Semantic Web Data Mining & Analysis

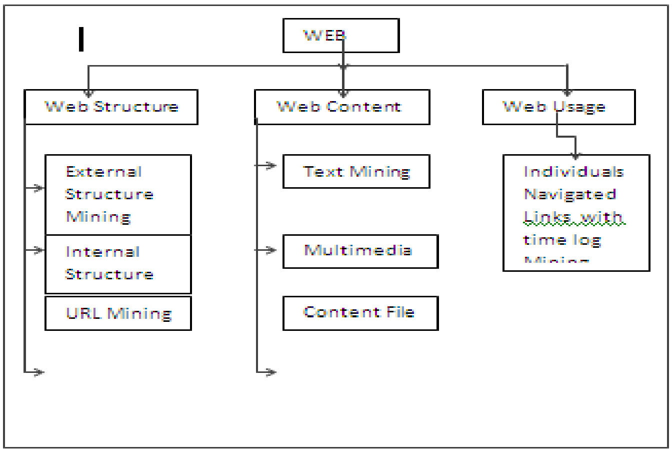
Abhishek Yadav[#1] Gaurav Srivastava[#2]
[#1]*MTECH(CSE), R.K.D.F,I.S.T, Bhopal. Affiliated to RGPV, Bhopal, M.P, India*
[#2]*R.K.D.F , I.S.T, Bhopal. Affiliated to RGPV Bhopal, M.P, India*

Abstract: *Semantic Web Mining combines two fast developing research areas: Semantic Web & Web Mining. In this relation, the research intension is to improve on the one hand, Web mining methods with new needs of semantic strategies and on another hand new strategic rule to make it fast and accurate. With tremendous development of WWW, it is making web experience more time spending to user. Hence semantic web mining has become necessary to apply some strategy so that valuable knowledge can be extracted and consequently returned to the user. Data extraction strategies and techniques when applied with web mining will provide a new way result to user query. Clustering will help to provide better satisfaction to user query with less surfing time.*

Key Words: *Web mining, data mining, semantic data mining, search query and surfing time.*

I. Introduction

Web mining can be defined as mining of the WWW to retrieve useful knowledge and data about user behavior, user query, content and structure of the web. In this paper, focus on processing of structured and unstructured data mining will take place. With tremendous development growth in website, web portal to provide downloadable data to user, required a lead to demand of a specific strategy to provide knowledgeable data to user and also useful to predict otherwise uncertain user behavior on the server. Semantic web is about machine-understandable web pages to make the web more intelligent and able to provide useful services to the users. This means that information on web pages may have to be mined so that the machine can understand the contents.



II. Literature Review

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. Meironget.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 andactual implementation is left as future research. Jichenget.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.

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. Singh et. Al[15] proposed the next agent based web mining but there is a scope to research in content file in contrast of unstructured data mining with concept of web. Multimedia mining already included in agent based web mining al[15] but the user timing log mining and file size mining can provide a better way to meet the requirements.

Literature review highlighted the fact that agent based systems have already been employed in various area of semantic web due to their promising features. Dimouet. 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.

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 written in 6different languages, then it can derive pattern of author name, context, geographical area of publishing (if any) and language in which it written. If context of those file 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 in user default language.

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.

Apart from this ontology database is more important that supports the overall objective of returning context relevant knowledge to the users.

