IJCSET |September 2012 | Vol 2, Issue 9, 1406-1409 www.ijcset.net | ISSN:2231-0711

Feature Selection and Clustering Approcahes to the KNN Text Categorization.

K.Gayathri, Dr.A.Marimuthu
Nrimala College of Arts and Science for Women
Coimbatore-18.

Abstract: Automatic text classification is a discipline at the cross roads of information retrieval machine learning and computational linguistics and consists in the realization of text classifiers. (ie) software systems capable of assigning text to one or more categories or classes, from a pre-defined set. This article will focus on the feature selection for, reducing the dimensionality of the vectors, after that we apply one pass clustering for group the related data's and then we apply classification technique like KNN for categorizations the data and finally evaluate the results by using precision, etc.,

Keywords: **Text Categorization, Pre-processing, LSI, one pass clustering, KNN.**

I. INTRODUCTION

As the volume of information is getting increased in the internet day by day there is a need for people to have the tools that finds, filter the information and manage the resource. It is highly difficult for the people to maintain the huge data manually and it is very time consuming to extrication techniques. Automatic text categorization is one particular tool to retrieve and make use of text information efficiently. Text classification is a learning task, where pre-defined categories labels are assigned to documents based on the like hood suggested by a training set of labeled documents. Text categorization methods proposed in the literature are difficult to compare [1]. The major challenges of text categorization in solving real world problem are hierarchical classification, imbalance corpus classification, classifying massive text data efficiently. Latent Semantic Indexing (LSI) techniques compresses document vectors into vectors of a lower-dimensional space whose dimensions are obtained as combinations of the original dimensions by looking at their patterns of co-occurrence [2].Although it was originally applied in the context of information retrieval, since then it has been successfully applied to a wide variety of text-based tasks. Clustering-a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data. The clustering or the cluster analysis is a set of Methodologies for classification of samples into a number of groups. Therefore, the samples in one group are grouped and samples belonging to different groups are grouped as another group. The input of clustering is a set of samples and the process of clustering is to measure the similarity and or dissimilarity between given samples. The output of the clustering is a number of groups or clusters in the form of graphs, histograms and normal computer results showing group number[3]. Among the numerous clustering algorithms proposed, single-pass clustering stands out in terms of both time and space efficiency. However, it is generally acknowledged that single-pass clustering has a major defect, namely its output depends on the order in which documents are presented. The single pass method has the advantage of simplicity, it is often criticized for its tendency to produce large clusters early in the clustering pass, and because the clusters formed are not independent of the order in which the data set is processed. It is sometimes used to form the groups that are used to initiate reallocation clustering. This paper presents a simple KNN algorithm adapted to text categorization that does aggressive feature selection. This feature selection method allows the removal of feature that add no new information given that some other feature highly interacts with them, which weak prediction capability. Redundancy and irrelevancy could harm a KNN learning by giving it some unwanted bias, and by additional complexity. By taking into account both the redundancy and relevancy of feature, we aim at providing solid ground for the use of KNN algorithms in text categorization where the document set is very large and the vocabulary diverse[4].Data sets used in the experiments are rarely same different studies usually use different portions of the test sets differently. More over classification will be performed using SVM. For the analysis and comparison of different results precision recall and F-measure are used.

II. FEATURE SELECTION

FS (Feature Selection) is an effective approach to reduce the size of feature space. Many researchers have proposed a lot such as DF (Document Frequency), IG(Information Gain), MI(Mutual Information) and combination of multiple methods. In these methods, features are selected according to the feature function value. The features with to the higher function value will be selected firstly. In the selection process, the size of feature set plays an important role in text categorization. Experimental results show that the performance of text categorization does not always increases with the growth of features. When the features are too much the performance of those feature selection methods may be lower. Thus, a reasonable measure of feature set can not only reduce the great number of processing overhead but also improve the effectiveness of classifiers [5].

