Volume 3, No. 7, July 2012

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH ARTICAL

Available Online at www.jgrcs.info

A NOVEL APPORACH FOR WEB INFORMATION GATHERING FROM CLOUD HOST USING WEBMINING ONTOLOGY

*1Mr. Saket Jain, 2Prof.Vineet Richharia and 3Mr.Abhishek Jain
*1Scholar,M,Tech (CSE), LNCT, Bhopal
Saketjains0071@yahoo.in
2HOD (CSE), LNCT, Bhopal
vineet_rich@yahoo.com
3Asst.Prof., LNCT, Bhopal
ajaries07@gmail.com

Abstract- Ontology defines a set of representational primitives with which a domain of knowledge is modeled. Cloud-based computing is an emerging practice that offers significantly more infrastructure and financial flexibility than traditional computing models. Ontologies are playing very important part in many areas such as intelligent information retrieval, knowledge management and organization, electronic commerce. Today, search engine crawlers are retrieving billions of unique URL's or web page. The main purpose of the Semantic web and ontology is to integrate heterogeneous data and enable interoperability among disparate systems. As the ontology is a standard, explicit and formalized description for shared conceptual model, the educational enterprises can be integrated by applying semantic express, shared knowledge described by ontology and automatic inference mechanism in educational institute management.

Keywords- unique documents, detecting replicate, replication, search engine.

INTRODUCTION

The increasing volume of data available on the Web as well as data's lack of structure, multidimensionality, large volume and dynamic evolution make information retrieval a tedious and difficult task. There are two ways to improve the quality of web mining, one is to use better mining technology for the existing resource and the other is to make the web resource more machine understandable for computer to process [1].

The last decades, 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. As introduced earlier the term "Semantic Web" encompasses efforts to build a new WWW architecture that enhances content with formal semantics. This will enable automated agents to reason about Web content, and carry out more intelligent tasks on behalf of the user. The relation between "ontology", "metadata" and "Web documents". It depicts a small part of the CIA world fact book ontology. with semantic annotations given in an XML serialization of RDF-based metadata descriptions. For the country and the organization there are metadata definitions denoted by corresponding uniform resource identifiers The URIs are typed with the concepts COUNTRY and ORGANIZATION. In addition, there is a relationship instance between the country and organization.

This scenario is common to many people on the World-Wide Web. A major problem with searching on the Web today is that data available on the Web has little semantic organization beyond simple structural arrangement of text, declared keywords, titles, and abstracts. As the Web expands exponentially in size, this lack of organization makes it very difficult to efficiently glean knowledge from the Web, even with state-of-the-art natural language processing techniques, index mechanisms, or the assistance of an army of data-entry workers assembling hand-made Web catalogs. In short, there is no effective way use the World-Wide Web to answer a query like: The chief intent of HTML and HTTP is to assist user level presentation and navigation of the Internet; automated search or sophisticated knowledge-gathering has been a much lower priority. Given this emphasis, relatively few mechanisms have been established to mark up documents with useful semantic information beyond document oriented information like "abstract" or "table of contents". Text indices suffer because they associate the semantic meaning of web pages with actual lexical or syntactic content From our previous example, searching a keyword index under "Laptop" yielded tremendous numbers of web pages, almost none of which are about living people named Laptop. "Laptop" has many uses besides being a last name. Although text indices are improving, the amount of information on the Web is also growing rapidly.

A major disadvantage of hand-built catalogs is the man hours required to construct them. Given the size of the World-Wide Web, and the rate at which it is growing, cataloging even a modest percentage of web pages is task. Additionally, the criteria used in building any catalog may turn out to be orthogonal to those of interest to a user.

Ad-hoc robots that attempt to gather semantic information from the web typically gather only the limited semantic information inferable from existing HTML tags. The current state of natural language processing technology makes it difficult to infer much semantic meaning from the body text itself at a reasonable rate (if at all). In our experience, even limiting a web robot's natural language understanding to a small topic like Computer Science Web pages still proves surprisingly difficult to implement, and like many ad-hoc methods, such algorithms are extremely brittle. Further, none of these approaches (except perhaps the last, for specific domains) allows for inferences about relationships *between* web pages, aside from simple facts about linkage. Sophisticated queries such as our initial "Laptop" example are therefore clearly out of reach.

