# Optimizing Web Mining using Multi-agent System

Shah Yesha B.
Pursuing M.E.,
Parul Institute of Technology
Vadodara

Prof G.B.Jethava
Head & Assistant Professor, IT dept.,
Parul Institute of Engineering and Technology
Vadodara

*Abstract:* **Data mining concepts and techniques when applied to world-wide web with its existing technologies are termed as web mining. With flooding of information on World-wide web it has become necessary to apply some strategy so that valuable knowledge can be extracted and consequently returned to the user. This paper proposes a multi-agent based Web mining model which is designed to improve the search efficiency of the web pages in compare to the ranked list of keywords based search engine. The proposed model generates valuable information from the web logs of the server by dividing the mining task into several parallel agents which coordinately work together and classify the web documents by which the mining efficiency is improved greatly.**

*Keywords:*web mining, multi-agent systems, web server logs.

## I.Introduction

Web mining is the data mining technique that automatically discovers/extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

Web mining can be categorized as below.

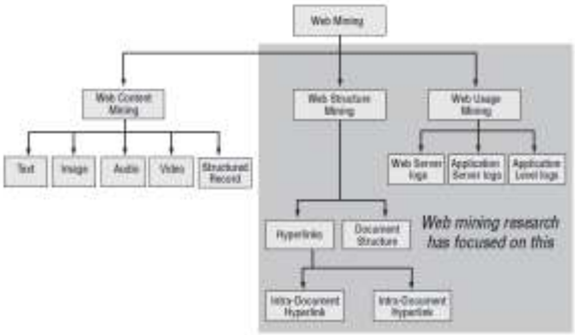


Fig 1. Web mining categories

**Web Content Mining:** - Web content mining is the process of extracting useful information from the contents of web documents. It is related to data mining. It is related to text mining because much of the web contents are text based. Text mining focuses on unstructured texts. Web content mining is semi-structured nature of the web. Technologies used in web content mining are NLP,IR.
**Web Structure Mining:** - tries to discover useful knowledge from the structure and hyperlinks. The goal of web structure mining is to generate structured summery about websites and web pages. It is using tree-like structure to analyze and describe HTML or XML.
**Web Usage Mining:** - Web usage mining is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It focuses on technique that can be used to predict the user behavior while user interacts with the web. It uses the secondary data on the web. This activity involves automatic discovery of user access patterns from one or more web-servers. It consists of three phases namely: pre-processing, pattern discovery, pattern analysis. Web servers, proxies and client applications can quite easily capture data about web usage.

## *II.Web Search Engines*

Many search engines are available on the Internet,each having its own characteristics and employing different algorithms to index, rank, and present Web documents. Examples of popular general-purpose search engines include AltaVista,Google, and Excite. These search engines allow users to submit queries and present the returned Web pages in ranked order.PageRank is an iterative algorithm used by Google that ranks web pages based on number and the PageRank of other websites and pages that are linked there so that good or desirable pages are linked together.Ranking items by its relevance thus helps in reducing the time required to find the required information. Further to find the set of matching items search engines collect metadata about group of related items beforehand, the process is called indexing. The index requires less storage and hence retrieval is faster and the iterative process works quickly.

## II.1 Web Crawlers

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion.Other terms for Web crawlers are *ants*, *automatic indexers*, *bots*, *Web spiders*,*Web robots*. A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the *seeds*. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the *crawl frontier*. URLs from the frontier are recursively visited according to a set of policies.[1]

But the problem with web crawlers is that the large volume of web documents on the World Wide Web implies that the crawler in a given amount of time can download only a few high prioritized web pages only. The crawling process is dynamic and hence the server side software generates a number of crawlable URLs due to which it becomes difficult to avoid retrieving duplicate content. The high rate of change implies that the pages may have been updated or deleted. Further the bandwidth for the crawls is neither infinite nor free so it becomes essential to conduct the crawl on the web in an efficient and scalable way.

