International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2015): 3.791

Review Paper on Web Structure Mining

Deepika Bhadoria[1], **Dr. Pradeep Sharma**[2]

[1]Mahatma Gandhi Chitrakoot, Gramodaya Vishwavidyalaya, Satna, Madhya Pradesh, India

[2]Govt. Holkar Science College, Indore, Madhya Pradesh, India

Abstract: ***The exploration for the knowledge has directed to the latest discoveries and creations. Along with the realization of the WWW, this have became the basic for the all type of discoveries. And the web-browsers have become the tool in order to make data to be available at the finger-tips of users. So the Web-mining process has been represented along with various approaches here within several of applications. The web data includes web-pages, links of web, objects on the web and some web-logs. The Web-mining process is used in order to make the behavior of customer to be understood, also evaluate a particular website based on the information which is stored within files of web-log. The review over mining approaches have made with respect to Clustering, Classification, Sequence Pattern Mining, ARM and Visualization. The research work done by different users depicting the pros and cons are discussed. Here also represented some issues, the process, algorithms and the application of the web-mining process.***

Keywords: Web Structure Mining, WWW, Clustering, Classification, Sequence Pattern Mining

1. Introduction

With the growth of information resources presented over the WWW, that have become essential for the users in order to use the automated type of tools in order to find the required resources of information and also for tracking and analyzing the usage-patterns. Also these types of factors are creating the server-side and the client-side type of intelligent platform which may mine the knowledge. And the web-mining process may be referred as searches and the measures of the essential type of information among the WWW. Also the web is the collection of billion of documents. The web is very enormous, diverse, flexible, and dynamic.

The WWW continues to grow both in the huge volume of traffic and the size and complexity of Websites. It is difficult to identify the relevant information present in the web. Most of the contents in the web are unstructured in nature, but very little work deals with unstructured and heterogeneous data over the internet/Web. And the web-mining process is an application of the DM approaches in order to extract the knowledge among the Web-data consisting of web documents, hyperlinks between documents, logs of usage of the different websites, etc [1]. The significant web mining applications are website design, web search, retrieval of information, search-engines, network management, Ecommerce, business, AI, market places for web, web communities, etc.

Due to increase of information overloading it becomes difficult to map or index all information on the web that's the reason much of useful information gets dropped and remain un-indexed. There are so many techniques which are used to solve these issues like the retrieval of information, database and the processing languages. The web is an attracting domain of the research. This may help to extract the knowledge among the web-information. In which even one of structure or usage of data is applied within process of mining. Based on the analysis the web-mining may be divided into the given separate types. The Web can be taking an important place in human's life and day by day it increases the number of information based on the expectations of the customers using it. Daily Updating is needed to fulfill the needs of the users.

The crime can be performed over the Internet, by the use of Internet and also through the applications over Internet. Along with the emerging Internet gave raise to another type of revolution of the crimes in which perpetrators commit acts of crime and wrong doing over WWW. The Computer system-crimes are the term which embarrasses such as illegal downloading, creation or distribution of viruses, child pornography, cyber terrorism, online threats, frauds etc. It is difficult to identify the relevant information present in the web. Moreover, most of the contents in the web are unstructured in nature. These problems can be addressed by the emerging domain of the web-mining process. The major objective is to find and then extract the significant information which is concealed within the Web-associated information, particularly within the text type of documents that are published over the internet. And the process of DM works with the extraction of useful, meaningful and valuable information from huge collection of the data. And the Web-mining mechanism is the vital area in data mining where the interesting signatures are derived among the contents of web.

As a result large quantity of web data has been generated. DM approaches are applicable to mine interesting data. But we cannot implement the approaches of DM directly to the web data since the web data is unstructured or semi - structured. Thus here use the web-mining process that can be applied for the data of web. And the Web-usage type of mining [1, 2]extracts meaningful & interesting patterns of usage in web site which can be uses in a various ways like improvement of web sites, checking of fraudulent element, better understanding of user behavior etc. The process of Web-mining is again categorized into the other types like the web-content process of mining, Web-Structure process of mining and finally the Web-Usage process of mining [5]. By the use of the objects such as pictures, text, multimedia and so on the content oriented mining process is performed within the web. Within the Web-structure process of the mining is performed dependent over the structures such as the hyper-links. And for the web-usage process the mining approach is performed over the logs of web that may have the pattern of navigation of the customers. Also the analysis of this type of navigational methods may get traced for the benefit of the customers [4].