Instead of trying to glean knowledge from existing HTML, another approach is to give authors the ability to embed knowledge directly into HTML pages, making it simple for user-agents and robots to retrieve and store this knowledge. The straight forward way to do this is to provide authors with a clean superset of HTML that adds a knowledge markup syntax; that is, to enable them to directly classify their web pages and detail their web pages' relationships and attributes in machine-readable form using HTML.

Using such a language, a document could claim that it is the home page of a graduate student. A link from this page to a research group might declare that the graduate student works for this group as a research assistant. And the page could assert that "Laptop" is the graduate student's last name. These claims are *not* simple keywords; rather they are semantic tags defined in some "official" set of attributes and relationships (an *ontology*). In this example the ontology would include attributes like "lastName", classifications like "Person", and relationships like "employee". Systems that gather claims about these attributes and relationships could use the resulting gathered knowledge to provide answers to sophisticated knowledge-based queries. Moreover, user-agents or robots could use gathered semantic information to refine their web-crawling process. For example, consider an intelligent agent whose task is to gather web pages about Laptop. If this agent were using a thesaurus-lookup or keyword-search mechanism, it might accidentally decide that Helena Laptop's web page, and pages linked from it, are good search candidates for this topic. This could be a bad mistake of course, not only for the obvious reasons, but also because Helena Laptop's links are to the rest of the University of Maryland (where she works). The University of Maryland's web server network is very, very large, and the robot might waste a great deal of time in fruitless searching. However, if the agent gathered semantic tags from Helena Laptop's web page which indicated that Laptop was her last name, then the agent would know better than to search this web page and its links.

LITERATURE SURVEY

Nizar R. Mabroukeh [2] With the advent of social networks and tagging systems, The Internet has recently witnessed a big leap in the use of Web Recommendation Systems WRS. Based on users' likings of items and their browsing history on the world wide web, these systems are able to predict and recommend items and future purchases to users. They are being used now in various domains, like news article recommendation, product recommendation, and make-friend recommendation. WRS are still limited by several problems, of which are sparsity, and the new user problem. They also fail to make full use and harness the power of domain knowledge and semantic web ontologies. In this article, we discuss how an ontology-based WRS can utilize relations and concepts in an ontology, along with user-provided tags, to provide top-n recommendations without the need for items clustering or user ratings. For this purpose, we also propose a dimensionality reduction method based on the domain ontology, to solve the sparsity problem.

A content-based web recommendation system is proposed based on a domain ontology. It relies on user-provided tags, that are mapped to concepts of this ontology. Similarity measures are used during mapping and a matrix of items concepts is built offline, which is used later for online top-n recommendation. their system outperforms popular algorithms like Top Pop and NNCosNgbr. In Addition, a proposed novel dimensionality reduction solves the sparsity problem, and does not compromise the accuracy of the proposed system. they also show how the recommendation set is expanded using Spreading Activation over the ontology, taking into consideration the several available relations, which raises the accuracy of the proposed model.

There are two benefits in using an ontology over clustering of the tags. First, it saves the costly step of clustering, and second, a full ontology has a far better reasoning power than a topic taxonomy. In a full ontology there are several semantic relations that can be taken into consideration (as opposed to only is-a relation in a topic taxonomy) to provide better relatedness measures, and better interpretability. In addition, a similarity measure can be formulated that uses relation hierarchy for recommending only highly similar concepts.

Xiaohui Tao, Yuefeng Li et. al. [3] 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 a user local information. they, 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.

Xujuan Zhou, Sheng-Tang et. al. [4] It is well known that taking the Web user profiles into account can enhance the effectiveness of Web mining systems. However, due to the dynamic and complex nature of Web users, automatically acquiring worthwhile user profiles was found to be very challenging. Ontology based user profile can possess more accurate user information. their research emphasizes on acquiring search intentions information. they presents a approach of developing user profile for Web searching.

The model considers the user's search intentions by the process of PTM (Pattern-Taxonomy Model). Initial experiments show that the user profile based on search intention is more useful than the generic PTM user profile. Developing user profile that contains user search intentions is essential for effective Web search and retrieval.

