AI in Web Mining: a Systematic Literature Review

Liyachew Edeti
Software Engineer
Addis Ababa, Ethiopia
liyaliyanwf@gmail.com

Abstract— Web Intelligence is the application of Artificial Intelligence in Web Mining. Making the web intelligent enough. The web consists of noisy semi-structured and unstructured data to the utmost extent in a form of text, video, audio and different data formats. As data increases need for different types of information in different aspects varies. Different applications of data on the web lead to the web mining. Web became successor of data warehouse, where vast information mined through web mining process in a form of usage, content and structure. The mountainous data size on web needs intelligent approaches for Information Retrieval and manipulation purpose, i.e. integration of the field of Artificial Intelligence with data on the web applying different techniques. AI in Web Mining takes different types of optimization level based on the algorithms in all the techniques. Such as search engine best intelligence to search and filter information needed through web agents acting on behalf of users. AI algorithms like Page Ranking played the role to trust the top visited page and prioritize for inter and intra communication. The reach for web documents through personalization, gave further intelligence to identify user and to classify from the web logs on web server for further analysis. Though web data is bulky, query search shouldn't take much processing time. AI played the magnificent role to meet demands in sequential and parallel processing to retrieve information. Artificial intelligence algorithms in addition to machine learning techniques resulted with optimized solution to track, explore and predict user demand. With advancement of AI future BI (Business Intelligence) integrates more with web data than Big Data (transaction + interaction + observation), where more transaction conglomerate.

Keywords --- Web Intelligence, Web mining, Web content mining, Web usage mining, Web structure mining, AI, Personalization, Adaptive web.

I. INTRODUCTION

This systematic review assesses AI integration to web mining. The diversified integration of AI in web mining, that made collected set of machines on the WWW/web to interoperate intelligently and interact with the stimuli for desired response accordingly. The research question can be stated as follows:

RQ: How AI applied in web-mining?

The objective of the systematic review to investigate and explore AI effect in all the three types of web mining. That is, how AI enriched the web with intelligence or made web intelligent to perform efficiently.

Artificial Intelligence is the branch of computer science concerned with making computer behave like humans [6].

Currently Web mining could be viewed as the use of data mining techniques to automatically retrieve, extract, generalize, and analyze information [10].

Table 1. Steps in Evolution of Web Mining [12].

Evolutionary Step	Enabling Technologies
Data Collection (1960s)	"Computer, Tapes, Disks
Data Access (1980s)	Relational Database (RDBMS), Structure Query language (SQL),ODBC
Data Warehousing & Decision Support (1990s)	On-Line Analytic Processing (OLAP), Multidimensional Databases, Data warehouses
Data Mining (2000s)	Advanced algorithms, Multiprocessor Computers, Massive Databases
Web Mining (Emerging Today)	WWW, Internet, monumental scale Database

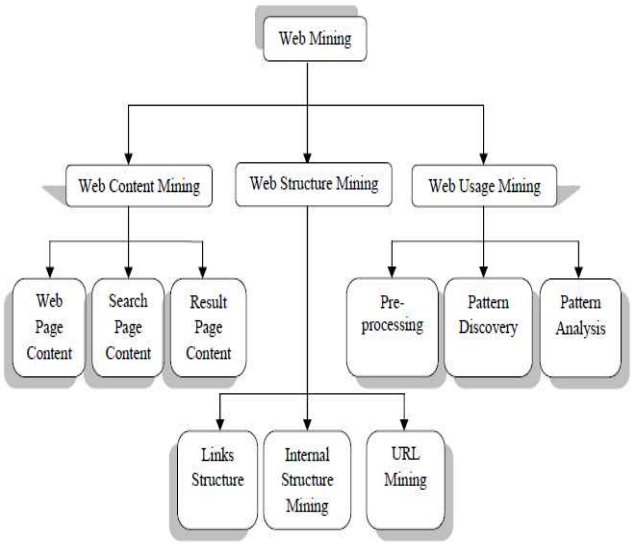

Fig. 1. Classification of Web Mining [12].

The Web mining analysis relies on three general sets of information: previous usage patterns, degree of shared content and inter-memory associative link structures corresponding to the three subsets in Web mining namely:

(i) Web usage mining,

It contains four processing stages including data collection, preprocessing, pattern discovery and analysis.

