Agent Based Framework for Semantic Web Content Mining

Aarti Singh
Associate Professor
MMICT&BM, M.M.University, Mullana, Haryana, India

Author Email: **singh2208@gmail.com**

Abstract

With flooding of information on WWW it has become necessary to apply some strategy so that valuable knowledge can be extracted and consequently returned to the user. Data mining techniques find their applicability in such scenario. Data mining concepts and techniques when applied to WWW with its existing technologies are termed as web mining. Web mining can change the way results are provided to user queries presently i.e. ranked list of keyword based results. This work focuses on proving agent-based framework for mining semantic web contents employing clustering techniques. Clustering will help provide user with query relevant cluster of web contents, which will better satisfy user requirement and will provide optimal utilization of web surfing time.

Keywords: *Web Mining, Content Mining, Multi-agent Systems, Hierarchical Clustering, OLAP.*

1. Introduction

Web mining [1] can be defined as mining of the World Wide Web (WWW) to find useful knowledge about user behavior, content, and structure of the web. It involves application of data mining techniques on the contents of WWW but is not limited to it. Tremendous growth of web-sites and text and multimedia contents on the WWW has lead to demand of a strategy which could provide knowledge from the vast data scattered over different servers and also could make useful predictions for otherwise uncertain user behavior. Web mining [3] is uniquely different from data mining as it works on web contents that are unstructured files or server logs in contrast to well-structured databases used in data mining. Web mining is defined as strategy for mining pattern P(UB) given the collection of web contents (wc), thus it is mapping M:wc->P(UB), where UB is user behavior. Here the user may be the consumer of the data & services and may be the contributor to it.
In literature [2,3,7] web mining is categorized as:

- ***Web Structure Mining***: is the technique to analyze and explain the links between different web pages and web sites. It mainly focuses on developing web crawlers. It works on hyperlinks and mines the topology of their arrangement.
- ***Web Content Mining***: focuses on extracting knowledge from the contents or their descriptions. It involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior.
- ***Web Usage Mining***: It focuses on digging the usage of web contents from the logs maintained on web servers, cookies logs, application server logs etc. It works on how

and when user moves from one type of content to other. Thus, it can provide association between different contents.

Figure 1 given below highlights the classification of web mining.

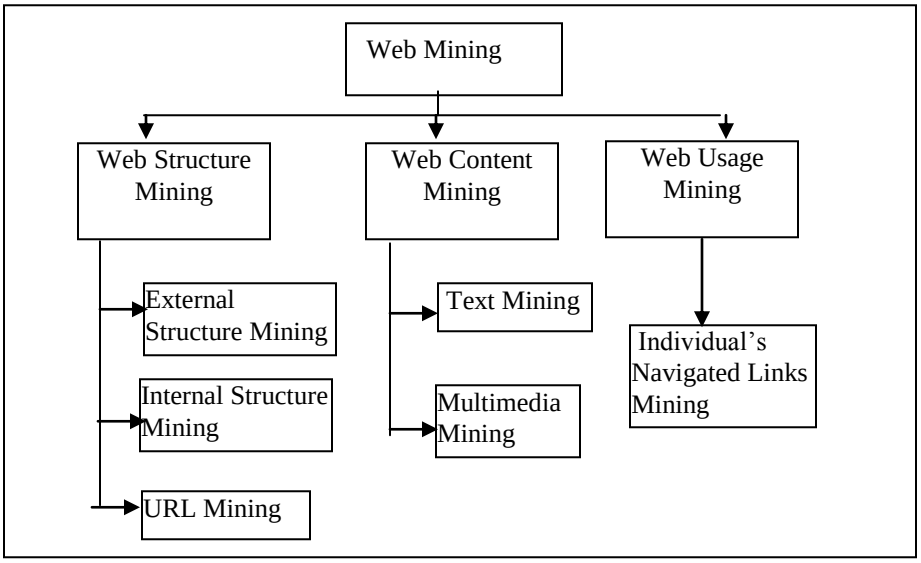

Figure1. Classification of Web Mining

In the present generation of WWW, the user is more interested in getting useful, relevant and knowledge oriented contents from the WWW. The paradigm is shifting from demand of information to demand for knowledge. The WWW is transforming into Semantic web which is knowledge oriented. Still we are away from complete realization of semantic web; web content mining when applied on semantic web contents can lead to discovery of knowledge that could be provided to end users to better serve their requirements. This work aim to propose agent based framework for mining contents of semantic web, which would provide query relevant knowledge using clustering technique. The paper is structured as follows: section 2 provides literature review, section 3 explains the proposed framework, and section 4 concludes the paper.

2. Literature Review

This section explores the literature on web mining and present status of employment of agents in it.

