2008 and Pursuing final M.Tech in NOVA College of Engineering Technology for Women. Her research work includes concepts of Data Mining and Data Warehousing.