Pattern Discovery- According the data preprocessing discovered the knowledge and implements the techniques to discover the knowledge like as machine learning and data mining procedures are carried out at this stage.

Pattern Analysis- pattern analysis is the process after pattern discovery. Its check the pattern is correct on the web and how to implement on web to extract the information on your web search / extract knowledge from the web.

(ii) Web content mining

The Web content mining refers to the discovery of useful information from web contents which include text, image, audio, video, etc.

Two approaches used in web content mining are Agent based approach and database approach. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, personalized web agents.

Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine information. Adapted web agents learn user preferences and discovers documents related to those user profiles. In Database approach it consists of well-formed database containing schemas and attributes with defined domains.

Web content mining has the following approaches to mine data (1) Unstructured text mining, (2) structured mining, (3) Semi structured text mining, and (4) Multimedia mining.

There are two tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking [4].

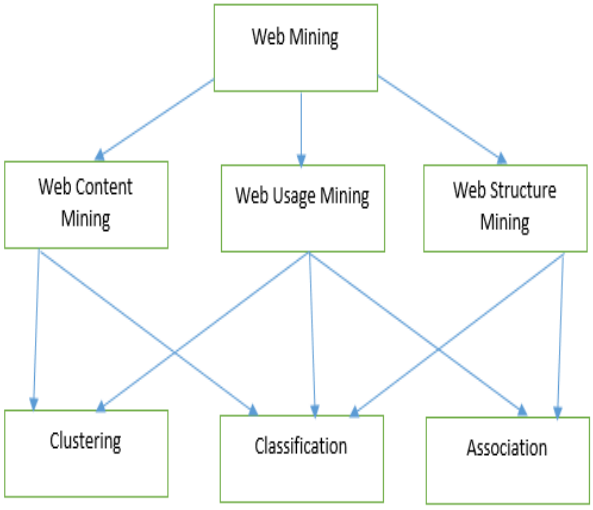

Fig. 2. Web mining is the application of the data mining techniques to discover patterns from the web [6].

(iii) Web structure mining.

We can define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's shown the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages/webs. [12].

The Web itself has been studied from two aspects, the structure of the Web as a graph and the semantics of the Web. Studies on Web structures investigate several structural properties of graphs arising from the Web, including the graph of hyperlinks, and the graph induced by connections between distributed search servants. The study of the Web as a graph is not only fascinating in its own right, but also yields valuable insight into Web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution. Studies of the semantics of the Web were initiated byTim Berners-Lee, the creator of the World Wide Web. The Web is referred to as the "semantic Web", where information will be machine-processible in ways that support intelligent network services such as information brokers and search agents [1].

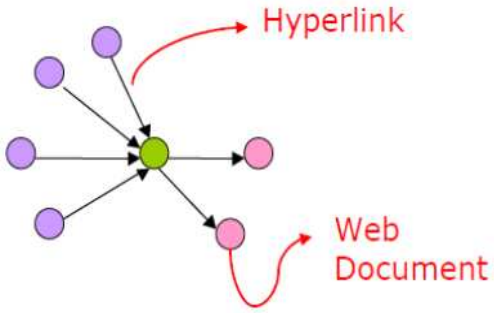

Fig. 3. Web graph structure [11].

Web structure mining is based on the link structures with or without the description of links. Markov chain model can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites. The goal of web structure mining is to generate structured summary about websites and web pages. It uses treelike structure to analyze and describe HTML or XML.

Web mining is work upon On-Line. In data mining data stored in (database) data warehouse and in web mining data stored in server database & web log [12].

Similar to the ants that do not have a global view of the environment, a web user is navigating the web without having information about the route followed by other users with the same objective. In order to apply an ACO algorithm to find the shortest path to a certain document or cluster of documents, information about target pages and routes can be kept on a special server [2].

Adaptive Web site is an attractive and challenging topic and the definition of this problem is provided in [Perkowitz and Etzioni, 1998] as follows: "adaptive Web sites are the Web sites that automatically improve their organization and presentation by learning from visitor access patterns" [3].

Web page Ranking is the main part of any information retrieval system. Web search engines are considered as intermediate between user and information repository. Search

engines used Crawler, spider, and indexer programs to present web pages [7].