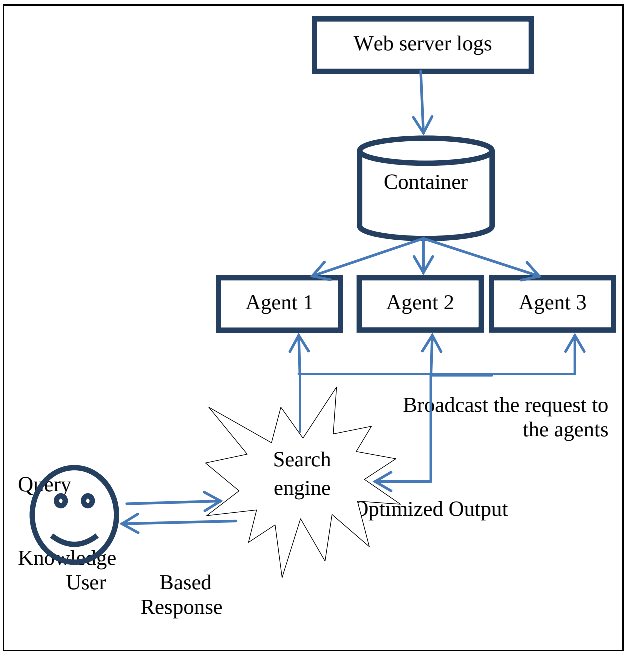
## II.2 Multi agent Systems

A **multi-agent system** (**MAS**) is a system composed of multiple interacting intelligent agents within an environment. Multi-agent systems can be used to solve problems that are difficult or impossible for an individual agent or a monolithic system to solve. Intelligence may include some methodic, functional, procedural or algorithmic search, find and processing approach.[1]

The agents in a multi-agent system have several important characteristics:

- **Autonomy**: the agents are at least partially autonomous
- **Local views**: no agent has a full global view of the system, or the system is too complex for an agent to make practical use of such knowledge
- **Decentralization**: there is no designated controlling agent (or the system is effectively reduced to a monolithic system).[1]

## III.Proposed Architecture



This work proposes anmulti-agent based web mining technique to optimize the process of web search. From the literature reviewed we found that that it is difficult to crawl through all the webpages in a given amount of time due to large volume of web documents. So to overcome this in our work we are going to classify the web documents on their attribute information and cluster them into multiple agents. As shown in the figure above only the specific agent has to be traversed so the result provided will be more optimized.
Here no metadata information needs to be collected as done in the indexing technique of web search to reduce the time of search. Here the database of the web pages contained in the container is clustered in a hierarchical format thus the web search starts from the seed to the lower levels in the hierarchy as per the attribute provided by the user.
As shown in the figure above the user's query is been broadcasted to all the agents. Each agent

processes the query and matches the entered keywords with the attribute type of that agent.Wherever the respective match is found only that agent processes the query further on its cluster of documents to provide a better result. The search time here is greatly reduced as the numbers of web pages to be processed are small. The search is alsooptimized and the results are moreappropriate than the normal indexed or ranked based search engines. The hierarchical database will further help in finding semantic web contents and inter-communication between the agents provides proper clustering of the contents. Even the network traffic will be reduced and bandwidth will no more be a constraint in web mining.

## IV.Conclusion

The work has proposed agent based mining of web contents with the aim to provide knowledge based response to the user. The next generation of world-wide web is knowledge oriented and to satisfy the customers web mining this is a promising solution.The agent based search of the web contents would improve performance and provide better results to the user.

## V. Future Work

Implementation of the proposed work is still under progress and is left as future work.

## VI. References

[1] www.wikipedia.com

[2] IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010 : Analysis of Server Log by Web Usage Mining for Website Improvement.

[3] International Journal of Advancements in Technology http://ijict.org/ ISSN 0976-4860.Vol. 3 No.2 (April 2012) © IJoAT :Agent Based Framework for Semantic Web Content Mining

[4] Sharma K., Shrivastava G. & Kumar V., ‘Web Mining: Today and Tommorrow’. In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[5] ZengminGeng Beijing Inst. of Fashion Technol., Beijing, China Xuefei Li Xiaodong Sun : ‘A Multi- agent based web mining model’.In proceedings of the IEEE 3rd International conference publications of PACCS.

[6] International Journal of Computer Applications (0975 – 8887)Volume 11– No.10, December 2010:’Constraint-based Web Log Mining for AnalyzingCustomersBehaviour’

[7] IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010 : “Analysis of Server Log by Web Usage Mining for Website Improvement” : By Navin Kumar Tyagi, A. K. Solanki and Manoj Wadhwa.

[8] “Web Mining: Information and Pattern Discovery on the World Wide Web” by R. Cooley, B. Mobasher, and J. Srivastava.