International Journal of Computer and Communication Technology

Volume 5 | Issue 4 Article 12

October 2014

Clustering Based Classification and Analysis of Data

NEERAJ SAHU
Singhania University Rajasthan, neerajsahu79@gmail.com

D. S. RAJPUT
Department of Computer Applications, MANIT, Bhopal (MP), India, Dharm_raj85@yahoo.co.in

R. S. THAKUR
Department of Computer Applications, MANIT, Bhopal (MP), India, Ramthakur2000@yahoo.com

G. S. THAKUR
Department of Computer Applications, MANIT, Bhopal (MP), India, ghanshyamthakur@gmail.com

Clustering Based Classification and Analysis of Data

[1]NEERAJ SAHU , [2]D. S. RAJPUT, [3]R. S. THAKUR, [4]G. S. THAKUR

[1] Singhania University Rajasthan
[2,3&4] Department of Computer Applications, MANIT, Bhopal (MP), India
E-mail : neerajsahu79@gmail.com[1] , Dharm_raj85@yahoo.co.in[2]
Ramthakur2000@yahoo.com[3] & ghanshyamthakur@gmail.com[4]

Abstract -This paper presents Clustering Based Document classification and analysis of data. The proposed Clustering Based classification and analysis of data approach is based on Unsupervised and Supervised Document Classification. In this paper Unsupervised Document and Supervised Document Classification are used. In this approach Document collection, Text Pre-processing, Feature Selection, Indexing, Clustering Process and Results Analysis steps are used. Twenty News group data sets [20] are used in the Experiments. For experimental results analysis evaluated using the Analytical SAS 9.0 Software is used. The Experimental Results show the proposed approach out performs.

Keywords *- Document clustering, Unsupervised learning, Supervised learning, Text Pre processing.*

I. INTRODUCTION

Document Clustering is an important issue in text mining. Clustering has been widely applicable in different areas of science, technology, social science, biology, economics, medicine and stock market. Clustering problem appears in other different field like pattern recognition, statistical data analysis, bio-informatics, etc.There exist Clustering based classification in the literature. Clustering based classification are mainly divided into two categories shown in Fig.1:

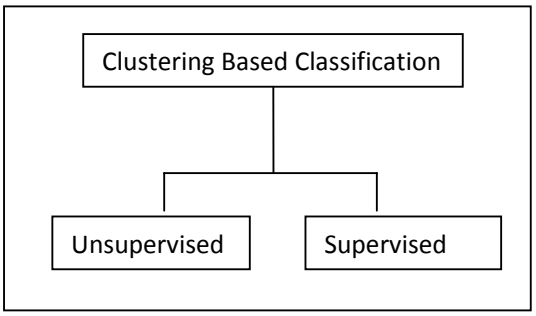

Fig. 1 : Tree structure with Clustering based classification

In recent years lot of research work has been done on Document Clustering. Some contributions are as follows:

In 2002 Beil et al.[1] worked to improve the cluster accuracy using frequent item based technique and find overlapping clusters and meaning full cluster label.

In 2010 Chun-Ling Chen,Frank S.C. Tseng,Tyne Liang[4],Fuzzy-based Multi-label Document Clustering (FMDC) algorithm concentrated on clustering accuracy and used frequent item based clustering concept and find overlapping cluster, semantic discovery and meaningful cluster label.The above mentioned work suffers from lack of efficiency and accuracy. The high complexity and low accuracy are still issues and challenges in the clustering. This motivates the study of Document Clustering.

The paper is organized as follows. Section-I described the introduction and review of literatures. In Section-II, the Unsupervised and Supervised Document Classification are described. In Section-III, Methodology of document clustering is described. In Section-IV, Experimental results are described. In Section-V, Evaluation measurement is described. Finally, we concluded and proposed some future directions in Conclusion Section.

II. UNSUPERVISED AND SUPERVISED CLASSIFICATION

Document clustering has been used to enhance information retrieval. This is based on the clustering hypothesis, which states that documents having similar contents are also relevant to the same query [1]. A fixed collection of text is clustered into groups or clusters that have similar contents. The similarity

between documents is usually measured with the associative coefficients from the vector space model, e.g., the cosine coefficient. Hierarchical clustering algorithms have been primarily been used in document clustering. The single link method has mostly been used, as it is computationally feasible, but the complete link method seems to be the most effective but is very computationally demanding [2].

Other methods than document vector similarity have been used for clustering. Neural models have been implemented for unsupervised document clustering [3].