Integrating ontology-based user profiles into the processing can be very beneficial for improving the efficiency of Web information search and retrieval.

Considering the Web user's search intention will assist building a more useful user profile. PTM is able to provide rich semantic relationship between the patterns. With the Web user's search intention derived from PTM, the new method shows a considerable improvement in terms of search effectiveness. This demonstrates that the user profiles based on the search intention can improve the information retrieval performance.

Yuefeng Li and Ning Zhong et. al. [5] It is not easy to obtain the right information from the Web for a particular Web user or a group of users due to the obstacle of automatically acquiring Web user profiles. The current techniques do not provide satisfactory structures for mining Web user profiles. they presents a novel approach for this problem. The objective of the approach is to automatically discover ontologies from data sets in order to build complete concept models for Web user information needs. It also proposes a method for capturing evolving patterns to refine discovered ontologies. In addition, the process of assessing relevance in ontology is established. they provides both theoretical and experimental evaluations for the approach. The experimental results show that all objectives they expect for the approach are achievable.

There is no doubt that numerous discovered patterns can befound from the Web data using data mining techniques. However, it is ineffective to use the discovered patterns in Web user profile mining due to the ambiguities in the data values (terms). The consequent result is that they obtain some inappropriate discovered patterns and many discovered patterns include uncertainties. they develop an ontology mining technique to provide a solution for this challenge. A discovered ontology in this research consists of two parts: the top backbone and the base backbone. The former illustrates the linkage between compound classes of the ontology. The latter illustrates the linkage between primitive classes and compound classes. they set up a mathematical model to represent discovered knowledge on the ontology. they also present a novel method for capturing evolving patterns in order to refine the discovered ontology. In addition,

The research is significant for WI since it makes a breakthrough by effectively synthesizing taxonomic relation and nontaxonomic relation in a mathematical model. It is also significant for data mining because it provides an approach for representation, application, and maintenance of discovered knowledge for solving real problems.

PROPOSED TECHNIQUE

Web mining is the use of data mining technologies to automatically interact and discover information from web documents, which can be in structured, unstructured or semi- structured form. XML has become very popular for representing semi structured data and a standard for data exchange over the web. The data based on XML is self described; it can be exchanged and handled without internal description. The web has become a major vehicle in performing research and education related activities for researches and students. There is tremendous amount of information and knowledge existing on the web and waiting to be discovered, shared and utilized. Ontology represents a set of precisely defined terms about a specific domain and accepted by this domain's community, ontology is an explicitly specification of a conceptualization. The RDF is a simple meta model for defining and exchanging information on the semantic web. We present an enterprise web framework regarding semantic web and mining in training institute, which can be used to not only improve the quality of web mining results but also enhances the functions and services and the interoperability of educational information systems and standards in the educational field. Mining the educational information on the web we are using new Semantic Web Mining technologies, such as Resource description Framework (RDF) and Web Ontology Language (OWL). Nowadays, the Web is rapidly growing and becoming a huge repository of information, with several billion pages and more than 300 million of users globally.

Indeed, it is considered as one of the most significant means for gathering, sharing, and distributing information and services. At the same time this information volume causes many problems that relate to the increasingly difficulty of finding, organising, accessing, and maintaining the required information by users. All these have affected greatly the way web-based applications are designed and implemented and e-Learning systems could not comprise an exception. Besides, among all other "e" movements, e-Learning is one of the fastest growing and universally accepted.For on line educational institute web site two important ontology's would need to be built one ontology describing all the educational services provided, with the relation between each other and the other ontology describing the web site. Thus semantic web ontology help build better web mining analysis in educational institute and web mining in-turns helps contract basis more powerful ontology in education. We propose a framework for personalised e-Learning based on aggregate usage profiles and domain ontology. We have distinguished two stages in the whole process, one of offline tasks that includes data preparation, ontology creation and usage mining and one of online tasks that concerns the production of recommendations.

The proposed system starts with mapping the click tags in a preprocessing step. Pages and their associated tags are stored in a database. Then, Wu and Palmer similarity measure [6] is used with WordNet6 as a thesaurus to compute the similarity between tags of each pages and each concept in the ontology, giving the similarity score sim(Tpi ; cj). To elaborate, each tag in pi is compared against each concept in the ontology, this is done by computing the similarity between the tag and the concept, both the tag and the concept are located as two words in WordNet (say n1 representing concept cj , and n2 representing a tag g from the set Tpi) and the Wu and Palmer.

