RESEARCH ARTICLE OPEN ACCESS

Methodologies of Adopting Algorithm for Network Convention Mining Through Neural System

[1.]C.Sadhana , [2.]Dr.L.Mary Immaculate Sheela
[1]*Research Scholar Computer Science and application, St peters University*
[2]*Professor, Department of Computer Application R.M.D Engineering College.*
Corresponding author: C.Sadhana

ABSTRACT: Web mining is an essential part of data mining. Web mining adopts a great part of the data mining mechanisms to discover potentially useful information. Web mining analysis depends on three common set of information such as patterns, shared content and inter-memory association link structure relating to three subsets in web mining namely Web usage mining, Web content mining, Web structure mining respectively. Data grouping or clustering is a standard mechanism for statistical data analysis, which is utilized as a part of numerous fields, consisting machine learning, data mining, pattern recognition, image analysis, and bioinformatics.

Clustering or Grouping aims to discover essential structures in data or document and arrange them into vital subgroups for further study and analysis. Existing techniques greedily select the following frequent item set which illustrates the following group to constrain the covering among the documents or data that comprise both the item set and some remaining item sets. As it were, the grouping or clustering outcome depends on the demand of grabbing the item sets, which in turns based on the avaricious heuristic. The technique does not take after a subsequent request of selecting groups or clusters.

To overcome the above issues, a novel approach Enhanced Self-Organizing Map (ESOM) is proposed for document clustering which offers highest effectiveness and performance. The proposed system is estimating similarity between documents or data and subsequently formulates a new criterion functions for a document or data clustering. 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.

Keyword: ESOM, data clustering, Web mining, , image analysis, bioinformatics

---

Date Of Submission: 25-05-2019 Date Of Acceptance: 07-06-2019

I. INTRODUCTION

1.1.1 Web Mining

The World Wide Web has become one of the essential media to store, share, and allocation of information. The fast development of the web has given a great chance to study client and framework action by investigating Web server log files. The Web Mining is the process of finding potential helpful and already unknown data from the Web server log information. The strategy is utilized to crawl the different URL (Unified Resource Locator) to recover the needed data, which empowers an individual or an organization to support commerce, comprehend the market flow, and new promotions floating on the Internet, and so forth. There is a developing pattern among different businesses and people to collect data through the internet. Here, URL can utilize the data depending on their advantage. Web mining is the usage of data mining method, which automatically finds and extracts data from web content and web services. Here, web mining is characterized by the following name as Web action, web server logs, and Web program action tracking.

1.1.2 Types of Web Mining

This section introduces the web mining categorization and their characteristics like web content, structure, and usage mining according to their application and feature of web mining. This section also expresses the information about structured, unstructured, and semi-structured data.

1.1.3Web 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

1.1.3.1 Unstructured Data Mining

It is one of the schemes for web content mining which is unstructured and helps to group cluster the huge volume of textual information. A number of the web site pages are in the form of content. According to this data, the inquiry can, and it can be recovered from web or internet. The unstructured data mining isn't essential that recovered data should be simply important, it might be unknown data. Here, a few instruments examined to recover related data from web or internet.

1.1.3.1.1 Textual Mining in Web Contents

Text or content Mining is a sub-area of data mining method. Data recovery from Web pages, it is a challenging task.Since, it includes various tokens, which are needed to recognize that specific tokens.Here, large amount of tokens have several issues due to complexity in process. To overcome the accuracy issues, the appearance of current devices are there in particular as a Support Vector Machine, Decision trees the outcomes. Fig.1.2 illustrates the information disclosure process in content or text mining with web contents.

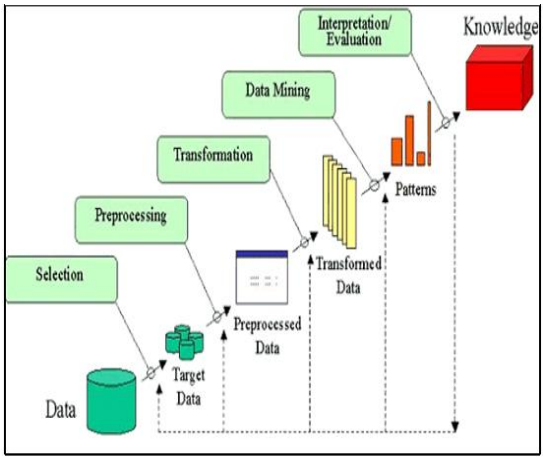

Fig.1.2 Information Disclosure Process in Content or Text Mining With Web Contents

1.2 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.