Within this analysis represented the view regarding the method in order to extract any essential and the significant detail over the web by the use of web-mining process and also here provides the extra knowledge along with the detailed comparison regarding the data-mining process. Within this paper also discusses about past, present and the future aspects of the web-mining process. In this also describe the online resources used for the retrieval of Information over the internet such as the web-content process of mining also the discovery of the access-patterns of the user through the web-servers such as by the web-usage process of mining which may enhance the drawbacks of DM process. Here also described about the web-mining by using cloud-computing mechanism like cloud-mining process. This may be considered as the future of the Web-Mining process.

In webpage the content-mining process of web is the type of search done via content. And the result of search of the content mining searches from the earlier results of search. If users search for any specific key word or any web page, number of links or result is displayed. But all the data which is displayed on the web is not relevant. So efficiently and effectively retrieve required data on the Web is becoming a challenge. The user issues the query terms (keywords) to the search-engine and also this may returns a set of pages that may be related to the query topics or terms. The web is very enormous, diverse, flexible, and dynamic. The WWW may continue to grow both in the huge volume of traffic and the size and complexity of Web sites. With the increasing use of information which is available within net it is difficult to identify the relevant information present in the web. Meanwhile much information is unstructured. It is necessary to use automated tool for obtaining the necessary information from the huge volume of the information or data.

The process of Web-content type of mining is the extraction of knowledge among the huge content of the documents along with the descriptions. And the web-document-text process of mining is the discovery process of resource dependent over the idea of indexing or the agents also it is dependent over the techniques that may also come within these categories. And the web-structure process of mining is an approach of deriving the knowledge by the WWW based organization and also from the links in between the references and the referents within Web/Internet. And in the end, web-usage process of mining, which is also referred as the Web-Log process of Mining, which is a method of deriving the interesting signatures within the access-logs of web.

2. Web Mining

The process of Web-Mining is a method of extracting the interesting or significant patterns that are useful and also some implicit details from the artifacts or from the activity that are related with the WWW.Also roughly there are three different types of fields for knowledge discovery which may have the web-mining process like the Web-Content approach of Mining, second is Web-Structure approach of Mining and last is Web-Usage approach of Mining. In the Web-content approach of mining is a method of deriving the information from various contents or documents of different types. The Web-mining is an integrated technology in which several research fields are involved, such as data mining, computational linguistics, statistics, and informatics, etc. Different researchers from different types of society are disagreed from each-other over what the Web-mining process is actually does [1].This may also have the images, text, audio, structured records, video, etc in the lists or tables form [5]. The research within the web-content can be mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages [6].

The process of Web-mining is also applied in order to gather important information or for creating the latest knowledge from those types of important data, also perform the personalization of information, may also enable learning regarding the customer or each user’s working also various other type of functioning. The WM process may uses the approaches of DM in order to automatically it can discover and also may extract the information among the WWW [2].Also various types of approaches such as retrieval of Information, extraction of Information, machine-learning process, etc may have also been applied within the last few years in order to find out the latest type of knowledge by the large volume of the data presented within the internet.

These types of approaches have also been got compared along with the web-mining process [5].And the retrieval of Information may be works through indexing the text and after this selects the essential information[8]. Such a repository contains not only text data but also multimedia objects, like the audio, images, video, etc. The DM process over the WWW may be expressed as the WM process that may have obtained various interests along with the fast growing volume of the information that are present over the web/internet.

The process of Web-mining can also consists of four important steps, they are, resource finding, data selection and pre-processing, generalization and analysis [3]. Resource finding is the process which is used to extract the data either from online or offline text resources. In data selection and preprocessing step, particular information among the derived sources of web are automatically selected and pre-processed.