Click Tags mapping: The web log is represented as click tags containing pairs of pages and their associated tags. Similarity is computed between each pages and each ontology leaf concept, and all similarity scores are stored in a matrix of pages concepts. Notice that the dimensionality of the matrix depends on the number of leaf concepts in the ontology.

Active user: As the active user arrives at a certain web page (or buys a certain product), the tags associated with this pages are retrieved, and a vector is generated, that is similar to one row of the online *WebPages database I_{db}* generated in the previous step. The vector shows the similarity between the active user tags and each concept. This vector is matched against each row in the matrix, and the top-n matching pages are used as the recommendation set.

Expanding data set: The recommendation set can be expanded by increasing n, and by expanding the active user vector using semantic relations in the ontology to include more concepts, not present in the matrix, from which recommendation can be drawn.

PROPOSED ALGORITHM

Algorithm:

```
Check web site validation
Connect web host
    Analysis of web hosted database
    WebPages database $I_{db}$,
    Domain ontology $O_d$,
        Check web base tags
        For each tags <Page $p_i$, its tags $Tp_i$ > in $I_{db}$
        For each concept $C_j$ in $O_d$
        Calculate Similarity ($Tp_i$; $C_j$ ), which is the average similarity between
        Each tag in $Tp_i$ and $C_j$
          Check data downloadable
            if <Page $p_i$, its tags $Tp_i$ > is valid then
              if <Page $p_i$, its tags $Tp_i$ > is duplicate then
              Leave date
              Else
              Download in result in $R_d$,
              $R_d$ [i; j] = Similarity ($Tp_i$; $C_j$ )
              End if
            End if
        End for
        End for
    Return $R_d$
Set top n recommended Page, $P_S$
Active user vector, $A_u$
Extended set, S
    Find the concept $C_j$ to which $A_u$ is highly similar
    Rank all relations of $C_j$ according to their relation hierarchy
    For each relation connecting $C_j$ with another concept $C_k$ do
        Instantiate concept $C_k$ to generate Pages
        Add pages to $P_S$, to get S
        Store page rank in $PR_i$
        End for
        Return S+
End
```

CONCLUSION

Web mining takes aspects of data mining and text mining and brings them together in the context of the world biggest information resource. The WWW web mining finds the info from web sources, may be from data warehouse and from own communal database. There is tremendous amount of information and knowledge existing on the web and waiting to be discovered, shared and utilized. Students, faculties, researchers require a lot of information about education and research activities.

Web mining is the use of data mining techniques to automatically discover and extract information from the web documents which can be structured, unstructured or semi structured from. XML has become very popular for representing semi structured data and a standard for data exchange over the web. The data based on XML is self described; it can be exchanged and handled without internal description. The core technique of semantic web mining is ontology. Ontology represents a set of precisely defined terms about a specific domain and accepted by this domain's community, ontology is an explicitly specification of a conceptualization.

REFERENCE

[1] Edward H.Y. Lim, Hillman W.K. Tam, Sandy W.K. Wong, James N. K. Liu and Raymond S. T. Lee,Collaborative Content and User-based Web Ontology Learning System, IEEE, 2009, Pg 1050-1055

[2] Nizar R. Mabroukeh and C. I. Ezeife, “Ontology-based Web Recommendation from Tags”, IEEE ICDE Workshops 2011

[3] Xiaohui Tao, Yuefeng Li, and Ning Zhong, "A Personalized Ontology Model for Web Information Gathering", Published by the IEEE Computer Society 2011

[4] Xujuan Zhou, Sheng-Tang Wu, Yuefeng Li, Yue Xu, *Raymond Y.K. Lau, Peter D. Bruza, "Utilizing Search Intent in Topic Ontology-based User Profile for Web Mining", IEEE/WIC/ACM International Conference, 2006

[5] Yuefeng Li and Ning Zhong, "Mining Ontology for Automatically Acquiring Web User Information Need", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 4, APRIL 2006

[6] Z. Wu and M. S. Palmer. Verb semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138. Association for Computational Linguistics, 1994.