Sharma et. al [1], Kosala et. al [3] and Eirinaki et. al [4] provided detailed review on web mining focusing on different dimensions of this field. [1] highlighted use of cloud computing in web mining , [3] focused on scope of agent technology in it whereas [4] provided details on web personalization through web mining. Bhatia et. al in [2] provided semantic web mining and suggested an ontology learning mechanism for the extraction of semantics through grammatical rule extraction technique. Meirong et. al in [5] proposed an agent based web mining model for e-buisness. Zhan et. al in [8] provided a multi-agent module working as knowledge crawler. Ting H.I. in [6] employed web mining for on-line social network analysis, however strategy for selecting appropriate sample size to reflect exact real social networks and

actual implementation is left as future research. Jicheng et. al in [7] proposed an agent based web text mining system for mining HTML based documents on the web, however it still lacks efficient algorithm for very large document collections and use of XML specifications.

Literature review highlighted the fact that agent based systems have already been employed in various area of semantic web due to their promising features. Dimou et. al. [9] developed an agent based framework called Biospider for developing and testing autonomous,intelligent & semantically focused web spiders. The framework takes the advantage of agent technology in distributing crawling load to a number of cooperating
spiders. Buccafurri et al. [10] proposed an agent-based recommender system based on a concept-graph model that represented user-behavior-dependent relationships among concepts. Singh et. al in [11] proposed an ontology agent based focused crawler (O-ABFC) which improves existing agent based focused crawlers by using ontology and contextual information in crawling. Use of ontology is emerging as a promising tool that eliminates simple keyword based crawling method as it introduces semantics or contexts in which a keyword is searched. Singh et. al in [12] proposed an intelligent & adaptive ontology mapping mechanism for providing an interface that facilitates agent interaction in homogenous as well as heterogeneous ontologies. Their work automates the ontology-mapping task using multi-agent system that not only overcomes the curse of already existing mapping mechanisms but also is time efficient.

Critical review of literature highlights this fact that agent technology has widely been employed in semantic web applications at various fronts and researchers have agreed on its applicability for mining semantic web contents. Although some efforts had already been made to propose application specific agent based solution in diverse areas like e-business[5] or for social networking[6], but there is no standard framework for semantic web content mining. Thus, there is scope of research in this direction. Upcoming section elaborates our proposed framework.

3. Proposed Work

This work proposes agent based Semantic Web Mining System (SWMS) which will provide classification and clustering of the web contents, thereby facilitating knowledge based response to the user and will highlight otherwise unnoticed patterns.
Figure 2 given below provides the high-level view of SWMS. It mainly comprises of Interface agent, collection agent supported with ontology database, content mining agent and clustering agent. Content mining agent works in collaboration with descriptive metadata agent and semantic metadata agent. Ecology of the agents contained is as follows:

- ***Interface agent (IA):*** It works as an interface between the search engine and the SWMS. It receives query given by the user and passes it on to the collection agent for match of relevant results. On receiving results from the collection agent, it passes it onto search engine for providing output to the user.
- ***Collection agent (CLA):*** Collection agent receives input query from the interface agent and explores ontology database for the meaning of the keywords or the context based meaning of the phrase. Once it is clear in which context user is searching for the information, content mining agent and the clustering agent are invoked to get suitable results.

- ***Content Mining agent (CMA):*** This agent works in coordination with indexes maintained by the search engines, further refining the information listed in indexes to extract knowledge from it. It focuses on the metadata contained in every document, which contains description of the contents known as descriptive metadata and information illustrating meaning/context of the content known as semantic metadata. It visits server indexes periodically to explore new contents and passes this information to descriptive metadata agent and semantic metadata agent for further handling.
- ***Descriptive Metadata agent (DMA):*** DMA is responsible for extracting the descriptive information such as title, date, size, type of the file etc. It maintains a table recording this information, on which text mining techniques are applied by CMA to extract useful knowledge, such as how many new web pages/files have been uploaded in a particular area in a specific year?
- ***Semantic Metadata agent (SMA):*** SMA focuses on recording semantic features of a document such as author name, context of document, organization concerned (if any) or domain of work. This information is recorded in semantic metadata table and is mined to obtain useful knowledge/pattern such as more addition of files in a specific context shows more research/development inclination of users in that area. Similarly, least attended area can also be discovered.
- ***Clustering Agent (CUA):*** Clustering agent works on the tables maintained by the DMA and SMA. It creates various clusters of the indexed documents such that inter cluster similarity is minimized and intra cluster similarity is maximized [7]. Clustering is different from text categorization or classification in the way that there are predefined classes in which documents have to be placed. Clustering does not follow any predefined taxonomy rather clusters emerge from the characteristics of the documents on their own. Clustering agent makes use of hierarchical clustering algorithm [14] for this purpose.

