Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFECTIVE FEATURE EXTRACTION FOR DOCUMENT CLUSTERING TO ENHANCE SEARCH ENGINE USING XML

P.AJITHA, DR. G. GUNASEKARAN
Research Scholar, Sathyabama University, Chennai, India
Principal , Meenakshi College of Engineering, Chennai, India
E-mail: hannahgracelyne @gmail.com, gunaguru@yahoo.com

ABSTRACT

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Clustering is done using lingo algorithm by extracting the data contents in the document. The data is stored in XML, which manages large volume of data. Lingo combines several existing methods to put special emphasis on meaningful cluster descriptions, apart from identifying document similarities. The steps involved in this process are designing the term-document matrix and then extracting the frequent phrase using suffix arrays. Readable and unambiguous descriptions of the thematic groups are an important factor of the overall quality of clustering. The Lingo algorithm consist of five phases, they are Pre-processing, Extraction of Frequent phrase, Induction of Cluster label, Discovery of Cluster content, Final cluster formation.

Keywords: *Lingo, XML, SVD, LSI, Phrase matrix.*

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into meaningful information to improvise profit by cutting overhead cost. Data mining software is one such analytical tool for analyzing data from various perspectives, to classify it, and cluster the contents. Technically speaking text mining is the phenomenon of finding correlations among various elements in Structured and Unstructured databases.

Text mining process is an interdisciplinary subfield of Data Mining. It deals with the computational process of discovering patterns in unstructured datasets, involving the usage of areas like database systems, statistics, machine learning, and artificial intelligence. Text mining process does extraction of information from an unstructured data set and converts it into a comprehensive structure for further use.

It involves database and data management aspects, inference considerations, data pre-processing model, complexity considerations, visualization, interestingness metrics, complexity priorities, post-processing of explored structures, and online updating. The term data mining is a buzzword which is frequently misused to mean any form of large-scale data. It is generalized to any kind of computer decision support system, including machine learning, business intelligence and artificial intelligence.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies. The data mining step might find multiple groups in the data that can be used to retrieve more precise prediction results by a decision support system. The data collection, data preparation, result interpretation and reporting belong to overall KDD process and not to data mining. Data mining works to analyze data stored in data warehouses that are used to store the data being analyzed. Data would be screened in Data pre-processing phase and is a vital phase in data mining process, since not doing this may produce unpredictable results.

2. EXISTING METHODOLOGIES

Existing feature selection methods predominantly concentrates on identifying most relevant features. Feature selection plays a major role in document clustering. The number of features in a document varies from hundreds to thousands.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

But an efficient algorithm must select the most relevant features. In [12], the author mentioned that in addition to feature relevance, feature redundancy also plays major role in clustering. So a method to extract feature redundancy and differentiate it with feature relevance was identified and a test to correlate the effectiveness and efficiency for feature relevance and feature redundancy was conducted.

A feature instance is typically described as an assignment of values to a set of features and to one of the possible classes in class label C. The task is to induce a hypothesis that accurately predicts the labels of novel instances. The classification of document is predominantly based on feature extraction. Generally more features will increase the processing time and in practice they may cause confusion with the training data. Feature Extraction plays a major role in text mining and an active field in research and development for decades in statistical pattern recognition, machine learning, data mining and statistic. For example let G be some subset of and be the value vector of F. The goal of feature selection can be stated as selecting G as a minimum subset in such a way it is more or less equal or close to F. It is the probability distribution of different classes given the feature values in G is the original distribution given the feature values in F. From the class concept perspective, it is order of the day that many features are not informative since they are either irrelevant or repetitive. In alternative term, with few relevant and non-redundant features learning can be more effective and efficient. However, probable exponential growth of feature subsets with growing number of dimensions is possible.

Finding an optimal subset is usually challenging and it is found to be NP hard for many problems related to feature selection. Wrapper model and Filter model are identified as two types of Feature selection methods. The Wrapper model determines the goodness of the selected subsets, by making use of predictive accuracy of a predetermined learning algorithm. As the number of features increase, these methods turn out to be computationally expensive. In contrast the Filter model separates feature selection from classifier learning by selecting feature subsets that are independent of any learning algorithm. It in fact depends on training data metrics such as, distance, dependency, consistency, information and dependency. One other key problem of feature selection is Search. To balance the computational efficiency and trade off of result optimality, various search strategies such as heuristic, complete, and random search have been researched for evaluation of generated candidate feature subsets.

Robust intelligence is best known for its ability to learn because of the evolving and adapting capabilities. Machine learning aids computer systems to learn, and improve efficiency. Feature selection facilitates machine learning by aiming to remove irrelevant features. Feature interaction presents a challenge to feature subset selection for classification. This is because probability of correlation with the target concept increases when it is combined with some other features. However the unintentional removal of these features may result in poor classification performance. Also having feature interactions in common is computationally challenging. However, the existence of feature interaction in a wide range of real-world applications demands practical solutions that can reduce high-dimensional data while perpetuating feature interactions.

