EXTRACTION OF WEB INFORMATION BY USING PERSONALIZED ONTOLOGY TECHNIQUE

ADAPA HIMABINDU [#1] & V.SARALA[#2] & D.D.D.SURIBABU[#3]
[#1] M.Sc Student, Master of Computer Science, D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.
[#2] Assistant Professor, Master of Computer Science, D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.
[#3] Head & Associate Professor, Dept of CSE, D.N.R. College of Engineering, Bhimavaram, AP, India.

Received: January 11, 2019 Accepted: February 13, 2019

ABSTRACT: *The term ontology is mainly represented as the process of extracting the valuable information from a large data source. Now a day for any sort of information, we try to search the web and extract the documents to and from the web. Due to fast increase and dynamic nature of the web, it has become one of the main challenges to traverse all URLs in the web documents and to handle these URLs. At this point we try to analyze the advantages of ontology, which is used to represent user profiles in personalized manner. In order to represent the user profiles, many models have utilized only knowledge from either a google information repository (GIR) or sometimes from local information repostitory (LIR). In this paper, we for the first time want to propose a new personalized ontology model is proposed for knowledge representation and reasoning over user profiles. This proposed model use google information repository (GIR) or sometimes from local information repostitory (LIR) for extracting the useful information. By conducting various experiments on our proposed model we finally came to a conclusion that personalized ontology model is best suited for extracting the exact information from web rather than similar data from the web.*

Key Words*: Ontology, Data Extraction, Global Knowledge Repository, Data Repository, Local Information Repository.*

I. Introduction

Now a day's world wide web (www) has rapidly increased its users from the past decades. In the past decades the information available on world wide web has exploded rapidly with a great range of topics and different categories. One thing that remained as a major challenge is how to collect the required information from the web. There are many search engines that are available now in order to extract the useful information for the given search keyword. And almost each and every individual search engines return more than 1,600 results per user query, where only thirty to forty percent of links are relevant or related to the user search query and remaining all are somewhat ir-relevant information that will be matched and displayed.

The world knowledge base (WKB) and a user's local instance repository (LIR) are used in the proposed model. The WKB is nothing but collecting the information or search query related data from the live data base like google, and for this we need to have internet connection for extracting any data related to user search query keyword. From that WKB we try to construct the personalized ontologies.Here the Local Information Repository (LIR) is mainly used for extracting the user background knowledge i.e. area of interest of that user who got registered for extraction of data. Here the proposed ontology model is evaluated by comparison against various benchmark models. The evaluation results show that the proposed ontology model is successful in design of personalized web information gathering systems.

II. Background Work

In this section we mainly discuss about the background work that was carried out in identifying the ontology learning environment. Now let us discuss about the ontology learning environment in detail as follows

Ontology Learning Environment

Ontology learning environment is nothing but dividing the extraction procedure into two phases.

1) User Background Knowledge Extraction
2) On Topic Keyword Extraction

Here in the first stage, the user need to create a profile in order to store his personal details along with area of interest and once if the user try to store his basic profile details along with area of interest. Then the user background knowledge can be extracted from that profile information.

Once the user background knowledge is extracted now the ontology learning environment try to find out the possible links that are relatively matched with the on topic search keyword. In this stage the ontology learning environment divides the search query into individual parts and then search process is done based on each and every individual search keyword.In this level all the links which are matched with the search keyword are extracted and send to the OLE environment in which the OLE try to apply a filter between the stage 1 and stage 2.

At this point, those links which are having exactly matching with stage 1 user background knowledge and the topic result from the stage 2,those will be filtered and send to the web mining results and those which are partially or not matched in any case will be ignored at this level.