The long computation time has always been the problem when using document clustering on-line. More recently fast algorithms for clustering have been introduced to use for browsing through collection when the user has little information about the collection and wants to brows for topics [4]. Suffix Tree Clustering is new clustering method, which creates clusters based on phrases shared between documents, works fast and intended for Web document clustering [5]. Different projections techniques, LSI and truncation, have been investigate to speed up the distance calculations of clustering [6]. An interesting application of clustering is topic clustering, i.e. clustering documents returned from a specific query, using -means clustering [7]. Effectiveness of five hierarchical clustering algorithms have been examined: single link, complete link, group average, Ward's method, and weighted average [8]. Single link is the only that badly compared to the others, but the results are very much dependent on the data set.

Supervised Document Classification:- Pattern recognition and machine learning has also been applied to document classification. As before the term frequency is used as feature. A number of classifiers have been used to classify documents. An example of these classifiers are neural networks [9,10], support vector machines [11], genetic programming [12], Kohonen type self-organizing maps [13], hierarchically organized neural network built up from a number of independent self-organizing maps [14], fuzzy -means [15], hierarchical Bayesian clustering [16], Bayesian network classifier [17], and naive Bayes classifier [18].

Some of these classifiers can be used with unsupervised learning, i.e., unlabeled documents, but the accuracy of a classifier can be enhanced by using a small set of labeled documents [18]. The aim is to use a classifier which need small amount of manually classified documents to be generalized.

The use of semi-supervised machine learning has emerged recently [18,15]. The learning scheme lies somewhere between supervised and unsupervised, where the class information are learned from the labeled data and the structure of the data from the unlabeled data.

The performance of four document classification methods have been measured: the naive Bayes classifier, the nearest classifier, decision trees and a subspace method [19]. The naive Bayes classifier and the subspace method outperform the others.

1. Supervised: In Supervised classification method, a set of predefines classes are given.
2. Unsupervised: In Unsupervised classification methods, a set of predefine classes are not given. This is also known clustering.

III. METHODOLOGY

A. Document Collection

In this phase we collect relevant documents like e-mail, news, web pages etc. from various heterogeneous sources. These text documents are stored in a variety of formats depending on the nature of the data. The datasets are downloaded from UCI KDD Archive [20]. This is an online repository of large datasets and has wide variety of data types.

B. Text Pre-processing

Text pre-processing means transform documents into a suitable representation for the clustering task. The text documents have different stop words, punctuation marks, special character and digits and other characters. After removing stop words, word stemming is performed .Word stemming is the process of suffix removal to general word stems. A stem is a natural group of words with similar meaning. In text-pre-processing we performed the following task:

a) Removal of HTML tags and special character

b) Removal stop words

c) Word stemming

C. Dimension reduction

High dimension is the greatest challenge of document clustering, so dimension reduction became major issue for clustering. This module performs two functions- indexing and feature selection. In indexing method we assign the value to the terms in the documents. After indexing, feature selection method is applied. Feature selection is the process of removing indiscriminate terms from the documents to improve the document clustering accuracy and reduce the computational complexity.

D. Word Stemming

Morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of Information Retrieval applications. For this reason we have been developed,

which attempt to reduce a word to its stem or root form. The key terms of a document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.

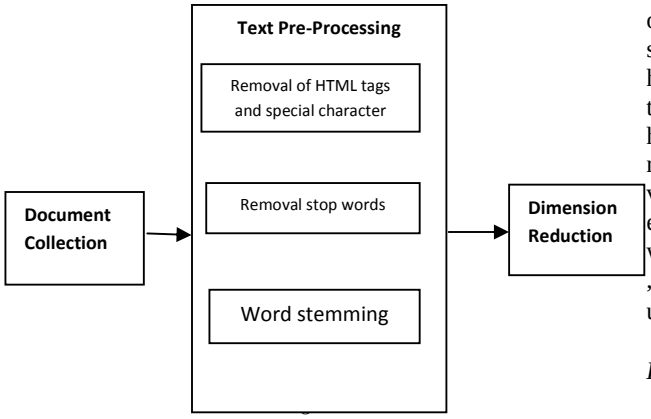

Figure 2. Steps of Methodology

*E. **Stop Words***

Stop Word list is probably the most widely used stop word list. It covers a wide number of stop words without getting too aggressive and including too many words which a user might search upon. This wordlist contains 429 words it is given below:

about,above,across,after,again,against,all,almost,alone,along,already,also,although,always,among,an,and,another,any,anybody,anyone,anything,anywhere,are,area,areas,around,as,ask,asked,asking,asks,at,away,back,backed,backing,backs,be,became,because,become,becomes,been,before,began,behind,being,beings,best,better,betweenbig,both,but,by,came,can,cannot,case,cases,certain,certainly,clear,clearly,comecould,did,differ,different,differently,do,does,done,down,down,downed,downing,downs,during,each,early,either,end,ended,ending,ends,enough,even,evenly,everevery,everybody,everyone,everything,everywhere.