According to (Tilak Patidar and Aditya Ambasth, 2016), discussion on Deep Web Crawling and AJAX crawling hold active interest and opens scope for more detailed and accurate web crawling. This inclines towards the field of Artificial Intelligence, empowering the spiders with human like selection intelligence [9].

WEB INTELLIGENCE has been presented as the usage of advanced techniques in ARTIFICIAL INTELLIGENCE and INFORMATION TECHNOLOGY for the purpose of exploring, analyzing and extracting knowledge from web data. WI explores the fundamental roles as well as practical impacts of Artificial Intelligence (AI) and advanced Information Technology (IT) on the next generation of web-related products systems, services and activities [6].

II. METHODS

A. Information Sources

Information sources which included books, journals, conferences, theses, webpages, presentations and technical reports and public electronic databases including IEEE, ACM, Science Direct, Citeseer, GOOGLE and Google Scholar checked for further investigations. It is tougher to have electronic or hard copy on "AI in web mining". It needs to adjoin different domains. Even AI by itself is an advanced topic and newer field to explore. The same to Web intelligence. However, the title makes crystal enough to understand an applied AI in web. Since AI is a wide concept to study, however combining with web mining it comes to web intelligence. Checking for the types of web mining presented with any AI algorithm type being sticky to the core concept web data mining/processing in different ways.

B. Study Selection

The study selection was made using words in the research question in detail, and related words as web intelligence also used. The review paper selection process made based on inclusion and exclusion criteria check lists stated as below.

i. Inclusion Criteria

- The primary studies must discuss AI algorithms, Deep Learning (DL) or neural networks and ML (Machine Learning) techniques in application to at least one of the classification of web mining.
- The primary studies concentrate to discuss the research question as focal point to some extent or fully.
- Unpublished but discuss web mining classification and contain beneficial idea in the knowledge domain.
- No publication year but with ISSN.

ii. Exclusion Criteria

- A paper discussing DL (Deep Learning) or neural networks, ML (Machine Learning) or AI algorithms individually or all together without web mining application concept. Since Deep Learning and Machine Learning are subsets of AI.
- Duplicate of information provided with the other primary studies.
- Author of the paper not mentioned.
- It's a book or magazine.

C. Data Extraction

Based up on the study selection data extracted from each primary studies included as the author cited for further analysis in the systematic review. For every aspect each detail of the paper read line by line to understand the concept. Any related concept that augment the study used according to the author or through IEEE citation standard.

D. Quality Assessment

Eleven papers assessed 9 of it published with reputable science journals and 3 unpublished, but useful domain knowledge inclusion. Since, Web intelligence is a newly developing research field yet there are challenges in getting too attached papers to the title under investigation. Since document analysis used, bias from experiment and subjective professional knowledge bias may not exist to the level of validity threat,. All the papers listed there in the references section. In all the papers content organized professionally by incorporating knowledge to some depth and being as guide to further enquiry to search more and analyze.

E. Data Synthesis

Data synthesized basically from defining the root words where importance is most to the detail description and application in the study. However, to prove the study objective in problem solving "AI in Web Mining" every detail learnt inductively to the general idea offering intelligence to the web (Web Intelligence).

III. RESULTS

Twelve papers selected passing through the selection criteria to the final systematic review. Almost all the papers done by experts of IT domain. Document analysis used as techniques to explore the objective of the research question, i.e, how AI applied to make information retrieval of modern days by far easier than felt by any user. Without different techniques applied in different level the DRIP/Data Rich Information Poor/ may happen definitely.

There are many web intelligence features discussed, better to recall and get the glimpse how the web getting by far matured to manage itself.

According to Ioana Moisil [2], almost all web mining tasks are using artificial intelligence techniques and algorithms in order to perform efficiently.

Pattern discovery, pattern analysis, unstructured text mining, structured text mining, semi-structured text mining, and multimedia mining are few of the tasks in web mining process.

Let us keep on listing again few of the techniques listed of the many from the primary studies: Clustering, Classification, Naïve Bayes, SVM, Markov Chain Model.