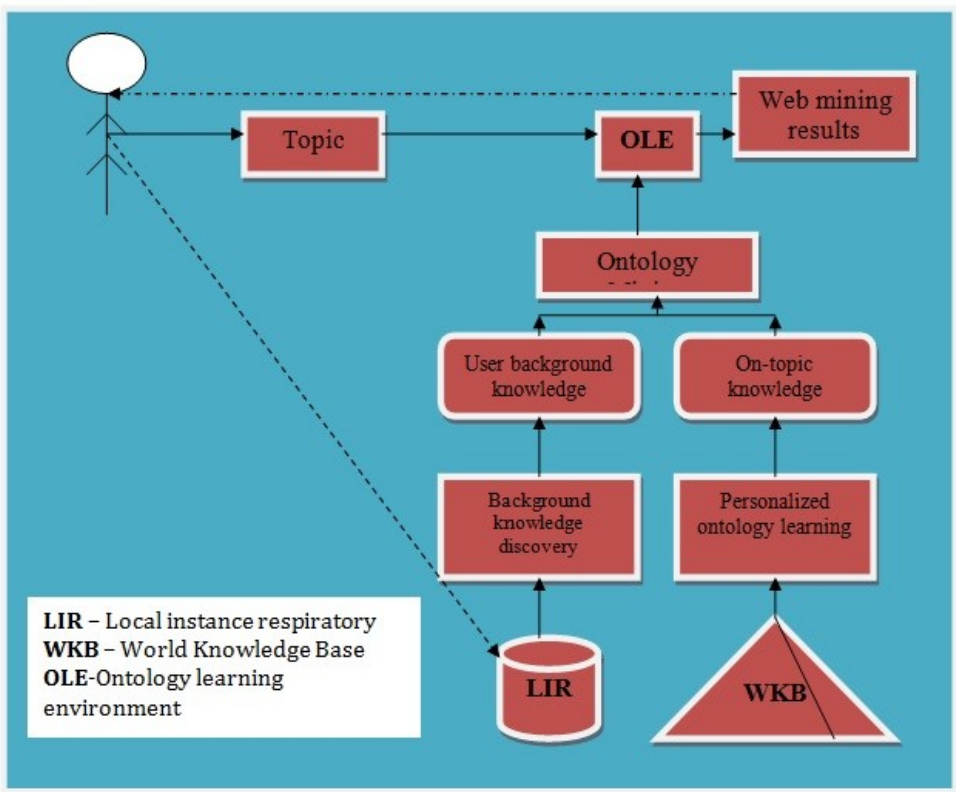

Figure 1. Denote the Flow of Personalized Ontology Model

From the above figure 1, we can clearly get an idea that our proposed personalized ontology model has two sources for extracting the information. One is LIR (Local Information Repository) and other is WKB (World Knowledge Base).From the above figure we can clearly find out that a user gives a topic or keyword as a search query and then the query enters into OLE (Ontology Learning Environment) for extracting the related data. Now the OLE will undergo the process of ontology mining by the two sources like LIR and WKB.From the LIR the user's AOI or area of interest is been extracted and from the WKB the search input query is extracted from the google[8].Once the query data which is extracted from WKB directly come to the buffer reader where the main logic is performed at the buffer area. In the buffer area all the query related links will be extracted and they will be filtered with the matched user interest field which was specified by the users at the time of registration. Now both the fields will be undergo the process of filtering and only the links which match will both LIR and WKB are extracted and all other links will be discarded at the buffer level. This is the way how each and every user individual area of interest is taken into consideration for extracting the valuable information from the web[9].

III. Preliminary Knowledge of Personalized Ontology Model

In this section we will mainly discuss about the preliminary information about the personalized ontology creation for extracting the useful or related data for the given area of interest. Now let us discuss about the personalized ontology creation in detail.

WORLD KNOWLEDGE REPRESENTATION

World Knowledge Representation is the process of extracting the valuable information from a large data source. World knowledge is mainly composed with knowledge possessed by people and acquired through their experience and education. Also, the term "world knowledge" is nothing but used for lexical and referential analysis.As we know that in our proposed application the WKR is nothing but user query keyword is send to the google server and in turn it will try to extract the related information from that WKB[10].

ONTOLOGY CREATION

General the user interactions are mainly extracted from web via user interests. For extracting the user interaction, we need to develop an ontology model to assist the users. Generally all the interactions are extracted in the form of URLs ,where the URLs are of two types: Positive URL and other is negative URL. Thus for a given topic we have both positive URLs and negative URLs of equal size.So when a request request a search topic, he will get both set of positive and negative URLs.We mainly concentrate on crawling or finding only the positive and related URLs rather than all ir-related URLs.Hence such a type of extraction is most related for ontology creation.