face,faces,fact,facts,far,felt,few,find,finds,first,for,four,from,full,fully,further,furthered,furthering,furthers,gave,general,generally,get,gets,give,given,gives,go,going,good,goods,got,great,greater,greatest,group,grouped,grouping,groups,had,has,have,having,he,her,here,herself,high,high,high,higher,highest,him,himself,his,how,however,if,important,in,interest,interested,interesting,interests,into,is,it,its,itself,just,keep,keeps,kind,knew,know,known,knows,large,largely,last,later,latest,least,lesslet,lets,like,likely,long,longer,longest,made,make,making,man,many,may,me,membermembers,men,might,more,most,mostly,mr,mrs,much,must,my,myself,necessary,need,needed,needing,needs,never,new,new,newer,newest,next,no,nobody,non,no,one,not,nothing,now,nowhere,number,numbers,of,off,often,old,older,oldest,on,once,one,only,open,opened,opening,opens,or,order,ordered,ordering,orders,other,others,our,out,over,part,parted,parting,parts,per,perhaps,place,places,point,pointedpointing,points,possible,present,presented,presenting,presents,problem,problems,put,puts,quite,rather,really,right,right,room,rooms,said,same,saw,say,says,second,seconds,see,seem,seemed,seeming,seems,sees,several,shall,she,should,showshowed,showing,shows,side,sides,since,small,smaller,smallest,so,some,somebody,someone,something,somewhere,state,states,still,still,such,sure,take,taken,thanthat,the,their,them,then,there,therefore,these,they,thing,things,think,thinks,this,those,though,thought,thoughts,three,through,thus,to,today,together,too,tooktoward,turn,turned,turning,turns,two,under,until,up,upon,us,use,used,uses,very,want,wanted,wanting,wants,was,way,ways,we,well,wells,went,were,what,when,where,whether,which,while,who,whole,whose,why,will,with,within,without,work,worked,working,works,would,year,years,yet,you,young,younger,youngest,your,yours.

*F.**HTML Tags:***

<*!doctype*>,<*a*>,<*abbr*>,<*acronym*>,<*address*>,<*applet*>,<*area*>,<*b*>,<*base*>,<*basefont*>,<*bdo*>,<*big*>,<*blockquote*>,<*body*>,<*br*>,<*button*>,<*caption*>,<*center*>,<*cite*>,<*code*>,<*col*>,<*colgroup*>,<*dd*>,<*del*>,<*dfn*>,<*dir*>,<*div*>,<*dl*>,<*dt*>,<*em*>,<*fieldset*>,<*font*>,<*form*>,<*frame*>,<*frameset*>,<*h1*>,<*h2*>,<*h3*>,<*h4*>,<*h5*>,<*h6*>,<*head*>,<*hr*>,<*html*>,<*i*>,<*iframe*>,<*img*>,<*input*>,<*ins*>,<*isindex*>,<*kbd*>,<*label*>,<*l,egend*>,<*li*>,<*link*>,<*map*>,<*menu*>,<*meta*>,<*noframes*>,<*noscript*>,<*object*>,<*ol*>,<*optgroup*>,<*option*>,<*p*>,<*param*>,<*pre*>,<*q*>,<*s*>,<*samp*>,<*script*>,<*select*>,<*small*>,<*span*>,<*strike*>,<*strong*>,<*style*>,<*sub*>,<*sup*>,<*table*>,<*tbody*>,<*td*>,<*textarea*>,<*tfoot*>,<*th*>,<*thead*>,<*title*>,<*tr*>,<*tt*>,<*u*>,<*ul*>,<*var*><*!--*>

IV. EXPERIMENTAL RESULTS

In this paper the unstructured datasets are used. The datasets are downloaded from UCI KDD Archive [20]. This is an online repository of large datasets with wide variety of data types. This repository has twenty newsgroups dataset for text analysis. This data set consists of 20000 messages taken from Usenet newsgroup. The subset of twenty newsgroups is mini newsgroup. We have done our experiments on 20 newsgroup datasets. Each category contains 1000 documents, so there are 20000 documents for experiments. The five categories Computer Hardware,Computer Graphics,Medical,Sports and Automobile are used in first experiment.