2.1 Latent semantic indexing:

Latent semantic Indexing technique compresses document vectors into vectors of a lower-dimensional space whose dimensions are obtained as combinations of the original

dimensions by looking at their patterns of co-occurrence. Although it was originally applied in the context of information retrieval, since then it has been successfully applied to a wide variety of text-based tasks. These studies of LSI have mostly used standard text collections in information retrieval, some of them having simpler document models. The purpose of this paper is to investigate the use of LSI on some real-life text documents, namely patent documents. Our motivation came from the continuous growth of patent documents database in the recent years, an increase which requires the development of new and efficient methods for classification and retrieval of patent documents.

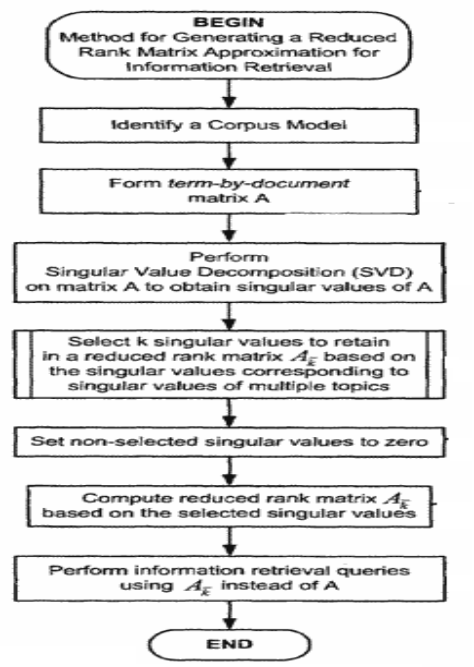

Fig .1

Latent Semantic Indexing is a statistical technique which tries to surpass some limitations imposed by the traditional Vector Space Model (VSM). In VSM, which uses the so-called bag-of-words representation of documents, the collection of text documents is represented by a terms-documents matrix

$$A = [aij]\varepsilon R^{txd}$$

where each entry aij corresponds to the number of times the term i appears in document j. Here t is the number of terms and d the number of documents in the collection. Therefore a document becomes a column vector and a query of a user can be represented as a vector of the same dimension. The similarity between the user's query vector and a document vector in the collection is measured as the cosine of the angle between the two vectors. LSI is a method for dimensionality reduction because it transforms the original terms-documents vector space into a new co-ordinate system of conceptual topics, a lower dimensional space that captures the implicit higher-order structure in the association of terms with documents . Both sets of documents and terms will be projected onto this new low-dimensional space spanned by the true factors or concepts, instead of representing documents as vectors of independent words. In order to obtain the space of concepts, i.e. the space of true representation of words and documents, LSI uses a truncated Singular Value Decomposition (SVD) applied to the terms-documents matrix A described above.

Given a t×d matrix A, where m = min(t; d), the singular value decomposition of A is defined as

$A = USV^{T,}$

where U is a t_m orthonormal matrix ($U^TU = Im$), whose columns define the left singular vectors, V is a d × m orthonormal matrix ($V^TV = I_m$), whose columns define the right singular vectors and S is a m× m diagonal matrix containing the singular values of A decreasingly ordered along its diagonal: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_m = 0$, where r = rank(A). This decomposition is unique up to making the same permutations of columns of U, elements of S and columns of V(rows ofV^T).

To reduce the noise and redundancy, LSI, taking as input the terms-documents matrix described above, uses a truncation of SVD which consists in retaining only the largest k singular values and deleting the remaining ones which are smaller and thus considered unimportant. We remove also from U and V the columns corresponding to the small singular values and we get

$$A_k = U_kS_kV_k^T$$

where S_k is a k × k diagonal matrix containing the largest k singular values as entries, U_k is a t×k matrix of the corresponding left singular vectors as columns and V_k is a d × k matrix whose columns are the corresponding right singular vectors[6].

