Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Enhanced Clustering Approach for Online and Offline web Content Search

Chhaman Lakheyan *,
*Student of M.Tech Computer Science,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Usvir kaur
Assistant Professor,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Abstract- This paper about extraction of data related to biomedical text. This approach is a concept of online and offline database. In traditional method, when the user is online then they get the data easily but when there is no connection of internet then user cannot get the data. In this current approach user can get the data in offline mode from the local database. In local database, all data saved that the user previously used. When user enter the keyword, that already used then it will get from local database inspite of server or global database. The local database contains data that is relevant to the user. So the results can conclude that the time taken by this approach is less.

1. INTRODUCTION[1,2]

***Web Usage Mining*: -- Pattern Discovery and its applications**

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses. A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future[2].

Why Web Usage Mining-In this paper, we will emphasize on Web usage mining. Reasons are very simple: With the explosion of E-commerce, the way companies are doing businesses has been changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us a cost-efficient and effective way of doing business. The growth of some E-businesses is astonishing, considering how E-commerce has made Amazon.com become the so-called "on-line Wal-Mart". Unfortunately, to most companies, web is nothing more than a place where transactions take place.[1] They did not realize that as millions of visitors interact daily with Web sites around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on.

CLUSTERING AND TECHNIQUES-

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).[2]

It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Clustering is the process of organizing objects into groups whose members are similar in some way. It can be considered the most important unsupervised learning problem which deals with finding a structure in a collection of unlabeled data. [4].

A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Hard clustering is the techniques in which any pattern can be in only one cluster at any time. Soft clustering is the technique which permits patterns to be in more than one cluster at any time.

III. TEXT MINING

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. (Nathan Harmston, Wendy Filsell and Michael P.H. Stumpf ,6th August 2010)

Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).[5]

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index Text mining is defined as the automatic discovery of new, previously unknown, information from unstructured textual data. There are various clustering approaches that can be applied to cluster the biomedical keywords extracted from full text articles, some of them are k-means, k-median, Hierarchical Clustering Algorithm, Nearest Neighbor Algorithm etc. Here we are using modified fuzzy C mean clustering algorithm.

2. Releated Work

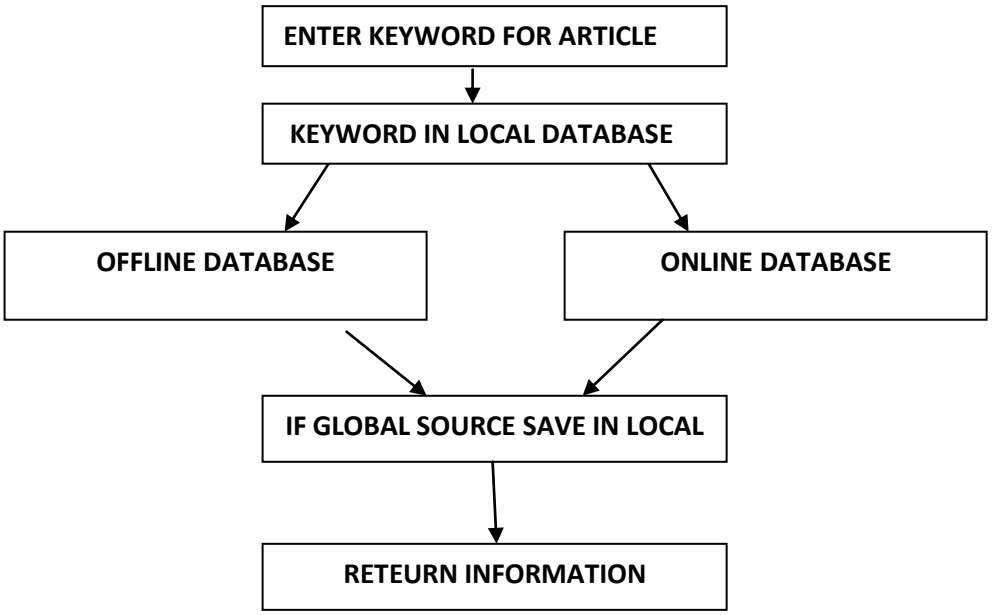
Researcher find in the preveious work that Extraction of text from biomedical literature is an essential operation. Given that there have been many text extraction methods developed; this paper presents a novel technique that employs keyword based article clustering to further enhance the text extraction process. The development of the proposed algorithm is of practical significance; however it is challenging to design a unified approach of text extraction that retrieves the relevant text articles more efficiently. They proposed algorithm, using data mining algorithm, seems to extract the text with contextual completeness in overall, individual and collective forms, making it able to significantly enhance the text extraction process from biomedical literature.