A pattern is a character string. All keywords can be written in both the upper and lower cases. It is used to extract hidden information from not-structured or semi-structured data. This aspect is fundamental because much of the web information is semi-structured due to the nested structure of HTML code, much of the web information is linked, and much of the web information is redundant. It should not include images, tags, and buttons. The extracted content should be stored in some file. But it should not include any HTML tags. The constructed ontology is personalized because the user selects positive subjects for personal preferences and interests as by selecting do-main names. This model is developed for four domains as–

1. General
2. Health
3. Education
4. Entertainment

It also allows entering 4 URL addresses at a time which pro-vides parallel processing for finding relative URL's.It also avoids time delay since providing parallel processing of input. It also counts every time how many URLs are searched at once, type of protocol, Hash code, web page content ,time of download. It also maintains local database which is used when user is offline and world knowledge base is searched when user is online.

IV. Implementation Phase

Implementation is a stage where the theoretical design is converted into programmatically manner. Here in this stage the application is mainly divided into several modules and each and every module differs from one another. In this proposed thesis, we try to divide the application into following four modules, they are as follows:

1. User Profile Creation & Authentication
2. Gathering Local & Global Information
3. Decompose The Categories
4. Comparison Between Web Search And User Profile

1. User Profile Creation & Authentication

In this module we describe create about user profile. When user comes first time for browsing we will create new profile to that particular user. Here in our proposed application the user profile is created with all his/her basic details along with a main field like AOI (Area of interest).Once the user profile is created he need to login with his valid id and password in order to search the output in a desired manner.

2. Gathering Local & Global Information

In this module we will collect the local and global information's about user. Local knowledge based on local instance repository, Here the user input query also known as ontopic is searched or extracted from the GIR database and the relevant AOI field of that searched user will be extracted from the LIR.

3. Decompose the Categories

Once the LIR and GIR information is extracted, now the personalization is done between LIR and GIR.Once the process on ontology learning is finished, now the desired result is decomposed into the list with all useful and exact links that are matched with on-topic keyword. Those which are not matched with the AOI are treated as separate list and they will be decomposed into separate list

4. Comparison between Web Search and User Profile

In this module we will do comparisons between user personal profile and real time web search result. So we will get different result to different user. When using a search engine, users typically formulate ambiguous queries which contain between one to three key-words. The search results that are returned from the search engine may satisfy the search criteria but often fail to meet the user's search intention.

V.Conclusion

In this proposed thesis we finally implemented a personalized ontology model for representing user background knowledge (I.e. individual area of interest of user's) 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.As an extension we also included the concept of rating chart for the proposed application which will clearly show the number of users who were interested in extracting the appropriate information based on their individual area of interest. This rating chart will tell the performance of our proposed application with respect to their individual area of interest.

VI. References

1. R R.Navigli, P.Velardi, and A.Gangemi, ?Ontology Learning and Its Application to Automated Terminology Translation, IEEE Intelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.
2. J.D. King, Y. Li, X. Tao, and R. Nayak, ‚Mining World Knowledge for Analysis of Search Engine Content,? Web Intelligence and Agent Sys-tems, vol.5, no.3, pp.233-253, 2007.
3. Trajkova and S. Gauch, ?Improving Ontology-Based User Profiles,Proc. Conf. Recherche d'Information Assistee par Ordinateur (RIAO '04), pp. 380-389, and 2004.
4. N. Zhong, ?Representation and Construction of Ontologies for Web Intelligence, Int'l J. Foundation of Computer Science, vol. 13, no. 4,pp. 555-570.
5. T. Tran, P. Cimiano, S. Rudolph, and R. Studer, ‚Ontology-Based Interpretation of Keywords for Semantic Search,? Proc. Sixth Int'l Semantic Web and Second Asian Semantic Web Conf. (ISWC '07/ASWC '07), pp. 523-536, 2007.
6. 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.
7. https://grakn.ai/pages/documentation/building-an-ontology/basic-ontology.html#identifying-entity-types
8. G. M. Voorhees and Y. Hou, ‚Vector Expansion in a LargeCollection,? Proc. First Text Retrieval Conf., pp. 343-351.
9. 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.
10. 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.