1.3 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 anEnhanced 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.

1.4 Methodologies of the Study

The thesis introduces Enhanced Self Organization Map (ESOM) to improve document clustering and offers highest effectiveness and performance. The proposed system is estimating the similarity between documents or data and subsequently formulates new criterion functions for a document or data clustering. 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 algorithm 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.

II. ENHANCED SELF ORGANIZING MAP ALGORITHM

In this chapter, a new system model named as Enhanced Self Organizing Map Algorithm (ESOM), according to its advantages and implementation part is presented in detail. Here, implementation of the model procedure is categorized into following sections namely: Pre-Processing of Document or Data, Web Content Extraction, Relevant Data or Document Identification, Cluster Formation, Formation of Histogram Document or Data Similarity Prediction and Enhanced Self-Organizing Map (ESOM) algorithm.

2.1 System Architecture Of ESOM Algorithm

The ESOM exhibits the step-by-step workflow procedure which improves clustering efficiency and data similarity. The methodology provides document clustering and offers highest effectiveness and performance. The proposed system is estimating similarity between documents or data and subsequently formulates new criterion functions for a document or data clustering.

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.

2.1.1 Pre-processing of document or data

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.

The frequency of every measure the weblog document features and weights, and it removes non-informative features such as numbers, stops words and special characters. The pre-processing step also utilized for weighting the weblog usage documents and their similarities. The pre-processing of data contains tokenization, stop words removal and stemming processes. Fig.2.1shows the data pre-processing steps.

2.1.1.1 Meta-Tokenization

Meta-Tokenization is the way of breaking a flow of content into words, expressions, symbols, images, or other significant components called meta-tokens.The objective of the meta-tokenization is the investigation of the strings in a sentence. The list of meta-tokens progresses input for additionally preparing, for example, parsing, or text mining. Tokenization is helpful both in linguistics (where it is a type of content division), and software engineering, where it elements of the lexical examination.

Textual information is just a block of characters at the beginning. All procedures in data recovery require the expressions of the data collection. Subsequently, the prerequisite for a parser is a meta-tokenization of web log usage documents. This may sound inconsequential, as the content is already stored in machine-comprehensible formats. A few issues are still left, similar to the elimination of punctuation marks. Different characters like brackets, hyphens, and so on require processing too. Besides, meta-tokenizer can provide for reliability in the web log usage documents. The primary utilization of meta-

tokenization recognizes the essential keywords. The irregularity can be a unique number and time formats. Other issues are abbreviations and acronyms, which must be changed into a standard form.

2.1.1.2 Eliminating Stop words

Numerous words in web log usage documents repeat very recurrently, however, are trivial as they are utilized to combine words in a sentence. The web log usage document word is usually comprehended that stop words do not add to the specific context or word of textual documents. Because of their high recurrence of the event, their occurrence in content mining presents an obstacle in understanding the substance of the web log usage documents.

2.1.3 Cluster Formation

Clustering is a separation of data or document into groups of similar entities. Representing the data or document through the smaller amount of clusters radically drops specific fine details, but accomplishes simplification. The same data or documents are grouped jointly in a cluster if their cosine data or document similarity computation is less than a particular threshold. Clusters formed by considering the similarity of the documents. Fig.3.6 exhibits the cluster formation of HTML or Web log usage documents.

2.4 Proposed Summary

The system has highly concentrated on similarity between documents or data and subsequently formulates new criterion functions for a document or data clustering. 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. Hence, the proposed design explains the diagrammatic representation of the following data. The developer to get a clear concept of the logical and the analytical view of the proposed ESOM method to be utilized in the real-time applications utilizes the design. The research study addresses the result and discussion along with a comparative analysis of proposed ESOM mechanism.