Application of these AI techniques on web give rise to adaptive web sites, web agents, web page ranking, web search engines, crawler, spider, indexer programs, web crawling, deep web crawling, AJAX crawling,

Web content mining uses clustering and classification. Clustering and classification data mining technique

objectively works based on similarity, and attribute selection respectively.

Significant web pages can be identified; also users that have common interests, i.e. using the identical clusters of linked pages. Till 1996, pages were fetched based on content similarity [11].

Crawlers are universal, topical and focused. Adaptive topical crawlers are the most sophisticated and they are designed using different machine learning techniques, in particular classifiers to guide them through the web [2].

The co-occurrence of links in web pages (user selections) was used to compute a matrix of link strengths [2].

Association works on frequency or co-occurrence. Web usage mining uses the three data mining techniques clustering, classification, and association techniques. Web structure mining uses classification and association techniques. By far clustering, classification and association are machine learning techniques. Machine learning is the sub set of Artificial Intelligence.

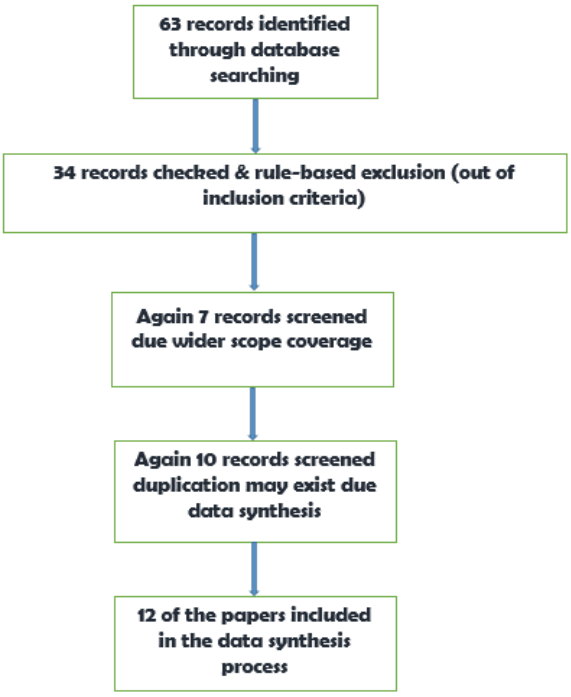

Fig. 4. Flow Diagram of the study selection.

IV. DISCUSSIONS

These days the application of AI in web mining has gone far more than to realize in comparison to the span of its invention 1950's. Even if steps ahead to execute without AI in web mining the management of web documents become cumbersome to handle. As data production across the web service increments too much, tremendous results achieved by using AI in Web Mining. Almost all of the papers affiliated with the positive impact of web intelligence, though security issue needs care. Let us summarize exploration made due the systematic review as follows:

The Systematic review study based on the primary studies organized in a way to brief and elaborate how the web became intelligent using different methods classified under algorithms and techniques in AI. As we have discussed the role of AI, it gives objects the ability to think and act by reasoning either rationally or logically, that is how AI made the web to act intelligently. AI offered web the dynamic features privileged inherently from nature.

The search for discovery of patterns (the relationship among data) in web makes too many applicability from simple to complex in problem solving. The global business benefitted too much in the case of web intelligence to be competitive in the market and put guidance plan to sustain in the market maximizing profit through business analytics. The web contributed for research and development indirectly to enhance the human fate in making living.

This paper contributed in some extent to the newly emerging field of web intelligence in exploring taxonomies and ontologies based on literature from the domain knowledge.

A. *Literature Survey of AI in Web mining*

(Y.Y. Yao, Ning Zhong, Jiming Liu, and Setsuo Ohsuga) [1] Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates web pages as being hubs or authorities. Many other HITS based algorithms have also been published. The most famous and powerful of these algorithms is the Page Rank algorithm. Invented by Google co-founder Larry Page, this algorithm is used by Google to organize the results of its search function. This algorithm helps determine the relative importance of any particular web page by counting the number and quality of links to a page. The websites with more number of links, and/or more links from higher-quality websites, will be ranked higher. It works similar to determining the status of a person in a society of people. Those with relations to more people and/or relations to people of higher status will be accorded a higher status.