The challenge here is to design a special data structure for feature quality evaluation and to employ an information-theoretic feature ranking mechanism to efficiently handle feature interaction in subset selection. Experiments were conducted to evaluate the approach by comparing with some representative methods, perform a lesson study to examine the critical components of the proposed algorithm. Also to obtain insights, and investigate related issues such as ranking, data structure, scalability, and time complexity in exploring interacting features. The rapid advance of computer technology and the ubiquitous use of them have provided a platform for humans to expand capabilities in Services, Production, research, and communications. In this process, immense quantities of high-dimensional data are accumulated challenging state-of-the-art machine learning techniques to efficiently produce useful results.

Machine learning can benefit significantly from using only relevant data in terms of learning performance and learned results such as improved comprehensibility. A widely applied technique for finding relevant data is feature selection, which studies algorithms of finding relevant features among many extant ones. Feature selection finds its pervasive application in many real-world domains: ranging from computational biology, biomedicine, text processing, image analysis, to services. Successful applications of feature selection also bring about new a challenge, one of which are to search for interacting features that often function together and is elusive to efficient solutions. In the case of high-throughput microarray data, for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

instance, finding interacting features is to search for a small number of pertinent genes from hundreds of thousands of ones, reflecting the immune response mechanism involving both antigen presentation and immune operate some pathways. The overarching need for finding interacting features goes beyond the current methods to keep pace with data of increasing dimensionality.

The new demands motivate us toward fundamental research in the design and development of novel and integrated methods of intelligent search for interacting features. Once over fitting happens, the gap between the estimated and the true accuracy becomes big, and the performance of the classier can deteriorate significantly. Over fitting poses a serious problem especially for learning on data with high dimensionality. In order to find probably approximately correct hypothesis, PAC learning theory gives a theoretic relationship between the number of instances needed in terms of the size of hypothesis space and the number of dimensions.

3. PROPOSED METHODOLOGIES

In proposed system, Clustering is based on the content of the data using lingo algorithm which is in XML and can manage large volume of data. In addition to discovering similarities among documents, several existing methods are combined by Lingo to stress on the importance of meaningful cluster descriptions. The steps involved in this process are designing the term-document matrix and then extracting the frequent phrase using suffix arrays

The Lingo algorithm consist of five phases, they are Pre-processing, Extraction of Frequent phrase, Induction of Cluster label, Discovery of Cluster content, Final cluster formation.

1. Pre-processing

The pre processing phase includes stemming, stop words and stop labels. Stemming is the process of folding grammatical variations of words into their "base" forms. Carrot2 tool uses built in set of stemmers. Stop words include the terms that are meaningless in the language (i.e. "is", "this" in English).It is often desirable to filter out certain frequently occurring expressions that should not be considered as cluster. This resource provides means of avoiding such cluster labels.

2. Extraction of Frequent Phrase

The frequently occurring terms and phrases in documents are found in this phase. There are some predefined thresholds given. Should frequency of the terms and phrases exceed these threshold values then it can be considered as a frequently occurred terms and phrases. The advanced method adds an extra step that involves finding the synonyms of the frequent terms and phases.

3. Induction of Cluster Label

This phase of lingo first computes the term document matrix for the frequent terms. Once done, using singular value decomposition, the term document matrix is decomposed. Then using this decomposed matrix it finds the abstract concepts from document and then apply phrase matching. The abstract concept can then be used as cluster labels according to some thresholds.

4. Discovery of Cluster Content

This phase of lingo then assigns the content of the document or the input snippets to the clusters which are labelled in previous phase.

5. Final Cluster Formation

Finally, the clusters are scored using label score and member count. Then clusters are sorted according to these cluster scores.

Advantages of proposed methodology are Low Time Consuming process and Effective search is achieved.

3.1 Pseudo-Code of the Lingo Algorithm

1. D ← input documents (or snippets)

STEP 1 Pre-processing

2. For all d □ D do
3. Identify the Sentences of d.
4. If the language of document is identified then
5. Remove the stop word and apply stemming algorithm in the document d;
6. End if
7. End for

STEP 2 Extraction of Frequent phrase

8. Concatenate all documents;
9. Pc ← discover complete phrases;
10. Pf ← p: {p ∈ Pc ∧ frequency(p) > Feature Frequency Threshold};

STEP 3 Induction of Cluster label

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

11. B ←feature-document matrix of features with occurrence greater than the feature frequency Threshold and not belong to stop words.