III. RESULTS AND DISCUSSION

3.1 ESOM Algorithms

3.1.1 Programming Environment

The ESOM methodology deployed in JAVA programming environment. The implementation of the ESOM method is deployed on a laptop with Intel Dual Core Processor (1.836 Hz), 2Gigabyte memory, and Window 7 Ultimate system. The ESOM methodology is implemented with the Java Development Kit (JDK) 1.8 JAVA (with JDK 1.8) [102] and the NetBeans 8.0 development environment. The Proposed methodology is estimated with three kinds of clusters such as 3 clusters, 4 clusters, and 5 clusters data.

3.1.1.1 JAVA

The Java programming language is a high-level programming language that can be characterized by all of the following buzzwords like Simple, Architecture neutral, Object-oriented, Portable, Distributed, High performance, Interpreted, Multithreaded, Robust, DynamicandSecure. The Java programming language is an extraordinary language in which a program is both compiled and interpreted. At first, it transfers a program into a middle-level language called Java byte codes in which the interpreter interprets the platform of independent source code with the programming compiler. The interpreter parses and runs each Java byte code instruction on the computer. Compilation process starts when interpretation occurs each time the program is executed.

Every Java interpreter, whether it is a development tool or it comes with java web browser. A web browser can be run with applets, is an implementation of the Java Virtual Machine (JVM). Java byte codes assist in constructing write once, anywhere possibilities. JVM can compile the program into byte codes on any platform that has a Java compiler. The byte codes can be run any java implemented Virtual Machine (VM) environment.

In details, it means that as long as a computer has a Java virtual machine. The same program is written in the Java programming language can run on Windows 2000, a Solaris workstation, or on an iMac. The Java Application Programming Interface (Java API) has introduced in java through a virtual machine. The Java API is bone of java platform, which is ported to many hardware-based platforms. The Java API has the large collection of a built software component, which produces many useful capabilities, such as Graphical User Interface.

3.2.2 F-Measure

The F-measure is a collection of precision and recall, here assumed to clustering estimation purposes. Precision (i, j) is the proportion of the amount of related weblog usage documents or data to the entire number of weblog usage documents recovered by j^{th} cluster belongs to i^{th} class for an inquiry. Recall (i, j) is the proportion of the amount of related weblog usage documents or data

retrieved by j^{th} cluster belongs to i^{th} class for a query to the whole amount of related weblog usage documents or data in the total aggregation. The proposed method is described as a mathematical model for F-measure in equation (4.2). F-measure (F-m) is evaluated as:

$$F-m=\frac{1}{n}\sum_{i}^{k}\frac{n_i}{\max_j F(i,j)} \quad (3.1)$$

$$F(i,j)=\frac{2*\mathrm{Re}call(i,j)*\mathrm{Pr}ecision(i,j)}{\mathrm{Pr}ecision(i,j)+\mathrm{Re}call(i,j)}$$

Where F (i, j) is every single cluster and n_i is the total number of weblog usage document for i^{th} class. A higher value of F-measure specifies better clustering for weblog usage document or data.

3.2.3 Similarity

In the absence of a few external data such as class tokens, the cohesiveness of clusters can be utilized to compute cluster similarity of weblog usage documents or data. The cluster cohesiveness computation is utilized the weighted similarity of the internal cluster similarity of weblog usage documents or data. The proposed methodology is defined as a mathematical model for data similarity or document similarity in equation (4.3). The similarity (S) is estimated as:

$$S=\frac{Number of Matching documents or data}{Number of Documents} \quad (3.2)$$

Where F (i, j) is every single cluster and n_i is the total number of weblog usage document for i^{th} class. A higher value of F-measure specifies better clustering for weblog usage document or data.

IV. STATISTICAL ANALYSIS OF HTML OR WEB LOG USAGE DOCUMENTS

To vast numbers of web users the phrase "relevant ranked search outcomes" is a mystery. A better phrase may have been "statistically significant query results." Adopting such a strategy, the application of statistical investigation against strings has its data recovery benefits over straight Boolean logic. For instance, table 4.1 shows the three HTML documents or web log usage documents comprising of various strings.