During generalization, the data-mining process and the machine-learning approaches are used to discover basic patterns among the each type of web-sites also from various sites simultaneously. Based on these kinds of the information a Web-Mining technique may be consists of 3 processes referred as Content, structure and the Usage process for Mining. In this content type of mining may deals with the raw data that is available on the web. The web structure mining mainly deals with the structure of the web sites [6].

The Usage type of mining process may consist of mining any usage features of customers of the different Web-applications. It is in a semi-structured format so that it needs lots of pre-processing and also parsing can be done before any actual process of extraction for any essential information or data. Both sides are face problems while dealing with several web-data. Therefore Usage process of mining may retrieve useful data. But there will be many copies of the same useful information or data which is available. Hence the usage types of mining can makes use of SOM model cluster only the similar data and eliminate redundancy.SOM process is one of the unsupervised learning method in the family of the ANN and this may also be applied within the usage type of mining process for getting similar data and avoid redundancy.

Earlier people used to communicate through postal services, purchase the products from nearby markets, and gather information from news papers and magazines. Even people do business and banking transactions manually through paper work. But in today era we have a vast ocean of data which we called as internet or web. This huge library of data originates as a result of modernization and globalization of data over internet.

All the activities discussed above, which we used to do manually earlier, now becomes the part of internet and hence the result of raising the data within the web. The process of WM may helps in order to understand the behavior of customer also it supports in evaluating the efficiency of the websites and also the researches have made within the content type of mining process which is also support to increase the business. The content type of mining approach may determine the result of search by the search-engine. And searching done manually may consume so much of time. If the data is to be observed within huge volume, so it is complicated to detect the useful data. As current within each domain of the life, various manual works are replaced by some type of techniques. Similarly up gradation can be happened within the internet functioning also.

The approaches of Web-mining may generate result after the long procedure of the research and also development of products. This type of evolution may begin if the volume of data is maintained within the files of computer and so the databases are also emerging at the exceptional rate. Simultaneously the users of those type of data are also expecting very complicated information to be obtained from them .Also the marketing-manager may not be satisfied by this simple type of listing of the marketing related contacts whereas required the brief information regarding the customer's previous purchases also the prediction of future purchases [7].

3. Types of Web Mining

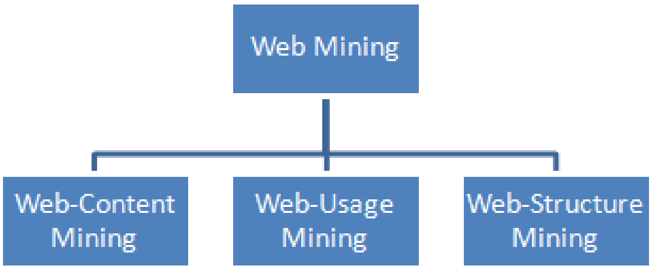

Figure 1: Types of WM process

The mechanism of Web-mining is again categorized into the 3 different mechanisms as first is Web-content mechanism for mining process, Web-Structure mechanism for mining process, Web-Usage mechanism for mining process [8]. By the use of objects such as pictures, text, multi-media and so on the content oriented approach of mining is performed within the website. Within the Web-structure oriented mining concept, it is oriented over the complete structure of the content such as the hyper-links presented in the content. Here is some brief explanation over the different types of the mining processes:

3.1 Web-Content Mining

This is a mechanism of extracting the essential information through the contents presented over the various web-documents. And the Content of data is a group of the several facts that a web-page is composed of with those contents. This can provide useful and interesting patterns about user needs and contribution behavior. Web data contents may involve the different types of data. Those types of data are Images, Texts, Audios, Metadata, Videos, Hyper-links, etc. This may discovers the essential information among the web contents or data or the documents present over internet. It involves mainly three processes [9].

There are some issues also found within the text type of mining process consist of the discovery of topic and then tracking, also the extraction of the association-patterns, perform clustering process of the documents of internet and may also be the classification process of the web-pages. The activities of research over this suggested domain may have created basically over the approaches that are developed within the other domains like the Information-Retrieval process and Natural-Language-Processing mechanism. Whereas there may also found an appropriate body of the work within extracting the knowledge from the images within this domain of the image-processing concept and also for the computer vision, may also be in the implementations of these approaches for which the content-mining can have the limited functions. In the web-content-mining process it is distinguished from the two separate aspects like on the basis of retrieval of information view and the View of Database.