3. Our Problem Statement with Purposed Work

In this work of thesis the case where user enters the old keyword again, then the user will be awarded with the data which is available at the offline data-base. However, for the sake of ease, user is developed that in case the user is offline data required can be extracted from offline database.

The work will be processed in a robust form and will reduce the overhead of the user in search the relevant information from web or offline databases. The tactic which is used is called fuzzy C means gathering in pre-processing of data. The data is searched by the user and is deposited in the offline database. If the user is not associated to internet then query will be forwarded to offline database removing the information from offline database.

4. METHODLOGY OF PURPOSED WORK



This logic is implemented with FUZZY C MEANS ALGORITHMS as:

The Algorithm Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function. The fuzzy c-means algorithm is one of the most widely used soft clustering algorithms. It is a variant of standard k-means algorithm that uses a soft membership function.

Fuzzy C-Means (FCM) clustering algorithm is one of the most popular fuzzy clustering algorithms.
FCM is based on minimization of the objective function Fm(u, c):

$$F_m(u,c) = \sum_{k-1}^{n} \sum_{i-1}^{C} (u_{ik})^m d^2(x_k, c_i)$$

Clustering is the process of organizing objects into groups whose members are similar in some way. It can be considered the most important unsupervised learning problem which deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Hard clustering is the techniques in which any pattern can be in only one cluster at any time. Soft clustering is the technique which permits patterns to be in more than one cluster at any time.There are various clustering approaches that can be applied to cluster the biomedical keywords extracted from full text articles, some of them are k-means, k-median, Hierarchical Clustering Algorithm, Nearest Neighbor Algorithm etc. Here we are using modified fuzzy C mean clustering algorithm. Here the proposed algorithm is responsible for extracting keywords present in the full text biomedical article store these keywords in a relation. Then the actual work of algorithm begins, it starts clustering of keywords. The algorithm initially picks some keywords that are extracted. It groups the full text articles based on these keywords. It means each cluster contains only those articles which contain that keyword as their part. Then it starts using fuzzy C mean clustering to combine the clusters together on some similarity measure. Here we combine two clusters if their similarity measure is greater than or equal to a specified threshold value. The proposed Algorithm repeats this process until no more changes are made to the clusters. Finally the proposed algorithm stores all the clusters in an xml file.

5. RESULTS AND COMPARSION

The results show the time of local database vs global data base, error rate and precision.

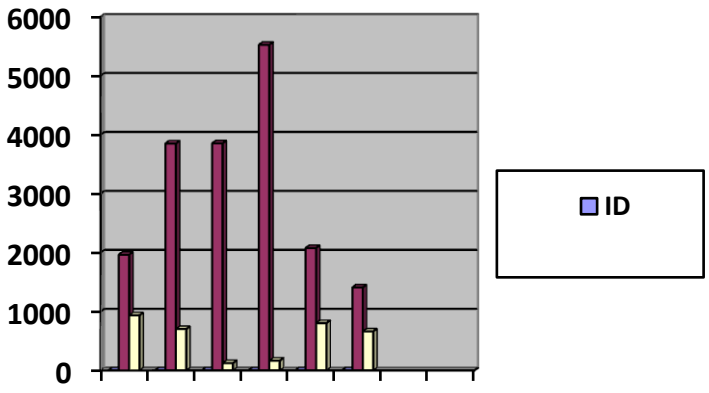
Precision: Evaluation of results via precision: PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.
Precision : Relevant link/(relevant link+total link)

Accuracy: Accuracy refers to the closeness of a measured value to a standard or known value. For example, if in lab you obtain a weight measurement of 3.2 kg for a given substance, but the actual or known weight is 10 kg, then your measurement is not accurate. In this case, your measurement is not close to the known value.

Precision: Precision refers to the closeness of two or more measurements to each other. Using the example above, if you weigh a given substance five times, and get 3.2 kg each time, then your measurement is very precise. Precision is independent of accuracy. You can be very precise but inaccurate, as described above. You can also be accurate but imprecise. In below this table, there shows six keywords and their url display time from local and global database. When user enter the keyword, if it is present in local data then it will come from here otherwise it will come from global data base. Here comparison shows that the local database access time is less than the global database.

Table 1. Display URLTimes of Local database and global data base.

