RECENT DEVELOPMENTS IN WEB USAGE MINING ALGORITHM : THEORY AND APPLICATIONS

Dr.M.CHARLES AROCKIARAJ
Asst.Professor, Dept.of Computer science
Christ College of Arts And Science, Kilachery,
Tamilnadu, India.

ABSTARCT:

World Wide net is a massive store of links and websites. It provides large quantity of knowledge for the web purchasers. The event of net is nice as regarding 1,000,000 pages area unit additional daily. Users' accesses area unit derived in net logs. attributable to the good usage of net, the net log files area unit increasing at a quicker rate and therefore the vary is turning into huge. net Usage Mining relates mining techniques in log knowledge to extract the performance of users that is employed in several applications like Support to the planning, E-commerce, changed services, pre-fetching etc. net usage mining has 3 phases as preprocessing, pattern detection and pattern learning. journal knowledge is usually strident and confusing, therefore preprocessing and pattern analysis is an important technique before mining. For learning patterns gathering area unit to be created professionally. This paper is presents work worn out the net usage mining. Finally a look of varied applications of net usage mining is bestowed. net Usage Mining has change into a dynamic region of study in field of information mining attributable to its crucial values. This paper affords a widespread spoken language of the all the stages in net Usage Mining and issues with connected works during this analysis areas.

***Keywords---Data mining, Web Mining, Web Usage, Mining and Data Preprocessing*.**

I. INTRODUCTION

In this planet of data experience, accessing information is that the chiefly continual task. Day once day we embrace looking varied forms of data that we tend to need and what we tend to achieve? Presently surf the web and also the most popular data is thru America on just one click. Nowadays, World Wide net or internet is taking part in such a vital responsibility in our daily living life that it's very difficult to live while not it. the web has partial a lot of to each users (clients) and also the computer vendors. the web site vendors square measure capable to attain to any or all the intentioned users everywhere the country and globally. They are serves to their purchasers twenty four hours. On the opposite half clients are gaining those services [1]. Information in net Usage Mining is achieved in server logs, proxy logs, browser logs, and simultaneously from a group's information. This information set fluctuate in requisites of the positioning of the data resource, the category of information existing, the realm of people on or once that knowledge was no heritable, or method of execution [1].

Web Usage Mining could be a branch of net Mining, that could be a division of information Mining. The method of mining sizable and precious knowledge or information from huge information is called knowledge Mining. net Usage Mining extracts (mines) the usage attributes of the purchasers of net services. This achieved records will then be helpful during a totally different approaches as an example, examination of false necessities etc [2].Web Usage Mining is well thought-out as a constituent of the massive business Intelligence during a business. it's apply for selecting business advances through the skilful use of net services. It is incredibly important for the client Relationship Management (CRM) as a result of it assurance purchasers performance until the interface among the organization and the shopper is bothered [3].

1.1 Web Content Mining

The web content mining is the procedure to extract the data from web depends on content or web content. Web content information is the gathering of realities a website page which includes the web content. The web content mining might comprise set of content, pictures, audio, video, or structured based records like lists and tables. Utilization of web content has been most part reached broadly. There are a few problems noticed in content mining containing topic revelation and tracking, extraction association designs. The web content mining groups web content and characterize the website pages. There are numerous research works on web mining topic have drawn strongly on strategies designed in different areas, for instance, Information Retrieval (IR) and Natural Language Processing (NLP).

There is some previous important work to extract the information from pictures in the fields of picture preparing and PC vision. In any case, the application of web content mining has a few constraints. The objective of content web mining is to outline the categorization and clusters of web content. Content mining aim is to offer valuable and fascinating prototypes about client prerequisite and contribution behavior. The web content mining commonly focuses on the learning disclosure, which comprises conventional gatherings of content records, collections of multimedia files, for example, pictures, audios, and videos, which are embedded or connected to the Website pages. A portion of the essential web content mining schemes are as per the follows:-

- Unstructured Data Mining.
- Structured Data Mining
- Semi-Structured Data Mining
- Multimedia Data Mining

1.2 Smart Web Spiders

Web spiders are sorts of crawlers which recover the data from the WWW. Commonly, crawlers are utilized to make a duplicate of all visited URL for later processing by an internet searcher. The web spider lists the downloaded web site pages to give quick inquiry. Crawlers can also be used for robotizing maintenance tasks on a Web like a verifying link or validating HTML source.