An IWIS (Intelligent Web Information Systems) should be able to perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement. Intelligent agents are computational entities that are capable of making decisions on behalf of their users and self-improving their performance in dynamically changing and unpredictable task environments. Intelligent Web Agents (WA) are software programs that primarily serve two important roles: a). autonomous entities for exploring and exploiting Web-based services, and b). prototype entities for exhibiting and explaining Web-generated regularities.

(IOANA MOISIL) [2] WM is heavily used for e-education and e-business, as the WWW is again their main platform. Almost all web mining tasks are using artificial intelligence techniques and algorithms in order to perform efficiently. In the following we will briefly describe some of the most valuable AI techniques, the multi-agent technology and swarm intelligence algorithms. Web mining is an important and challenging activity that aims to discover new, relevant and reliable information and knowledge by investigating the web structure, its content and its usage. In our paper we have presented only two main AI techniques: the multi-agent systems and swarm intelligence, with some of their applications in web mining. The mining tasks are so complex that they cannot be efficiently performed without the support of appropriate advanced AI techniques.

Swarm Intelligence (SI) systems are in fact simple agents that are interrelated, being able to communicate one with another and to interact with their environment. The community

of agents carry out a distributed problem solving. They follow simple rules and there is no centralized control. Examples from nature of SI include ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling.

(Oznur Kirmemis Alkan and Pinar Senkul) [3] Initially, the system ask the users to provide the information regarding their reason to enter the Web site, or in other words, what they are seeking for towards using the Web site. In addition to this information, before users leave the Web site, they are asked to provide whether they have found what they were looking for. After that, those users' navigation paths together with their feedback are used in order to create suggestions for future visitors that seek the same content. The resulting suggestions are presented by highlighting the already existing hyperlinks.

In that project, the user's final objective as well as his next step is aimed to be discovered. A model for the user is built, partly based on the information that the user provides about him and also from his navigation paths. Using this information, direct links to pages, that are considered to be liked by the users, are presented to them. In addition to the consideration of specific likeliness, hyperlinks that lead to pages of potential interest to each visitor are highlighted.

(R.Malarvizhi and K.Saraswathi) (2013) [4] Some algorithms have been proposed to model the Web topology such as HITS, PageRank and improvements of HITS by adding content information to the links structure and by using outlier filtering.

(Pradnyesh Bhisikar and Prof. Amit Sahu) (2013) [5] The task of user and session identification is find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions.

A user layer containing users' profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents.

User queries can be enriched by adding new properties from the available domain ontology.

(Rahul Pareek) (2014) [6] An essential aspect of educational servers' adaptivity is personalization – they should be able to personalize interactions with each learner by keeping track of his recent visits/activities and relating the topics he learns and the sites he accesses during different learning sessions. Moreover, an intelligent educational server should actively help the learner and interact with him when executing these tasks. Since educational servers are interconnected, a specific server may personalize the session with a particular learner by prefetching the material the learner needs from other servers. This is an adaptive process based on observations of the learners surfing behavior.

(Seema Rani and Upasana Garg) (2014) [7] Author developed new method to index the web pages using an intelligent search strategy in which meaning of the search query is interpreted and then indexed the web pages based on the interpretation.

(Subhendu kumar pani, Deepak Mohapatra, and Bikram Keshari Ratha,Topic) (2010) (8) specific crawlers have become important tools to support applications such as specialized Web portals, online searching, and competitive intelligence. These crawlers are designed to retrieve pages that are relevant to the triggering topic. Generally employing single crawler to gather all pages is inevitably difficult. Therefore, many search engines often run multiple processes in parallel to perform the task. We refer to this type of spider as a parallel spider. This approach can considerably improve the collection efficiency.

(Tilak Patidar and Aditya Ambasth) (2016) [9] Politeness policy refers to not overwhelming a particular web domain with concurrent requests. It demands separating URLs by web domain with sufficient time interval to avoid banishment of the crawler.

(Tulasi Gayatri Devi, and Aparna KS) (2016) [10] Web mining, however, is very different from data mining in that the former is based on Web-related data sources, such as semi-structured documents (HTML, or XML), log, services, and user profiles, and the latter is based on more standard databases. In Web-based intelligent information systems, however, the new feature is that the set of inputs is very large. To avoid searching all inputs in KBs, the challenging issue is to quickly discard most non-relevant inputs.