3.2 Web-Usage Mining

This type of mining approach is also referred as the Web-log-mining process. This is also applied for discovering user's navigation-patterns among the web-data and prediction of user behavior while they interact with the web. It helps to improve large collection of resources. There are typical applications of this type of mining process like web type of personalization, adaptive websites and user modeling. This type of usage-mining process is an implementation of the various approaches of DM over the web-log containers to discover knowledge about user behaviors [10].

This may also expressed as the process of discovery of the access-signature of the user through usage-logs of the web. This may also targeted over several type of DM approaches in order to understand it and analyze the patterns of search. This type of mining process may also referred as automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of the communication of the user by using the resources of Web over more than one type of Websites. And the goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site.

There are four main mining techniques that can be applied to Web access logs to extract knowledge, but we will focus on algorithms oriented over the ARM concept and sequential mining because of their complexity, applicability and popularity.

3.3Web-Structure Mining

This type of Mining approach is used to recognize the association in between the various web-pages that are linked through the information or through the direct links or connections. Major objective of this type of mining process is to derive the existing type of un-known associations in between the various web-pages. The usage type of Mining, can examine the log data stored indifferent formats in the web server, proxy server and client caches. So the structure oriented mining approach may apply to reduce the two major issues of the WWW caused because of its huge volume of information. And the first issue is not related with the results of search at all. The pertinence of the searched information have become mis-construed because of the issues that the search-engines are basically applied for the low precision conditions. And second issue is the inefficiency to create the index of large volume when the information is given over the internet. This may create the less volume of the recall along with the content type of mining approach. So this type of minimization has applied along with the process for discovering some models within the Web based hyper-link architecture offered by this type of mining technique [8].

Some issue for this type of structure oriented mining process is related with the architecture of hyper-links presented over the Web. And those Links get analyzed by the researchers in previous time. Though, by this emerging interest within the Web-mining mechanism, various researches have may for the analysis of structure have also increased so these serious efforts may result in the latest growing research domain which is referred as Link-Mining, that is situated at the communication level of various works like the analysis of link, hyper-text or the web-mining process, inductive-logic-programming, relational-learning process, graph-mining approach, etc.

4. Web Structure Mining

Within the Web graph of the typical type of Web-pages may have the structure includes the nodes and some hyper-links in the form of edges that are connecting the associated pages. The suggested WSM approach is a procedure of finding the structure of information store over the Web. This type of structure mining approach is targeted over the hyper-link oriented structure of a Web technique. Various objects in this are linked with each other by any way. Just by implementing the basic processes also by supposing that some events are not dependent so they may cause incorrect conclusions. This type of mining structure is an approach of applying the graph-theory in order to observe the node along with their structure of connections of the websites.

This structure oriented mining can helps the users to retrieve the relevant documents by analyzing link oriented structure of any type of Web content. And some issues for these types of mining in order to work with any structure of given hyper-links in the Web. In this analysis of Link is also a domain for researchers. And the Web may consist of not only of pages, but also of hyperlinks pointing any page to the any other type of pages. It discovers the specific structure of links that is hyperlinks provided at some level of inter-documents and also in order to generate the structural conclusion regarding any Web-site or any Web-page.

This concept is used for retrieving pages that are not only relevant but are also of high quality, or authoritative on the topic. Though, by the emerging attention within the Web mining concept, the analysis of the structure have also been increased also their efforts can have generated in the latest type of research domain referred as the Link-Mining [8], that is also placed in the middle of the content within the analysis of link, also the hyper-text is placed and performed the web-mining through some learning and the inductive type of programming approach and through graph-mining processes.

This can be further categorized into two different types based on the kind of structure information applied.

4.1 Types:

4.1.1 Hyperlinks:

This is the unit of structure which may links a location within the Web-page with the other locations, that are belongs to the similar Web-page or over the different type of Web-pages. Any type of hyperlink which may connect with the different sections of similar page is referred as Intra-Document-Hyper-link also the hyper-link which may links the two type of pages is referred as an Inter-Document Hyperlink. There has been a significant body of work on hyperlink analysis, of which [23] provide an up-to-date survey.