A. 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.

III. Proposed Model Framework

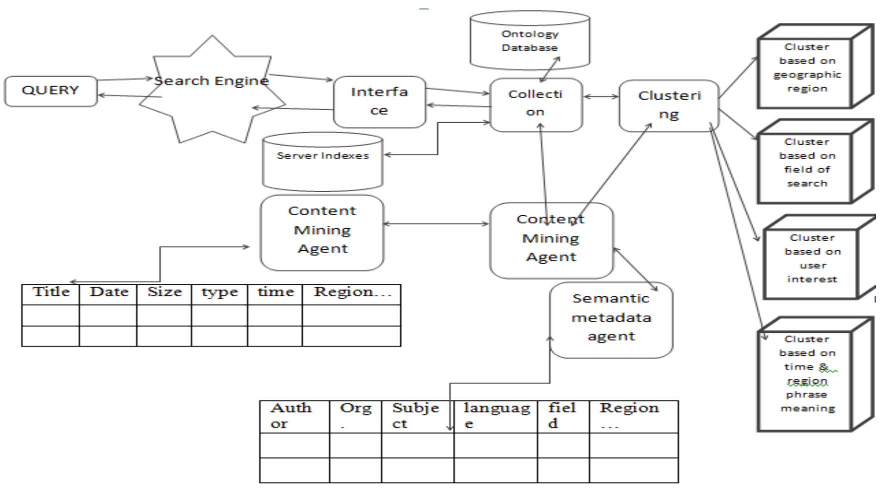
This framework proposes agent based Semantic Web Mining System (SWMS). It will provide classification and clustering of the web contents according to user navigating links and time when navigating to other pages, thereby facilitating knowledge based response to the user and will highlight otherwise unnoticed patterns. It mainly comprises of Interface Agents, 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.

Let us take an another example of web page searching like if user enters a query or phrase which contains the multiple meaning of that phrase. Ontology database will be searched for that query to mean and once meaning is derived it will hit to the DMA to descriptive metadata and then CMA, same process as above but here from user side, IA will send the information of place and top to bottom listing of the hitting link by user and ontology database will store it. And another time when users enters same query but hits on another link it will record the behavior of that site and next time it will proceed to provide correct result. As in example, A phrase entered "Hotel in India" and this was queried in morning or before noon and user clicked to hotel to stay in that hotel but after noon or in evening user queried same phrase but clicked on hotel for dinner, lunch. This will recorded and will be stored in ontology database and next time when user enters same phrase after noon o in evening it will show the topmost hotels ready to dinner or lunch according to user behavior in past.

A. Ecology of the agents

- Interface Agent(IA)**:** It works as an interface between the search engine and the SWMS. It receives the requested query from user and passes it to Collection Agent to provide compatible 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 query from interface agent and explores ontology database for the meaning of the keyword or context based meaning of phrase of the required query. Once, meaning cleared, it invokes the content mining agent and the clustering agent to get suitable result.
- Content Mining Agent (CMA):this agent works in co-ordination with indexes maintained by search engines, further process to get extract knowledgeable data from listed index. It focuses on metadata contains description of the contains known as descriptive metadata. Information illustrating meaning of the content known as semantic metadata. It visits server indexes periodically to check and update information added to it and passes this metadata to descriptive metadata agent and semantic metadata agent to mine it.
- Descriptive Metadata Agent (DMA):It is responsible to extract the descriptive information such as title, date, size, type of the file, geographical reason etc.It maintains the table data recording this information on which text mining techniques are applied by CMA to extract useful knowledge such as number of new web pages, file and types of these files are added. With this a new system will propose to search according type of file and search entered hitting time and priority according to hitting priority to any link.
- Semantic Metadata (SMA):SMA focuses on recording of semantic features of documents such as originator name, substance of document, organization concerned or domain of work. This propaganda 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 related to each line by line search in whole document pages and also discover the geographical area(if any) and languages in which document prepared according to user area.
- 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 stratified clustering algorithm [14] for this purpose.

High Level view of semantic data mining



Title	Date	Size	type	time	Region…

Auth or	Org .	Subje ct	languag e	fiel d	Region …

IV. Expected Outcomes

Here in our research we are experimentally going to provide a simulation of semantic web mining system where query will work with different type of conditions related to past behavior of query and ontology database and will help to provide accurate result.

V. Conclusion

This work has proposed agent based solution for mining semantic web contents, with aim to provide knowledgeable result based on time, context, region and language. The next generation of web will knowledge oriented but user dependent to provide accurate result according to particular phrase. 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 MultiagentInterface'.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.

[15] Aarti Singh, 'Agent Based Framework for Semantic Web Content Mining'. Published in International Journal of Advancements in Technology,Vol. 3 No.2 (April 2012), ISSN 0976-4860.