Spiders utilized different algorithms like breadth-first search, genetic methods to recover the data. Spiders include numerous applications like constructing search databases, individual searches, website page reinforcements, etc. Here, fig.1.3 shows the workflow diagram of smart web spiders in details.

1.3 Semi-Structured Data Mining

Semi-structured information is intersection point form website pages and database networks: it manages with content, and next with the database. The layouts of that information are developing from inflexibly structured relational tables with numerals and strings to empower the normal illustration of compound real-world entities like books, movies, papers, and so forth, without distributing the application writer into expressions. Emergent descriptions for semi-structured information are deviations from the Object Exchange Model (OEM). In OEM, information is developing in nuclear or compound entities. Here, nuclear substances may comprise numbers or string, and complex entities comprise different entities through labeled edges.

2. WEB CLIENT AND SESSION IDENTIFICATION

The activity of web client and session identification is correlated to various web client sessions from the original web server access sheet. Web client's identification is utilized to recognize who get website page and which web site pages are accessed. The target of session identification is to separate the website page accesses of each client at once into unique sessions.

Here, pattern evaluation is another fundamental step in data pre-processing. Because of a few reasons, the outcome in patterns inadequacy, particularly illustration, local cache, operator cache, "post" mechanism, and web server browser's "back" button can influence the result of various imperative accesses not recorded in the web log usage file.

The Uniform Resource Locators (URL) quantity might be recorded and less than the genuine one. Here, the client can be accessed the patterns which are not entirely preserved in the web access log file. To discover the travel pattern, the missing website pages pattern should be attached. The most essential aim of pattern achievement is to achieve work. The better result of data pre-preparing, mining patterns mechanisms should be enhanced.

3. STATEMENT OF THE PROBLEM

Grouping aims to discover fundamental structures in data or document and categorize them into crucial subgroups for further study and investigation. Depending on the Hierarchical Clustering model, the use of Expectation-Maximization (EM) method in the Gaussian Mixture method entirely the constraints and generate the two sub-clusters consolidated when their cover is the biggest is described.

Previous methods avariciously pick the following frequent item set which represents the next group to limit the covering between the documents or data that include both the item set and some remaining item sets. As it were, the clustering outcome based about grabbing the item sets, which in turns relies upon the avaricious heuristic. The technique does not take after a subsequent request of selecting groups.

3.1 Objective of the Study

The primary objective of the proposed approach is to Enhanced Self Organization Map (ESOM) for enhancing the data or document similarities and clustering process of web usage log files. The ESOM algorithm is also reducing the dimensions of data or document, estimates the data or document similarity, and computes the number of clustering data or document using HTML (Hypertext Markup Language) or web log usage documents. The proposed method offers a reliable and effective solution for document similarity prediction and clustering process of web usage log files. The research objectives are as follows:

•To design an Enhanced Self Organization Map (ESOM) algorithm for improving data or document similarities and clustering process of web usage log files

•To preprocess web usage log files and to utilize clustering process.

•To estimate similarity between documents or data, and subsequently formulate new criterion functions for document or data clustering.

•To check how much a data or document similarity measure overlaps with the real class labels and investigate useful similarity measure for data clustering.

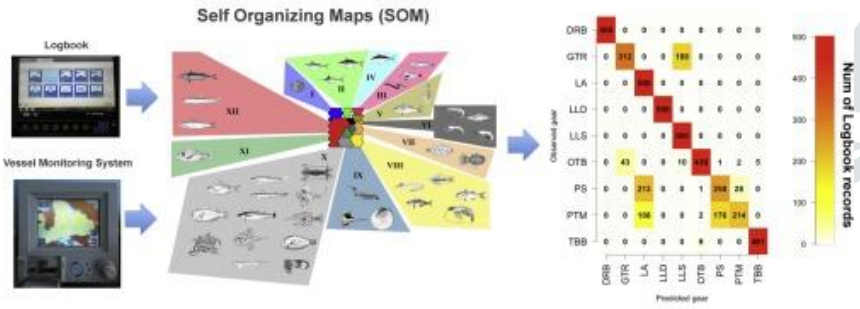
•To optimize ESOM neural network learning and improve the effectiveness of clustering and save computing time of clustering process.