4.1.2 Document-Structure:

Additionally, the content presented in the Web-pages may also arrange in the tree type of structured pattern, this is dependent over the several types of HTML-tags or the XML-tags in any page of web. Mining efforts here have focused on automatically extracting the document object model structures out of documents [5, 7].

The WSM is also a category type of the web-mining process for the data, which is a type of tool that is applied in order to recognize the association in between the Web-pages that are connected through the information or may be with the direct connection. And this type of data structure is introduced through the supplying the web architecture model via the techniques of database for the Web-pages. So this type of connection may enable any search-engine in order to pull the data which is related with the query for search which is directly connected with the linking page of any Web-site on which the content is placed. This type of work may be performed by the use of the spiders scanning process over Web-site, also retrieving content from the home-pages, after which is links the detailed via the links of reference in order to bring some particular pages in front that are having some desired details.[7]

Major objective of this type of structure mining process is to derive the existing type of unknown association in between the Web-pages over web. And this structure type of DM may enable to use it in the business purpose in order to connect the details of their Web-sites in order to allow the navigation function and also cluster the information over the site-maps. Also this may enable their users to have the capability to use the required information via the association of keyword along with the content oriented mining process. Here also may define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's show the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages/webs. It is shown the link one web page to another web page.

4.2 Processes in Web-Structure-Mining:

4.2.1 Link-oriented process of Classification:

This process is latest type of upgraded one for the basic type of DM works in order to link the domains. This process is targeted over the anticipation of categories of the web-pages, and also dependent over the words which may found over the pages, or links in between the pages, over the html-tags, anchor-text may also be over the attributes that are present over the web-pages.

4.2.2 Link-oriented Cluster-Analysis process:

Objective within this process is finding the naturally found sub-classes. And the data in this is got fragmented into the groups, in which same type of objects are merged with each other and the different type of objects are also grouped separately within the other groups. It is a different type of process in which this type of analysis is considered as unsupervised and also it may be applied in order to find out the hidden-patterns among the data.

4.2.3 Type of Link:

Approximately a wider range of works are considered in this process of prediction for available links, like anticipating the pattern of the links in between the two different entities or may anticipating the objective of the links.

4.2.4 Strength of Links:

In this the Links may be related along with their weights. And each link should have unique weight

4.2.5 Cardinality of Links:

Major work of this process is to anticipate the links present in between the objects.

5. Optimiztion Techniques in WSM

Here are some basic types of Optimization techniques used in WSM are described below:

5.1 Clustering-technique:

This approach of Cluster observation or simply referred as clustering is basically the process of creating the groups of the set of the objects as a manner such that the objects within the similar groups are termed as a cluster. And it is a primary task of explanatory DM process that basic type of approach is applied for the analysis of statistical data which is used within various domains like pattern, machine-learning approach, picture analysis, data retrieval & Bioinformatics.

5.2 Page-Ranking Technique:

The ranking process is basically a relationship in between the group of the items like for two given points either the first is ranked to be higher than or ranked to be lower than or may be it ranked to be equal-with the second point. Mathematically, it is unnecessary to be use the total number of objects as the two separate objects may have similar type of ranking. And the page-rankings is completely merged. The Page Rank leads a better approach that can calculate the importance of web page by simply including the number of pages that are linking to it. The above calculated links are called as back-links. In case a back-link generates from a key page and then this link is given higher weight-age than those which are coming from non-important pages. So link from the first page to the other page is measured as a vote.

5.3 Hyper-link induced-topic-search algorithm:

This approach is suggested highlights and defined with the two attributes like hubs and authorities. It uses the process of link-analysis that may rates any web-page and also invented developed an approach which enable the use of linking structure for web so that it may find out and then rank the pages that are related with a specific concept. This approach is also follows the concept of search-engine called Ask. The HITS algorithm uses the Sampling and Iterative steps.