We performed our experiments on five newsgroups- Computer graphics, Computer hardware, Automobile, Sports and Medical. In this

research the 80% dataset are used as training dataset and 20% dataset are used as test dataset.

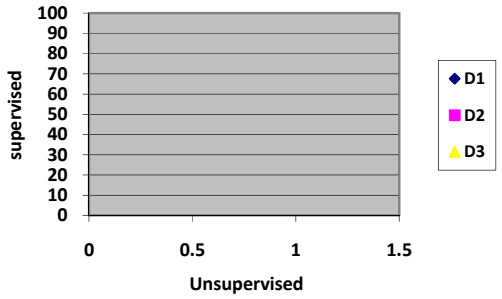

Fig.2 D1, D2 , D3 Data Set with Clustering based classification

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India, Endt.No. 2427/CST /R&D/2011 dated 22/09/2011.

REFERENCES

[1] van Rijsbergen, C. J. Information retrieval. Butterworths, 1979.

[2] Willett, Peter. Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing and Management Vol. 24, No 5, p. 577-597, 1988.

[3] MacLeod, K. An application specific neural model for document clustering. Proceedings of the Fourth Annual Parallel Processing Symposium, vol.1, p. 5-16, 1990.

[4] Douglass Cutting, David R. Karger, Jan O. Pedersen and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In Proceedings of ACM/SIGIR, p. 318-329, 1992.

[5] Zamir, O. and Etzioni, O. Web document clustering: a feasibility demonstration. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 46-54, 1998.

[6] Schutze, Hinrich and Silverstein, Craig Projections for efficient document clustering. SIGIR Forum (ACM Special Interest Group on Information Retrieval), p. 74-81, 1997.

[7] Sahami, Mehran; Yusufali, Salim and Baldonado, Michelle Q.W. Real-time full-text clustering of networked documents. Proceedings of the National Conference on Artificial Intelligence, p. 845, 1997.

[8] Burgin, R. The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. Journal of the American Society for Information Science, Vol.46 (8), p. 562-72, 1995.

[9] Li, Wei; Lee, Bob; Krausz, Franl and Sahin, Kenan. Text Classification by a Neural Network. Proceedings of the 1991 Summer Computer Simulation Conference.Twenty-Third Annual Summer Computer Simulation Conference,p. 313-318, 1991.

[10] Farkas, Jennifer. Generating Document Clusters Using Thesauri and Neural Networks. Canadian Conference on Electrical and Computer Engineering, Vol.2, p. 710-713, 1994.

[11] Joachims, Thorsten. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Machine Learning: ECML-98. 10th European Conference on Machine Learning, p. 137-42 Proceedings. 1998.

[12] Syingen, B. Using genetic programming for document classification. FLAIRS-98. Proceedings of the Eleventh International Florida Artificial Intelligence Research, p. 63-67, 1998.

[13] Hyotyniemi, H. Text document classification with self-organizing maps. STeP '96 - Genes, Nets and Symbols. Finnish Artificial Intelligence Conference, p. 64-72, 1996.

[14] Merkl, D. Text classification with self-organizing maps: Some lessons learned. Neurocomputing Vol.21 (1-3), p. 61-77, 1998.

[15] Benkhalifa, M., Bensaid, A. and Mouradi, A. Text categorization using the semi-supervised fuzzy c-means algorithm. 18th International Conference of the North American Fuzzy Information Processing Society - NAFIPS, p. 561-5, 1999.

[16] Iwayama, M. and Tokunaga, T. Hierarchical Bayesian Clustering for automatic text classification. IJCAI-95. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol.2, p. 1322-7, 1995.

[17] Lam, Wai and Low, Kon-Fan Automatic document classification based on probabilistic reasoning: Model and performance analysis. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol.3, p. 2719-2723, 1997.

[18] Nigam, Kamal; Maccallum, Andrew Kachites; Thrun, Sebastian and Mitchell, Tom. Text Classification from Labeled and Unlabeled Documents using EM. To appear in the Machine Learning Journal 1999. Draft.

[19] Li, Y.H. and Jain, A.K. Classification of text documents. Computer Journal, 41 (8) , p. 537-46, 1998.

[20] www.kdd.ics.uci.edu

◈◈◈