•To reduce the Entropy (E) and enhance F-measure and Similarity compared than their previous methods.

4. SYSTEM ARCHITECTURE OF ESOM ALGORITHM

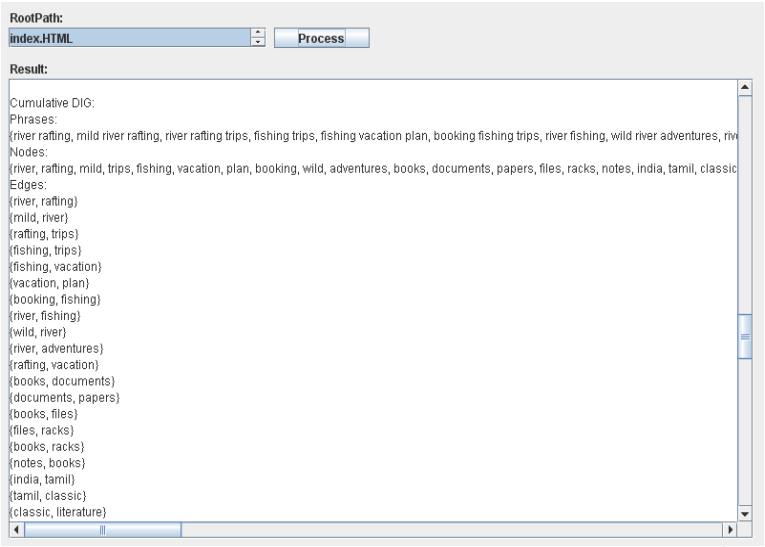
The principle of this analysis is to verify how much a data or document similarity measure overlaps with the real class labels and investigate useful similarity measure for data clustering. The ESOM is mainly focused on analyzing and generating usage of cluster overlapping phenomenon to plan cluster integrating criteria. The system improves the effectiveness of clustering and saves computing time of clustering process.

The primary objective of pre-processing step is improving the quality of features and minimizes the complexity of mining process at the similar time. The pre-processing stage is reading the input weblog usage document and its partition into elements such as tokens, phrases, attributes, etc. The weblog document structure is illustrated as a graphical model.



4.1 Web Content Extraction

The module used to discover the paths or flow of document structure of HTML document or weblog usage documents. The composition containing tokens, phrases, and attributes. Web content extraction is described as the identification or partition of meta-tokens in an HTML document or web log usage file. The meta-token is extracted for all the nodes in the tree structure. Following Fig. shows the content extraction of HTML or web log usage documents.



4.2 The Process of Cluster Formation

Clustering methodology makes a hierarchical decomposition of the provided set of data entities. Based on decomposition method, hierarchical methods are categorized as agglomerative (combining) or disruptive (splitting). The agglomerative method begins with every data point in a separate group or with a specific huge number of groups. Every progression of this strategy combines the two groups that are the most similar. Therefore, after every progression, the aggregate number of groups diminishes. This is frequent until the desired amount of groups is acquired, or just a single group remains. By dissimilarity, the disruptive strategy begins with all data entities in a similar group. In every progression, one group is partition into small groups, until a termination condition holds. Agglomerative methodologies are more generally utilized as a part of training. In this way, the similarities between groups are more researched.

5. ENHANCED SELF-ORGANIZING MAP (ESOM)

The Enhanced Self-Organizing Map (ESOM) is estimating the document or data similarity, reduces the dimensions of data, and computes the number of clustering document or data utilizing weblog usage files. ESOM is utilized for clustering the document or data without knowing the group. The ESOM algorithm offers to plot similarities of document or data by grouping the similar data items in one or two dimensions. The method utilized to reduce the problem of dimensions. The ESOM algorithm integrates document or data clustering process, similarity prediction, and reduction of dimension process.

ESOM is initializing the weight vectors and selecting a sample vector randomly. The method is searching the mapping weight vectors to find the weights that can consider the best sample. Every weight vector has a location; it also has neighboring weights that are close to it. The selected weight is rewarded for performing better than a randomly selected sample vector. In addition to this reward, the neighbors of the weight are also rewarded.

From this step, it increases some small quantity as the number of neighbors, and it noticed every weight could decrease over the time. The entire process is repeated a huge amount of times, generally at least 1000 times. Then, the ESOM algorithm automatically (self-organizing) clusters documents or data for large scale of data.