5.4 Ant-colony-optimization algorithm:

In this algorithm the pheromone trails are used. Artificial ants may carry one or more objects and may drop them according to given probabilities. These agents do not communicate directly with each other, but they may influence themselves through the configuration of objects on the floor. Thus after a while, these artificial ants are able to create sets of same type of objects and a problem which is known as data clustering. [4] A traveling salesperson problem is perfect for ACO, because this problem closely resembles finding the shortest path to a food source. ACO results in premature convergence to a local optimal solution unless pheromone evaporation is implemented; a solution disappears after a period of time.

5.5 Genetic-Algorithm:

This approach is applied due to extra ordinary enhancement in the volume of the information over the WWW have increases the subject of crawling of a Web. While the targeted process of crawling may includes the mechanical type of categorization of Web page approach like WPC which is required in order to find out if the page is being taken for the approach or not. For this type of learning process, the Genetic-Algorithm is dependent over the mechanical type of WPC approach that uses the HTML and the conditions both that are related with each tag as the features of classification and also the it analyzes the best classifier obtained from any positive or the negative type of Web-pages within given training data-set. This system may classify the Web-pages through just estimating the relationship within the learned type of classifier and also some latest Web-pages. This may have various issues in order to recover those complexities by applying the Hybrid type of GA or PSO approaches.

5.6 Firefly Algorithm:

It is a Meta heuristic approach that is inspired through the glowing nature of fireflies [6]. To attract other fireflies is the main aim of firefly's flash. In this analysis formulating the algorithm through supposing situations like entire fireflies are unisex, such that anyone firefly may attract the entire type of fireflies. Brightness make them attractive accordingly, and also for the any two types of fireflies if one is less glowing then one will attract (and thus move) to the brighter one; here, distance increases makes decreases of brightness. And in case no single fireflies are more glowing as compare to the given firefly then this may be randomly move.

5.7 PSO algorithm:

In the PSO algorithm, the birds in a flock are shown as particles in n dimension. Best fitness value of particle at a location in the n-dimensional problem space represents one solution for the problem. When a particle updates it's position, another problem solution is generated and then new solution is evaluated by fitness function and the process is repeated until a stopping criteria is met. PSO has a random population matrix like the GA, but the rows in the matrix are called particles instead of chromosomes. Particles are potential solutions that move in a particular direction on the cost surface with a certain velocity. Particles update their positions and velocities using formulas based on the knowledge about the best solution achieved by each particle in its movements (i.e., personal best) and by the complete swarm of particles (i.e., global best).

6. Literature Review

In this paper [11] suggested an effective approach for the WPC technique of data-mining. This approach may be very efficient for the training-set which is the set following the way so that it may produce maximum sets. However the results of experiments of this analysis are also influencing so that this will be efficient in case it is applied with the huge sized data-sets along with the maximum type of classes. Also the available approach may needs to be some data used for the training also less time is required for computation of these approaches.

Within this paper [12] a significant amount of patterns can be retrieved through applying the approaches of DM in order to derive the information among the data from Web. Various types of approaches for DM have also been suggested within the past some years. And these approaches may consist of the frequent-item-set mining process, ARM, sequential-pattern-mining process, closed-pattern-mining, etc. Though, here presented the way to efficiently apply these types of discovered patterns is still an unsolved problem. Another typical issue is that only the statistic properties are used while evaluating the effectiveness of patterns. A comprehensive comparison of data mining approaches is implemented for the Web-mining purpose is performed in this analysis. And the results of experiment have represented that the closed-pattern approaches like the SCPM or the NSCPM may have perform more efficiently because of the using pruning technique within the stages of pattern discovery process.

This paper [13] gave an idea about web mining and how it can be utilized in an efficient way to improve the business. Customer behavior is very important for an organization to enhance the way of providing information to attract them. Analysis of significant information will be helpful for organization to develop promotions that are more effective, internet accessibility, inter – company communication, structure and productive marketing skills through the web-usage type of mining process. Pattern extraction and the web-content type of mining process are the best tools to know the customer and web behavior.