III. CLUSTERING METHODS

Clustering is a process of partitioning data into clusters of similar objects. It is an unsupervised learning process of hidden data. In text clustering, it assumes the similarity degree of the content of the documents in the same cluster is the most, while in different clusters to the least. Therefore, to preprocess the documents using clustering is useful for discovering the distribution and structure of corpus. The state-of-the-art clustering approaches were reported in the thorough survey. A majority of clustering algorithms proposed in previous literatures cannot handle large and high-dimensional data. However, incremental clustering algorithms with less time consuming can deal with it, since they are non-iterative and scan corpus in single pass. One pass clustering algorithm is a kind of incremental clustering algorithm with approximately linear time complexity. To build the classification model with the training text documents, we use one pass clustering algorithm to constrainedly cluster the text collections.

A. Single Pass Algorithm

The single pass algorithm is a data clustering algorithm based on the comparison of the maximum similarity between a particular data item and existing clusters with the critical similarity, parameter[7].

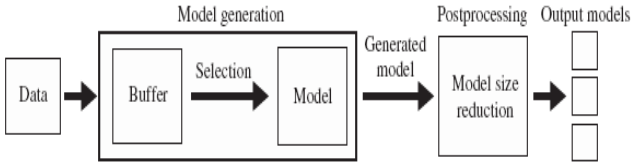

Fig.2

The single pass model generation is generated from the data by the single-pass algorithm. The model can then be post-processed in order to obtain a simpler representation, if desired. The post-processing can be performed without the original data. The result is a model that represents the same data but with fewer components. The benefit of using the first stages model takes less time and memory to process than the entire data set. The step illustrated in fig.2

The first stage of the algorithm tends to generate a model that has an excessive number of components. Although this does not affect the representation accuracy of the model, a more compact model with a smaller number of components could be desired for practical purposes. however it is not a simple task to restrict the size of the model during the first stager without affecting the quality of the model[7].

The process of clustering documents with single pass algorithm:

Input: a list of documents and the critical similarity

Step 1: Encode a list of documents into bags of words, numerical vectors, or string vectors

Step 2: Get the first document from the list of documents

Step 3: Create a cluster and include the document into it, as its representative document

Step 4: Repeat the following sub steps, 4-1, 4-2, and 4-3 for each of successive documents

Step 4-1: Compute the similarity between the successive document and the representative document of each cluster

Step 4-2: Obtain the maximum similarity and its corresponding cluster among existing clusters

Step 4-3: If the maximum similarity >= the critical similarity, include the successive representation otherwise, go to step 3

Output: a list of clusters including documents

Step 4 illustrates the process of clustering documents using the single pass algorithm.

A list of documents and the critical similarity are given as the input to this process. These documents are encoded into bags of words, or numerical vectors. A cluster is created and the first document is included in it, as its representative document which is used to compute the similarity between its cluster and another document. For each successive document, Sub-steps, 4-1, 4-2, and 4-3, in figure 4, are repeated. The representative document of each cluster indicates the document which is included initially, when the cluster is created. The maximum of these similarities and its corresponding cluster are determined. If the largest similarity is higher than or equal to the critical similarity, the successive document is included in its corresponding cluster. Otherwise, a new cluster is created and the successive document is included in the new cluster as its representative document. The process illustrated in step 4, thus, generates a list of clusters including documents, as its output of clusters is far less than the number of documents.

3.2.K-Nearest Neighbor Classification Approach:

There are many approaches to assign category to incoming text. In our paper, we implemented text-to-text comparison (TTC), which is also known as the k-nearest neighbor (KNN) KNN is a statistical classification approach, which has been intensively studied in pattern recognition over four decades. KNN has been successfully applied to TC problem, The KNN algorithm is quite simple: Given training and test documents, the algorithm finds the k-nearest neighbors among the training documents, and uses the categories of the k-neighbors to weight the category of the test document. The similarity scores of each neighbor document to the test document are used as a weight of the categories of the neighbor n document. If several of the k-nearest-neighbors share a category, then the pre-neighbor weights of that category are added together, and the resulting weighted sum is used as the likelihood score of that category with respect to the test document. By sorting the scores of the candidates' categories, a ranked list is obtained for the test document [8].