Yahoo was the first to introduce the concept of a "personalized portal," i.e. a web site designed to have the look-and-feel and content personalized to the needs of an individual end-user.

Web mining techniques can be apply for the Customization i.e. Web personalization. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc.

(K.Harish Kumar) [11] Web Mining generates individual user's profile to understand the needs of users. It checks for fraud. Helps in internet advertising and also provides retrieval of similar images.

The summarized literature by far display how the web becoming intelligent incredibly in day to day activities. Artificial intelligence in different forms made the web too much better to be accessed, even though implementation of such algorithms and techniques by far seem daunting task without peoples' immense co-operation in the research field.

B. Limitations

The issue of streaming not discussed in any of the papers as the web is full of streaming data. Better to study some of the AI algorithms in web mining practically based on experiments. The study lacks applied techniques performance measurement as it matters most. Experimental outcome for each techniques by taking algorithms as sample may help more to understand and check the achievement of AI in web mining. Comparison reveals what to improve in relative to the technological resources speed, size, and processing power the world of invention reached.

C. Conclusion

The application of AI in web mining is crucial as the web data increases for optimum performance. As the world built of

information in this era, imaging web mining without AI is to crumble civilization. AI augments humanity not overriding it.

As the primary and secondary studies showed there is an advancement of AI applied in web mining from time to time creating enabling environment to interact in the web, though data size increased in geometric progression.

AI changed the world perspective by changing the web discourse in collaboration with giants whom labored imaginatively.

The systematic review described its objective on the issue how AI applied in web mining, and the research question implied with rich and thick data triangulation. The major issues in the coming web generation can be fixed by further enhancement of fruitful results achieved these days on the field of Web Intelligence. The issue of security, and privacy better fall under construction. The study intertwined these domains based on the factual data collected from references.

ACKNOWLEDGEMENT

I want to give my sincere gratitude and appreciation to Dr. Mesfin Kifle, whom provided me the title of my systematic review as best to conduct study.

REFERENCES

[1] Y.Y. Yao, Ning Zhong, Jiming Liu, and Setsuo Ohsuga, "Web Intelligence (WI), Research Challenges and Trends in the New Information Age", unpublished.

[2] IOANA MOISIL, "Advanced AI Techniques for Web Mining," MATHEMATICAL METHODS, COMPUTATIONAL TECHNIQUES, NON-LINEAR SYSTEMS, INTELLIGENT SYSTEMS, ISSN: 1790-2769.

[3] Oznur Kirmemis Alkan and Pinar Senkul, "IntWEB: An AI-Based Approach for Adaptive Web", unpublished.

[4] R.Malarvizhi and K.Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), volume 4, August 2013.

[5] Pradnyesh Bhisikar and Prof. Amit Sahu, "Overview on Web Mining and Different Technique for Web Personalisation," Applications (IJERA) ISSN: 2248-9622, Vol. 3, pp.543-545, March -April 2013.

[6] Rahul Pareek, "Web Intelligence-An Emerging vertical of Artificial Intelligence," International Journal of Engineering and Computer Science, ISSN: 231-7242, Voil. 3, pp. 9430-9436, December 2014.

[7] Seema Rani and Upasana Garg, "A Ranking Of Web Documents Using Semantic Similarity And Artificial Intelligence Based Search Engine," International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, December 2014.

[8] Subhendu kumar pani, Deepak Mohapatra, and Bikram Keshari Ratha, "Integration of Web mining and web crawler: Relevance and State of Art," International Journal on Computer Science and Engineering Vol. 02, No. 03, 772-776, 2010.

[9] Tilak Patidar and Aditya Ambasth, "Improvised Architecture for Distributed Web Crawling," International Journal of Computer Applications (0975 – 8887), Volume 151, October 2016.

[10] Tulasi Gayatri Devi, and Aparna KS, "A Survey on Web Mining: Overview, Techniques, Tools, and Applications," International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4, January 2016.

[11] K.Harish Kumar, "A Study on Web Mining Types and Applications,"International Journal of Trend in Research and Development, Volume 3(5), ISSN: 2394-9333.

[12] Kavita Sharma, Gulshan Shrivastava, and Vikas Kumar, "Web Mining: Today and Tomorrow", unpublished.