The ESOM clustering process groups a data over various levels by generating a cluster tree. The tree is not a single set of the cluster, but it is a multilevel set of the cluster. The individual level of clusters is connected as next level of clusters. The method permits to decide the level or scale of the clustering process. The ESOM method identifies the document or data similarities among each pair of vectors in the clustering process.

6. RESULTS AND DISCUSSION

A critical factor in the accomplishment of any grouping or clustering algorithm is the similarity measurement received by the algorithm. With a specific end goal of group similar HTML or web log usage document phrases, proximity metric must be utilized to discover which groups (or clusters) are similar. There is an extensive number of similarity measurements announced in the literature. The idea of similarity is essential in approximately every logical field. For instance, in arithmetic, geometric strategies for evaluating similarity are utilized as a part of investigations of similarity and homothetic within related fields, for example, trigonometry. Topological techniques are connected in fields, for example, semantics. Graph hypothesis is broadly utilized for evaluating coloristic similarities in scientific categorization. The fuzzy set hypothesis has also designed its particular measurements of similarity, which discover application in regions, for example, management, medication, and meteorology. A vital issue in atomic biology is to estimate the sequence similarity sets of proteins.

An analysis or even a listing of the considerable number of utilizes of similarity is not possible. Rather, the perceived similarity is focused on. The degree to which individuals perceive two HTML or web log usage documents as similar generally influences their rational idea and behavior. Negotiations among politicians or commercial administrators might be seen as a procedure of data gathering and evaluation of the similarity of hypothesized and genuine motivators. The valuation for a fine fragrance can be comprehended similarly. The similarity is a core component in accomplishing a comprehension of factors that motivate behavior and mediate influence.

7. CONCLUSIONS

The ESOM algorithm measured the document or data similarity among two web log usage documents or HTML documents. Numerous desirable properties were embedded in this similarity measure. For instance, the data or document similarity measure was symmetric. The presence or absence of an attribute was considered more important than the dissimilarity among the values correlated with a current attribute. The document or data similarity measurement score will be increased, when, the amount of presence or absence attribute pairs minimizes. The two web log usage documents are the smallest amount of similarity score to each other if none of the attributes has non-zero values in both web log usage documents. In addition, it was desirable to consider the value of an attribute for its contribution for the document or data similarity between two web log usage documents. The ESOM algorithm was utilized to measure the similarity between two sets of web log usage documents. The effectiveness of clustering and similarity measurements were enhanced and provided an estimation value to decrease the difficulty involved in the estimation. The efficiency of our document or data similarity analysis was analyzed by applying ESOM algorithm on numerous real-world HTML documents or web log usage documents.

REFERENCES

1. *Abraham, A. and Ramos, V., "Web usage mining using artificial ant colony clustering and linear genetic programming", In Evolutionary Computation, CEC'03, The 2003 Congress on IEEE, Vol. 2, pp. 1384-1391, 2003.*

2. *Abraham, A., "i-miner: A web usage mining framework using hierarchical intelligent systems", In Fuzzy Systems, FUZZ'03, The 12th IEEE International Conference on IEEE, Vol. 2, pp. 1129-1134, 2003.*

3. *Abraham, A., "Meta learning evolutionary artificial neural networks", Neuro computing, Vol. 56, pp. 1-38, 2004.*

4. *Adeniyi, D. A., Wei, Z. and Yongquan, Y., "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics, Vol. 12, No. 1, pp. 90-108, 2016.*

5. *Admiraal-Behloul, F., Van Den Heuvel, D. M. J., Olofsen, H., van Osch, M. J., van der Grond, J., Van Buchem, M. A. and Reiber, J. H. C., "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly", Neuroimage, Vol. 28, No. 3, pp. 607-617, 2005.*

6. *Agrawal, R. and Srikant, R., "Privacy-preserving data mining", In ACM Sigmod Record, ACM, Vol. 29, No. 2, pp. 439-450, 2000.*

7. *Aguilar, J. S., Ruiz, R., Riquelme, J. C. and Giráldez, R., "Snn: A supervised clustering algorithm", In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, Berlin, Heidelberg, pp. 207-216, 2001.*

8. *Amatriain, X. and Pujol, J. M., "Data mining methods for recommender systems", In Recommender systems handbook, Springer, Boston, MA, pp. 227-262, 2015.*