According to this paper [14] suggested the semantic type of web features that may enable to add structure to the Web, while Web Mining can learn implicit structures. And the merged domain of the semantic type of web-mining process may enable the latest approaches to be applied in order to enhance both the domains. The semantic approach of web-mining may involve the integration of the concept of knowledge within the process of web-mining. Domain knowledge can be integrated within the process of web-mining in three ways- domain ontology acquisition, knowledge base construction, and knowledge-enhanced pattern discovery. And obtained benefits of research within various domains of industry like health-care, e-activities, privacy, knowledge-management, security, etc have presented here.

Within this paper [15] describe the issues of the widely used domain over the big-RDF data and also suggested here a latest summary-oriented solutions. This analysis may provides a brief summary over the level types from the RDF type of data while evaluation of query and this represented the summary in order to reduce the required area of the RDF data among the search region and also it formulate the queries of SPARQL for effectively using the data. Additionally, the suggested analysis may be regularly get upgraded with the updates in data. And the experiments performed over the RDF-benchmark and also over the real datasets of RDF have represented that solution provided here is more effective, portable over the RDF type of engines, scalable, etc.

In this paper [16] suggested web-content oriented mining process as a component of Data Mining. Whenever in this discussed about data, and conclude that there is a vast range of data over WWW. And to manage this vast range of data, also often need many tools that can retrieve the data as per the criteria. There are various tools available on the internet which mines the data according to their types like whether they are in a structured format or the semi-structured type data or may be un-structured type of data. Here also discuss the project work which has an ability to mine the data from web efficiently. In this paper, the previous type of approaches of web-content oriented mining process will also be discussed and the tree structure of a webpage document will also be discussed.

This paper [17] has presented the approach of web-content based process of mining to be applied in order to derive the properties of the product and this may labels the attributes within the obtained result. Also the labeling process enables to recognize and also providing naming to the attributes when process of information-retrieval is completed. Then the information-gained may be applied here for analyzing the product along with its explorations. The approach of web-content type of mining is just the merging of the data among several types of resources through analyzing the views of customers. And this paper has also represents the analysis over the suggested concept applied for the mining and its applications for mining. Here also represented few growing approaches that are applied for deriving the data from the various online shopping portals.

In this paper [18] describe an analysis regarding the domain of Web-structure based mining concepts applied for structure and view of data. Here points some confusion in between the DM approach and the web-mining approach. Web data is growing at a significant rate. Web Mining is fertile area of research. Many Successful applications exist. In this also suggest the subtask of web mining & future of the WM process. Now also work for the process mining and try to combine the usage based mining process along with the structure oriented mining processes. In this analysis also go for the mining from cloud. Whenever work on mining over cloud computing that time researchers may hesitate for the cost but that come very less by cloud mining. So, also say that cloud mining can be seen as future of web mining.

In this paper [19] suggested an efficient method to address some of the problems during web content extraction. In the proposed method extract the required signatures through eliminating the noise which is present in the web document. Proposed method shows better performance when compared with existing methods. In future also plan to extend the work to construct DOM tree (Graphical representation) after extraction of useful patterns. Here suggested a new method for web data extraction. It has three phases. In the first phase list of web documents are selected, second phase documents are preprocessed, in the final phase results are presented to users. Experimental results are compared with existing methods. Performance of proposed system is better than existing methods.

In this paper [20] describe a definition of web mining, research direction and benefits of the web-mining process along with its taxonomies. Here identified some of the issues and problems in this area which can needs further analysis and development. Web mining is applied to various fields E-Commerce, Information filtering, Fraud detection Education and research.

In this paper [21] have analyzed the approach of web based mining of data that is the implementation of techniques of DM in order to derive the knowledge from the data of Web which may contains hyperlinks within the documents, some text documents, logs of usage of the web-sites, and so on. Now in today's advanced world, web becomes an important part of many of all organizations, businesspersons and daily individuals. As a web data is of very many different formats, we have studied the characteristics of web data. As it is very much important to mine particular data from web, we have studied two effective techniques to mine this big data one along with the apache based Hadoop-Map-Reduce approach and second with visualization based technique called as Visual Web Mining.