ID	Keywords	Global Database urltime(nano sec.)	Local database urltime(nano sec.)
1	Skin	1962	936
2	Lung	3849	703
3	Cough	3852	125
4	Leg	5523	166
5	Joint	2076	803
6	Dark	1408	662



In this graph shows the local and global database access time. In x-axis URLTime shows in anon seconds and in y-axes shows IDs. There are two url time one for local database and other for global database. Their is comparison between global and local database time that it take for getting information from the databases. When user enter keyword if it is

present in local data base then data extract from here and its time is very less than the global database access time. If it is not present in the local database then it will extract from the global database, its access time is more than the local database access time

Error Rate: When user searching through the database, the sites mismatch our keyword searching

6. CONCLUSION AND FUTURE SCOPE

In our proposed work we are moving forward to establish the data pulling out method by FCM algorithm on all sort of data bases on both online and offline modes. Our work will be targeting on the biomedical data withdrawal with the notion of data mining and web mining. In this work of thesis for the benefit of user is developed that in case the user is offline and required data then the data can be extracted from offline database. If user enters the old keyword again then user will get the data from offline data-base otherwise data will get from online database and also save this data in the local database for future use. These results will fast extraction of data as data will be available and stored in offline data base.

Future scope

We can give more research on clustering algorithms to make them more better as data bases with large records are increasing and is required now. We can also integrate the FCM with Ranking algorithms to make the Fuzzy C-means clustering more efficient.With the ever increasing demand of the data bases in almost every fragment of our lives, there is an urgent need of researching on the various kinds of data bases available. As every one is aware of the fact that clustering algorithms are more efficient and appropriate techniques to accomplish them, we are heading forward to conduct further exploration in the field. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). We are efficient enough of availing more research on clustering algorithms to turn them better as data bases with large records are mounting and is required now. We can also put together the FCM with ranking algorithms to make the Fuzzy C-means clustering more efficient.

References

[1] M. La vanya & Dr.M.Usha Rani "Vision-Based Deep Web Data Extraction for Web Document Clustering".

[2] Pranit C. Patil1, (M.Tech. Computer, Department of Computer Technology, Veermata Jijabai Technological Institute, Mumbai-19, Maharashtra, India) Pramila M. Chawan2, (Associate Professor, Department of Computer Technology, Veermata Jijabai Technological Institute, Mumbai-19, Maharashtra, India), Prithviraj M. Chauhan (Project Manager, Morning Star India Pvt. Ltd., Navi Mumbai, Maharashtra, India) "Extracting Information From Tables of HTML".

[3] Nathan Harmston, Wendy Filsell and Michael P.H. Stumpf, "What the papers say: Text mining for genomics and systems biology".

[4] Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii and Douglas B. Kell,"Event extraction for systems biology by text mining the literature".

[5] Tamanna Bhatia, "*Link Analysis Algorithms For Web Mining*" ISSN :2229 - 423 (Print) |ISSN : 0976 - 8491 (Online) IJCST Vol. 2, Issue 2, June 2011.

[6] Sumit Vashishta, Dr. Yogendra Kumar Jain*" Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm"(IJACSA*) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011

[7] Hong-Jie Dai1,2, Yen-Ching Chang1 et al **"***New Challenges for Biological Text-Mining in the Next Decade***"** journal of computer science and technology 25(1): 169–inside back cover Jan. 2010

[8] D. Akila, Dr. C. Jayakumar, *" ENHANCED BONDING BASED WEB PAGE INFORMATION RETRIEVAL USING CLUSTERING ALGORITHM*".

[9] Neha Verma, Aditya Verma, Rishma and Madhuri, "Efficient and Enhanced Data Mining Approach for Recommender System".

[10] Rimmy Chuchra, M.tech (Computer Science) ,Lovely Professional University Phagwara, India, "Performance Analysis & Comparison b/w Enhanced K-Means & Orthogonal Partitioning (OC), based on proposed".

[11] R. Sagayam, 2 S.Srinivasan, 3 S. Roshni 1, 2, 3 Department Of Computer Science Govt. Arts College (Autonomous) Salem-7 2 Periyar University Salem-636011, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques."

[12] John Atkinson-Abutridy, Chris Mellish, and Stuart Aitken, Uiversity of Edinburgh, "Combining Information Extraction with Genetic Algorithms for Text Mining"

[13] Kalyani M Raval (B.Com, MSc IT), Lecturer in B.Com MIP and PGDCA M. J. College of Commerce, Maharaja Krishnakumarsinhji Bhavnagar University Bhavnagar, International Journal of Advanced Research in Computer Science and Software".

[14] Sriram Krishnan San Jose State University, "Clustering Algorithm for Enhanced Bibliography Visualization" .

[15] Nam Pham*, Bogdan M. Wilamowski Auburn University, Department of Electrical and Computer Engineering, Auburn AL, U.S.A, "IEEE Article Data Extraction from Internet" .

[16] International Journal of Scientific & Engineering Research, Volume 3, Issue 2, February -2012 1 ISSN 2229-5518, "Enhanced Content Based Image Retrieval Using Multiple Feature Fusion Algorithms" .