Apart from these agents, ontology database is an important component that supports the overall objective of returning context relevant knowledge to the users.

- ***Ontology Database:*** Ontology is defined as well organized knowledge scheme that represents high-level background knowledge with concepts and relations. Ontology based crawling [9] eliminates simple keyword based crawling method as it introduces semantics/context in which a keyword is being searched thus improving crawl efficiency. Most existing ontology focused crawlers use ontology as background knowledge and apply weights of concepts in the ontology to compute the relevance score (reader interested in design details of ontology database should refer [13]). The comprehensiveness of the ontology database can ensure context based information retrieval.

Content mining agent (CMA) periodically visits indexes maintained at different servers and provides the newly added documents to DMA and SMA, which update their table by recording appropriate features. CMA performs mining on these tables by using text-mining tools to get knowledge about the recorded documents. Let us consider, if an author name appears with 50 different files then it can derive pattern of authors work/interest. If the context of those files span across semantic web, agent technology, fuzzy logic and neural networks then CMA can draw the conclusion that author's field of work is artificial intelligence and thus whenever there is some query for artificial intelligence papers, the

work of this author may also be listed as part of output. Ontology database will help in this kind of context generalization. This kind of knowledge will also help clustering agent in creating various clusters. Once CMA is finished with mining of the indexed files, CUA starts clustering process. It creates clusters in the form of multi-dimensional cubes supported by OLAP tools. For instance a cluster will contain one author name, all his publications, year of publications. This cluster will be a part of another cluster based on organization of the author and then one based on country or geographical location. These clusters may be rolled down to view all dimensions and may be rolled up to have broader look at the contents. Thus with the help of clustering it will be possible to answer queries like papers published on wireless sensors in India or by Indian authors. Since all such papers will lie in the cluster having geographical location as India.