12. Σ, U, V ← SVD(B); {Product of SVD decomposition of B}

13. G ← 0; {Start with zero clusters}

14. N ← rank(B);

15. Repeat

16. G ← G + 1;

17. Q ← (Pg i=1 Σii)/(Pn i=1 Σii);

18. until Q < Threshold of the Candidate Label ;

19. P ← phrase matrix for Pf;

20. For all columns of UT g P do

21. find the largest component mi in the column;

22. Add the corresponding phrase to the Cluster Label Candidates set;

23. Score_Label ← mi;

24. End for

25. Calculate cosine similarities among all combination of candidate labels;

26. Identify the combination of labels that is beyond the Threshold of Label Similarity;

27. for all groups of similar labels do

28. Identify the label with maximum score;

29. End for

STEP 4 Discovery of Cluster content

30. For all M □ Label of Cluster Candidates do

31. Create cluster C related to M;

32. Add to C all the documents d where similarity to C is larger than the Snippet Assignment Threshold;

33. End for

34. In Separate group retain remaining unassigned document d;

STEP 5 Final Cluster Formation

35. for each clusters identify

36. Score_Cluster ← Score_Label × gCg;

37. End for

The general idea behind LINGO is to find the content of the clusters by deriving the descriptions from the meaningful descriptions of clusters. LINGO could use the Latent Semantic Indexing (LSI) in the configuration, to map documents to the already labelled groups meant for it. It helps in retrieving the best matching documents, given a query. Contents of the cluster will be returned, when a cluster label query is provided to LSI. This approach exploits the LSI's ability to identify any intensified semantic dependencies in the input set.

The architecture diagram shows the relationship between different phases of the system. The program searches and identifies items in a database that correspond to keywords or characters specified by the user. This is used especially for finding particular sites on the Internet. One of the main functions of XML editor is editing the XML. Plain text editor is used to edit the XML and all the codes are visible here. The various features like tag completion, menus and buttons for task are available in text editor. This is based on data supplied with XML tree or Document Type Definition (DTD) or the XML tree. Features dealing with element tags were provided by Text XML editors. Highlighting the syntax is a basic standard of any XML editor; that is, they colour element text differently from regular text. Many text XML editors provide Element and attribute completion based on a DTD or schema. User sends the request for the data retrieval to the data server. According to the keyword the main server will response to the client.

The following are the phases of this research paper, which is planned in aid to complete part of the research work.

Data set creation

XML Parsing

Clustering xml data

Web application

Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

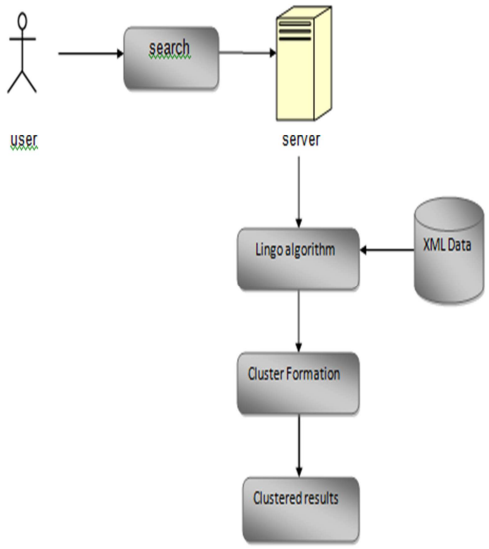

Figure 1: Outline of the Proposed Model

Data Set Creation

Data is stored in the xml.

Maximum number of data can be added in xml.

Xml Parsing

An XML parser converts an XML document into an XML DOM object - which can then be manipulated with a JavaScript.

Clustering Xml Data

Data is clustered using Lingo clustering algorithm.

Clustering process is based on the content of the document

Web Application

The user gives the query in the form of keyword to the search engine.

4. IMPLEMENTATION

Many of the text clustering algorithm follow the below steps. First they identify the content of the cluster and then they identify the label using the content. But lingo reverses this steps by first creating the cluster label and then assign the corresponding document to it. In Lingo algorithm, initially extract the frequent phrase because frequent occurring phrase may be the more relevant features.

Next, design the term document matrix based on the feature extracted to discover the topics. In the last, we assign appropriate documents to them by comparing group descriptions against extracted topics. The main requirement in a good web clustering algorithm is that the content and label of the formed group should be sensible to the user. Existing phenomena of performing cluster content discovery first, followed by determination of labels results in some groups' descriptions being meaningless to the people who use it. This is usually caused by the absurd content of the clusters. LINGO follows a fundamentally alternate method in discovering and describing groups to avoid such issues.

The main idea behind LINGO is to first identify sensible cluster descriptions. Then determine their content, based on the descriptions. LINGO can utilize Latent Semantic Indexing in the configuration, to assign documents to the earlier labelled groups Contents of the cluster would be returned, when a cluster label is provided as a query to LSI. This method would exploit LSI's capability to identify intensive semantic dependencies in the input set. Hence apart from retrieving the documents containing cluster label, extraction happens for the documents in which the same concept is described instead of the exact phrase. Semantic retrieval is rapidly declined by the minute size of the web snippets inputs, in web search results clustering.