7. Conclusion

Within this paper it is concluded with briefly describing the basics of computer technology with its contribution within various domains of mining of data over web along with its various types and also highlights few promising regions of the future analysis. In this paper given the brief description about the web-structure type of mining process (WSM) along with its various functioning and also explains types. This paper have described about the various techniques of the web- mining concept. This type of mining approaches has been proved very useful in the business world. The analysis have also discusses the techniques used for extracting information from different types of data available in the internet and how this extracted data can be used for mining purposes.

References

[1] T. Sunil Kumar, Dr. K. Suvarchala, "A Study: Web Data Mining Challenges and Application for Information Extraction", IOSR Journal of Computer Engineering (IOSRJCE), Vol 7, Issue3, Nov-Dec 2012, pp 24-29.

[2] D. Jayalatchumy, Dr. P. Thambidurai, ―Web Mining Research Issues and Future Directions –A Survey‖, OSR Journal of Computer Engineering (IOSR-JCE), Volume 14, Issue 3 (Sep. -Oct. 2013), PP 20-27.

[3] Pradnyesh Bhisikar, Prof. Amit Sahu, ―Overview on Web Mining and Different Technique for Web Personalization ‖, International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 2, March - April 2013, pp.543-545.

[4] Darshna Navadiya, Roshni Patel, " Web Content Mining Techniques-A Comprehensive Survey", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, December- 2012, pp.1-6

[5] S. Balan, P. Ponmuthuramalingam, " A study of various techniques of Web Content Mining Research Issues and Tools, International Journal of Innovative Research and Studies (IJRIS), Volume 2, Issue 5, May 2013, ISSN: 2319-9725.

[6] R. Lokeshkumar1, R. Sindhuja2, Dr. P. Sengottuvelan, "A Survey on Pre-processing of Web Log File in Web Usage Mining to Improve the Quality of Data" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014

[7] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi, ''Overview of Web Content Mining Tools", The International Journal of Engineering And Science (IJES), Volume 2, Issue 6, 2013.

[8] Rosli Omar; Abu Osman Md Tap; Zainatul Shima Abdullah, "Web usage mining: A review of recent works", Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference.

[9] Ananthi .J, " A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014

[10] Amit Pratap Singh1, Dr. R. C. Jain 2, " A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)Volume 3, Issue 3, May – June 2014 ISSN 2278-6856

[11] Keyur J. Patel1, Ketan J Sarvakar, "Web Page Classification Using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 7, July 2013

[12] Sheng-Tang Wu and Yuefeng Li, "Pattern-Based Web Mining Using Data Mining Techniques", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 3, No. 2, April 2013

[13] Abdul Rahaman Wahab Sait1 and Dr.T.Meyappan2, "WEB MINING – A CATALYST FOR E-BUSINESS", Advanced Computing: An International Journal (ACIJ), Vol.3, No.6, November 2012

[14] V. Sitha Ramulu, Ch. N. Santhosh Kumar, K. Sudheer Reddy, "A Study of Semantic Web Mining: Integrating Domain Knowledge into Web Mining", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[15] V. A. Chakkarwar and Amruta A. Joshi, "Semantic Web Mining using RDF Data", International Journal of Computer Applications (0975 – 8887) Volume 133 – No.10, January 2016

[16] Pooja Rohilla1, Ochin Sharma2, "Web Content Mining: An Implementation on Social Websites", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015

[17] Ananthi.J, "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 4091-4094

[18] Kaikala Anjani Sravanthi1, Yalamarthi Madhavi Lata, "Web Mining Using Cloud Computing", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 4, April 2013)

[19] V. Shanmuga Priya1, S. Sakthivel, "An Implementation of Web Personalization Using Web Mining Techniques", IJCSMC, Vol. 2, Issue. 6, June 2013, pg.145 – 150

[20] M. Vengateshwaran and E. V. R. M Kalaimani, "Web Mining Research Direction and Open Source Tools", Volume 4, Issue 7, July 2014 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering

[21] Pranit B. Mohata and Prof. Sheetal Dhande, "Web Data Mining Techniques and Implementation for Handling Big Data", IJCSMC, Vol. 4, Issue. 4, April 2015, pg.330 – 334