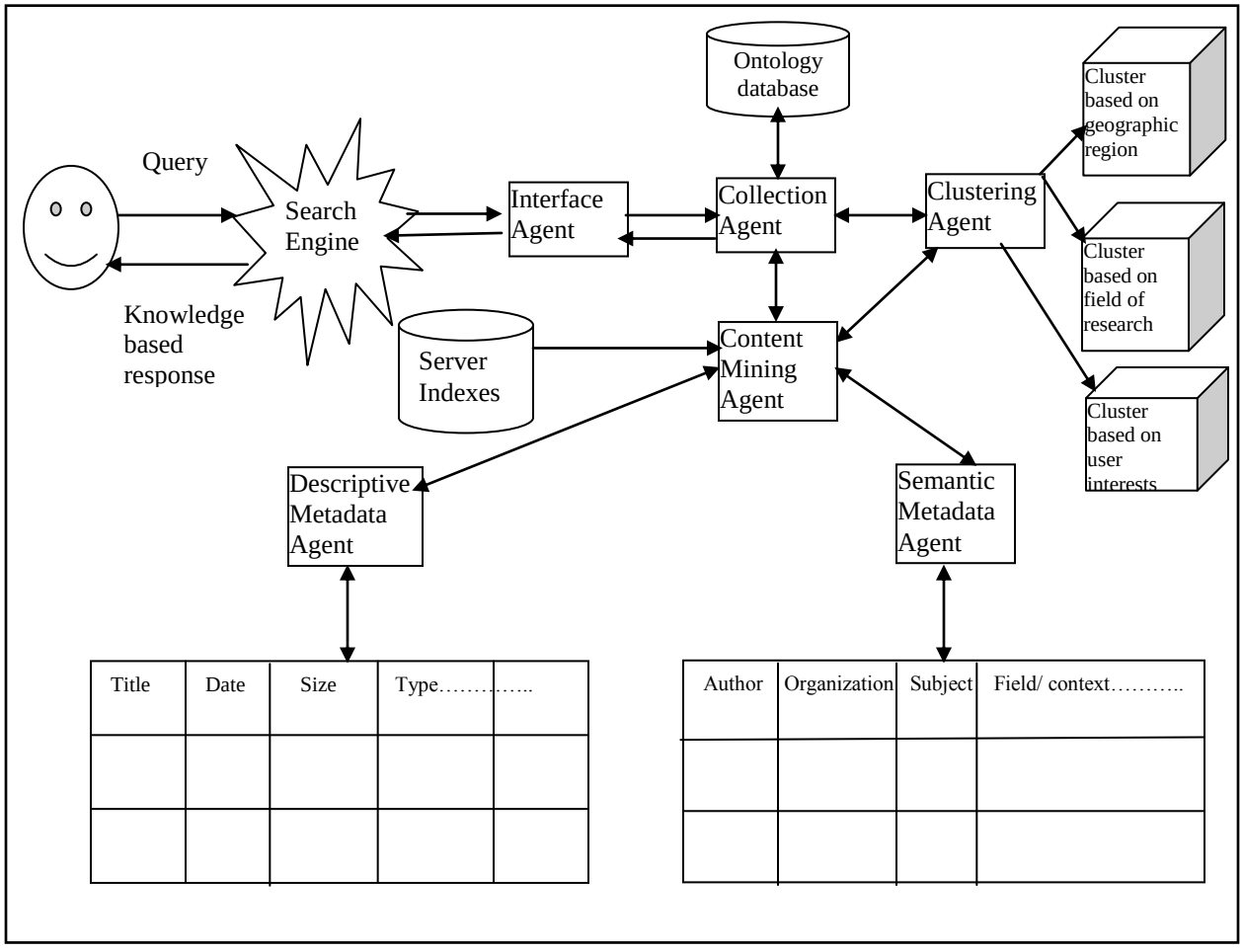

Fig. 2 High Level View of Semantic Web Mining System

Now if a user enters query say- Software agents by Wooldridge, with the help of SWMS, not only the same paper could be returned but also the user would be provided with the cluster containing other papers by Wooldridge. This kind of response will be more relevant to a

researcher interested in the related work by the same author which is otherwise unknown to the user.

4. Conclusion & Future Work

This work has proposed agent based solution for mining semantic web contents, with the aim to provide context based knowledge oriented results to the user. The next generation of WWW will be knowledge oriented and to satisfy the customers web mining is a promising solution. The amalgamation of web mining techniques with agent technology will lead to improved performance, reduced network traffic, and better results. However, implementation of this work is still under progress and is left as future work.

References:

1. Sharma K., Shrivastava G. & Kumar V., 'Web Mining: Today and Tommorrow'. In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.
2. Bhatia C.S. & Jain S., 'Semantic Web Mining: Using Ontology Learning and Grammatical Rule Interface Technique'. In IEEE 2011.
3. Kosala R. & Blockeel H., 'Web Mining Research: A Survey'. Published in ACM SIGKDD, Vol. 2, Issue 1,July 2000.
4. Eirinaki M. & Vazirgiannis M., 'Web Mining for Web Personalization'. Published in ACM Transactions on Internet Technology, Vol.3 , No. 1, February 2003, pp. 1-27.
5. Meirong T. & Xuedong C. , 'Application of Agent Based Web Mining in E-business'. Published in 2010 IEEE Second International Conference on Intelligent Human-Machine Systems and Cybernetics, pp. 192-195.
6. Ting I.H., 'Web Mining Techniques for On-line Social Networks Analysis'. In Proceedings of the 5th International Conference on Service Systems and Service Management, Melbourne, Australia, 30 June-2 July 2008, pp. 696-700.
7. Jicheng W., Yuan H., Gangshan W. & Fuyan Z., 'Web Mining: Knowledge Discovery on the Web'. In Proceedings of IEEE International Conference on System, Man and Cybernetics 1999 (IEEE SMC'99), Vol. 2 , pp. 137-141.
8. Zhan L. & Zhijing L., 'Web Mining based on Multi-Agents'. Published in proceedings of Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), 2003.
9. C.Dimou, A.Batzios, A.L.Symeonidis and P.A.Mitkas, 'A Multi-agent framework for Spiders Traversing the Semantic Web'. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.
10. F. Buccafurri, G. Lax, D. Rosaci and D. Ursino, 'Dealing with Semantic Heterogeneity for Improving Web Usage'. Data Knowledge Eng. Vol. 58, Issue 3, pp. 436–465,2006.
11. Singh A., Juneja D. and Sharma A.K., 'Design of Ontology-Driven Agent based Focused Crawlers'. In proceedings of 3rd International Conference on Intelligent Systems & Networks (IISN-2009),Organized by Institute of Science and Technology, Klawad, 14 -16 Feb 2009, pp. 178-181. Available online in ECONOMICS OF NETWORKS ABSTRACTS**,** Volume 2, No. 8: Jan 25, 2010.
12. Singh A., Juneja D., Sharma A.K., **'**Design of An Intelligent And Adaptive Mapping Mechanism For Multiagent Interface'.In Proceedings of International Conference on High Performance Architecture and Grid Computing Communications in Computer and Information Science (HPAGC'11), 2011, Volume 169, Part 2, 373-384, DOI: 10.1007/978-3-642-22577-2_51.
13. Singh A., Juneja D., Sharma A.K., 'General Design Structure of Ontological Databases in Semantic Web'. Published in International Journal of Engineering, Science & Technology, Vol. 2, Issue 5, pp. 1227-1232, 2010.
14. Karayannidis N. & Sellis T., 'Hierarchical Clustering for OLAP: The CUBE File Approach'. Published in The VLDB Journal — The International Journal on Very Large Data Bases, Vol. 17, Issue 4, July 2008.