Figure 2: XML Dataset creation

Precision of cluster content assignment is affected as a result of this. Pre-processing of the input set should happen, prior to identifying cluster labels

Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and contents. This phase should cover text filtering, stop words identification, stemming, and document's language recognition. It is also recommended that post-processing of the resulting clusters be performed to eliminate groups with identical contents and to merge the overlapping ones

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Clustering is based on the contents data on the document using lingo algorithm. The data's are stored in XML. XML manage large volume of data. Apart from discovering commonalities among documents, Lingo clubs various prevailing methods to emphasize on reasonable cluster descriptions. Thus document clustering using lingo clustering algorithm is very useful to retrieve information application in order to reduce the consuming time.

Figure 3: Clustered documents

The F-score on a class D depends on the parameter named precision and recall. Precision is defined as the number of documents correctly labelled belonging to the class D. Recall is defined as the number of documents correctly labelled as belonging to class D divided by the total number of documents that actually belong to D. Thus the F-score on a class D is defined as follows

$$\text{F-score} = \frac{2*\text{Precision}*\text{recall}}{\text{Precision}+\text{recall}} \quad (1)$$

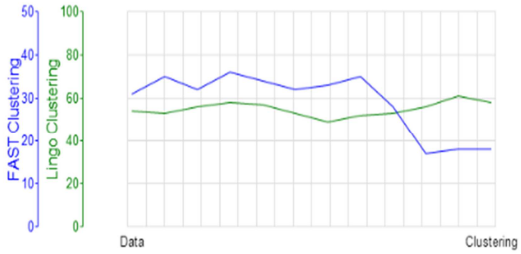

Figure 4: performance analysis

In existing algorithm, the cluster content discovery is performed first, and then, based on the content, the labels are determined. This may result in some groups descriptions being meaningless to the users, which in turn, is very often caused by the nonsensical content of the clusters themselves. To avoid such problems LINGO adopts a radically different approach to finding and describing groups. Lingo clusters the document and retrieve the clustered data faster than existing algorithm. The consuming time is very less than the existing one.

5. CONCLUSION

The cluster based searching is introduced for XML Key word search whereas xml manages large volume of data. It can elevate the adeptness of information retrieval. Thus document clustering using lingo clustering algorithm is very useful to retrieve information application in order to reduce the consuming time. LINGO fairs better for queries that wide range of subjects. For the common query all labels indicate sensible groups, in spite of larger number of clusters had been created. In case of the specific query several labels provide minimal clues for the cluster's content. In addition, for the specific query, more snippets had been mapped to the other group. This affirms that such queries are more challenging to cluster.

Lingo achieves impressive empirical results, but the task on the algorithm is yet to finish. Adding less sensible phrases would improve Cluster label pruning phase. Topic separation phase needs computationally costly algebraic transformations. It is interesting to find a way to trigger hierarchical association between topics. Finally, a more detailed evaluation methodology would be needed to expose caveats in the algorithm

REFERENCES:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[1] Almuallim.H and DietterichT.G,1992, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45.

[2] Almuallim.H and DietterichT.G,1994, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305.

[3] Arauzo-Azofra.A, Benitez.J.M, and Castro.J.L, 2004, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109.

[4] Baker.L.D and McCallum.A.K,1998, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103.

[5] Battiti.R,1994, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July .

[6] Bell.D.A and Wang.H,2000 "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195.

[7] Biesiada.J and Duch.W, 2008 "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249.

[8] Butterworth.R, Piatetsky-Shapiro.G, and Simovici. D.A, 2005, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584.

[9] Cardie.C,1993, "Using Decision Trees to Improve Case-Based Learn- ing," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32.

[10] Chanda.P, Cho.Y, Zhang.A, and Ramanathan.M, 2009, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355.

[11] Qinbao Song, Jingjie Ni, and Guangtao Wang, 2013, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data" VOL. 25, NO. 1, 1041-4347.

[12] Yu.L and Liu.U, 2003, "Efficiently Handling Feature Redundancy in High-Dimensional Data," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), pp. 685-690.

[13] Zhao.Z and Liu.H, 2007 "Searching for Interacting Features," Proc. 20th Int'l Joint Conf. Artificial Intelligence.

[14] S.Gowri , G.S.Anandha Mala, "Improving Intelligent IR Effectiveness in Forensic Analysis" in the Proceedings of the International Conference on Advances in Communication Network and Computing – CNC 2012 organized by ACEEE on Feb 24 th - 25 th, 2012 in Chennai, India. ISBN 978-3-642-35614-8