Table 4.1: HTML or Web Log Usage Documents having specific String

Web Log Usage Document 1	**Web Log Usage Document 2**	**Web Log Usage Document 3**
String	**String**	**String**
Airplane	book	Building
Blue	car	Car
Chair	chair	Carpet
Computer	justice	Ceiling
Forest	milton	chair
justice	newton	cleaning
Love	pond	justice
Might	rose	libraries
Perl	shakespeare	newton
Rose	slavery	perl
Shoe	thesis	rose
Thesis	truck	science

A search for "rose" against the corpus will revisit three hits, yet which one should begin reading from the latest HTML or web log usage file. The HTML or web log usage document is used by a specific author or in a specific format. Still, the corpus comprised 2,000,000 HTML or web log usage document and a search for "rose" restored an unimportant 100 the issue would remain. Which ones would it be a good idea for us to invest our significant time in accessing? Sure, it could limit our search in any quantity of ways. However except we are doing a known string search it is very likely the search outcomes will return more than to utilize and data proficiency skills will go up until this point. Ranked search outcomes, a list of hits depend on phrase weighting has proven to be a successful method for addressing this issue. All it needs is the application of fundamental arithmetic against the HTML or web log usage documents being searched.

4..1 Measurement of Document or Data Similarity

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 homothety within related fields, for example, trigonometry. Topological techniques are connected in fields, for example, semantics. Graph hypothesis is broadly utilized for evaluating cladistic 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.

Naturally, similarity has additionally assumed a primarily significant part in psychological experimentations and hypothesizes. For instance, in numerous testing's, individuals are requested to create direct or indirect judgments about the similarity of sets of HTML or web log usage documents. A variety of experimentation mechanisms are utilized in these investigations. Yet, the most widely recognized are to ask HTML or web log usage documents whether the phrases are the similar or different, or to request that they create a number, between 1 and 7, that matches their sentiments about how similar the HTML or web log usage document phrases appear (e.g., with 1 meaning dissimilar and 7 meaning fundamentally very similar).

The idea of similarity also plays a critical however less direct part in the modeling of numerous other psychological tasks. This is particularly valid in hypotheses of the detection, identification, and classification of HTML or web log usage documents, where a general assumption is that the more significant the similarity among a couple of HTML or web log usage document phrases, the more probable one will be mistaken for the other. Similarity also plays a vital part in the modeling of preference and preferring for HTML or web log usage document phrases.

4.2 Simulation Results

The ESOM algorithm is computed with the previous approach; namely, K-means discussed by Karol et al., 2013 allocates every point to the cluster whose center (also called centroid) is closest. The centroid is the average of the considerable number of points in the clustering. The centroid is coordinates are the arithmetical mean for every measurement independently over all the points in the group or cluster. However, it does not yield a similar outcome with every run, because of the subsequent clusters based on the initial random tasks. The task reduces intra-group variation, however, does not guarantee that the outcome has a worldwide smallest amount of variation.

V. CONCLUSIONS AND SCOPE FOR FURTHER STUDY

5.1 Conclusions

An Enhanced Self-Organizing Map (ESOM) algorithm presented for the document for data similarity, reduced the dimensions of data, and computed the number of clustering document or data utilizing weblog usage files. The algorithm has improved the effectiveness of clustering and saved computation time of clustering process. The ESOM algorithm offered to plot similarities of document or data by grouping the similar data items in one or two dimensions. The ESOM algorithm was automatically (self-organizing) clustered the documents or data for large scale of data. The ESOM clustering process groupedthe data over various levels by generating a cluster tree. The ESOM method identified the document or data similarities among each pair of vectors in the clustering process.

5.2 Scope for Further Study

During the research work, several potential ways of future research were identified. In future, this research work can be extended with document or data similarity prediction for the pdf, text, and word documents, etc. Choosing distinctive dimensional space and recurrence levels prompts to diverse precision rate in the grouping or grouping outcomes. How to extract the most effective features sensibly will be examined in future work. In future, it is planned to illustrate the ESOM algorithm depends on MapReduce mode, which can manage with a vast amount of datasets with a massive amount of hubs on Hadoop platform.

Other future work, it will be present in external assets, for example, Wordnet and Wikipedia, to compute the semantic sentence similarity. The ESOM algorithm can resolve the issues of the synonym and the multi-verbal word. In future, it can manage with other language issues. The algorithm is utilizing distance function of clustering algorithms, which help semantic strategies, and resolves the stemming issue for the innovative words.

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.

C.Sadhana" Enhanced Self Organizing Map Algorithm for Web Usage Mining Through Neural Network" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.06, 2019, pp. 01-07