IV. REUTERS 21578

The Reuters-21578 dataset and used the standard "modApté" train/test split. These documents appeared on the Reuters newswire in 1987 and were manually classified by personnel from Reuters Ltd. ModApt'e spli-9603 training and 3299 testing documents. Totoal-12902 documents. Out of 135 categories of documents only the top five were selected. Distinct words- 31715; Average number of words per document-126 words of which 70 were distinct.

Table-1

Category	#Train Docs	# Test Docs
Acq	1651	719
Grain	434	149
Ship	198	89
Crude	389	189
Wheat	212	71
Earn	2877	1087
Money-fx	538	179
Trade	369	117
Interest	348	131
Corn	181	56

Table-2

1	Acq	Mergers/Acquisitions
2	Grain	Grain(Commodity)
3	Ship	Shipping
4	Crude	Crude oil
5	Wheat	Wheat (Commodity)
6	Earn	Earning and earning for cast
7	Money-fx	Money/Foreign Exchange
8	Trade	Trade
9	Interest	Interest Rates
10	Corn	Com (Commodity)

V. PERFORMANCE METRIC

The evaluation of a classifier is done using the precision and recall measures .To derive a robust measure of the effectiveness of the classifier It is able to calculate the breakeven point, the 11-point precision and ”average precision” . to evaluate the classification for a threshold ranging from 0 (recall = 1) up to a value where the precision value equals 1 and the recall value equals 0, incrementing the threshold with a given threshold step size. The breakeven point is the point where recall meets precision and the eleven point precision is the averaged value for the precision at the points where recall equals the eleven values 0.0, 0.1, 0.2... 0.9, 1.0. ”Average precision” refines the eleven point precision, as it approximates the area”below” the precision/recall curve.

Resultant Table:

Without Feature Selection, without clustering		
Accuracy	Error Rate	Speed
90.04	9.96	40.43
93.94	6.06	29.54
95.91	4.09	20.19
With Feature Selection, with clustering		
93.94	6.06	29.54
95.91	4.09	20.19
98.05	9.99	42.45

CONCLUSION

Analyzed the text classification using the KNN with the Feature selection and one pass clustering techniques. . The advantage of the proposed approach is, the classification algorithm learns importance of attributes and utilizes them in the similarity measure. In future the classification model can be build, which analyzes terms on the concept sentence in document.

REFERENCES

1. Shengyijiang, Guansong Pang, MeilingWu, Limin Kuang; “An improved K-nearest neighbor algorithm for text categorization”.Expert systems with application 39(2012)1503-1509.WWW.elsevier.com/locate/eswa.
2. Fabrizio sebastion “Machine learning in automated text categorization” ACM computing surveys Vol.34, No.1 March-2002 p.p1-47.
3. Koteeswaran.S, P.Visu and J.Janet.”A review on clustering and outlier analysis Techniques in Data Mining. American Journal of Applied Sciences 9 (2):254-258-2012.ISSN:1546-9239©2012 Science Publications.
4. Pascal Soucy, Guy W.Mineau “A simple KNN Algorithm for text Categorization” Universite Laval,Quebec, Canada, 0-7695-1119-8/01©2001IEEE.
5. JinDai,Zhongshi He, and Feng Hu.”A High Performance Algorithm for Text Feature Automatic Selection” Proceedings of International Symposium on Information Processing, August 21-23, 2009.
6. Andress Moldovan, Radu Ioan Bot, Gert wanka,.”Latent Semantic Indexing for Patent Documents”.Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz,Germany.
7. Taeho jo “The Implementation of Dynamic Document Organization using the Integration of Text Clustering and Text Categorization” Ottawa-Carleton Institute for Computer Science, School of information Technology and Engineering(SITE).University of Ottawa,Ottawa,Canada.
8. Wael Musa Hadi, Fadi Thabtah, Hussein Abdel-jaber “A Comparative Study using VSM with KNN on text CategorizationData” Proceedings of the world congress on Engineering 2007 Vol-I WCE2007, July2-4-2007,London, U.K.