International Journal of Information Processing, 7(3), 83-91, 2013
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Text Classification and Distributional Features Techniques in Datamining and Warehousing

Srikanth Bethu[a], G Charless Babu[b], J Vinoda[c], E Priyadarshini[c], M Raghavendra Rao[c]

[a]Assistant Professor, Department of Computer Science and Engineering, Holymary Institute of Technology and Science, JNTU Hyderabad - 501 301 India, Contact: srikanthbethu@gmail.com

[b]Professor, Department of Computer Science and Engineering, Holymary Institute of Technology and Science

[c]Department of Computer Science and Engineering, Holymary Institute of Technology and Science, JNTU Hyderabad

Text Categorization is traditionally done by using the Term Frequency and Inverse Document Frequency. This type of method is not very good because, some words which are not so important may appear in the document. The term frequency of unimportant words may increase and document may be classified in the wrong category. For reducing the error of classifying of documents in wrong category. The Distributional features are introduced. In the Distribuional Features, the Distribution of the words in the whole document is analyzed. Whole Document is very closely analyzed for different measures like FirstAppearence, Last Appearance, Centriod, Count, *etc.*. The measures are calculated and they are used in tf*idf equation and result is used in k- nearest neighbor and K-means algorithm for classifying the documents. .

Keywords : K-means Algorithm, K-nearest Neighbour.

1. INTRODUCTION

Text classification is the task of automatically classifying set of documents into categories from a predefined set. This task has several applications related to selective distribution of information to information consumers, spam filtering and identification of document type. Automated text classification is good because it frees organizations from the need of manually organizing document. Text classification has gained importance because of the increased amount of documents over the years. Text documents should be categorized according to its contents. The aim of the paper is to classify the document in the correct category, *i.e.*, the document should be classified into correct type of document to which it belongs, for example the computer science type of document should be classified as computer science type of document.

1.1. Problem Definition

The classification of the text is traditionally done by the term frequency and the inverse document frequency, this method has lot of problems. Term frequency cannot classify the documents in the correct category because the unimportant words may appear more number of times and the document may be classified to wrong category. The below example will explain the problem.

1.2. Example:Bill Gates philanthropy costs him richest-man title

(i) Bill Gates attends a session at the World Economic Forum (WEF) in Davos January 28, 2011.

(ii) Bill Gates didn't lose his title as the world's richest man last year; he gave it away by plowing billions into his charitable foundation, experts say.

(iii) Forbes will release its 2011 billionaires

The algorithm is composed of the following steps:

(i) Randomly choose k data points to be the initial centroids, cluster centers.

(ii) Assign each data point to the closest centroid.

(iii) Re-compute the centroids using the current cluster memberships.

(iv) If a convergence criterion is not met, go to 2).

Examples for k means algorithm are shown in Figures 2 and 3. The k-means clustering technique has been implemented which is like with hierarchical initial set (HKM). The goal is to prove that clustering document sets do enhancement precision on information retrieval systems, since it was proved by Bellot and El-Beze on French language.

3. EXPERIMENTAL RESULTS

The corpus consists of 30 documents, which need to be classified. The 15 documents are of Institute for Electrical and Electronic Engineers, related to Computer Science Engineering. The remaining 15 documents are of conference papers of various topics like Medical, Civil Engineering, *etc.*.

The thirty documents are taken as test documents and the performance is calculated using both traditional method using term frequency and distributional feature that is using compactness, using the precision and recall measures. The precision of the traditional method and compactness give the same result, but the recall measure of distributional features gives the better result than traditional method of classifying the documents.

Distributional features give more accurate classification than the traditional term frequency method of classification of documents. Therefore distributional features are useful for classification of the documents.

Figure 4 shows the selection of document from the folder, the document which is required for analysis can be uploaded for system. Figure 5 shows the show stats screen this screen will show the complete results of the analysis done. This screen shows the first appearance, last appearance, compactness and the number paragraphs in the document are displayed.

Figure 6 shows the results of single document compactness histogram. The important word from the selected words is displayed. Figure 7 shows the results of multiple document classification. The document which is more relevant to the selected word is displayed.

4. CONCLUSIONS

The Classification of the documents should be done by using the distributional features. The measures should be used and all the features of the documents are to be extracted. The extracted features should be observed and the right decision should be taken in classification of document.

The right combination of the features should be taken to gain the full advantages of the Distributional features. The Features along with good classification algorithm should be used for classification of text.

REFERENCES

1. Xiao-Bing Xue and Zhi-Hua Zhou. Distributional Features for Text Categorization, *IEEE Transactons on Knowledge And Data Engineering*, 21(3):428-442, 2009.
2. Hyunsoo, Haesun Park and Kim Peg Howland. Dimension Reduction in Text Classification with Support Vector Machines, " *Journal of Machine Learning Research*, 6:1-17, 2005.
3. Li Youwen, Xia Shixiong and Zhou Yong. An Improved KNN Text Classification Algorithm Based on Clustering, " *Journal of Computers*, 4(3):230-237, 2009.
4. Chengqing Zong, Chu-Ren Huang, Shoushan Li and Rui Xia. A Framework of Feature Selection Methods for Text Categorization, " *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 692700, 2009.

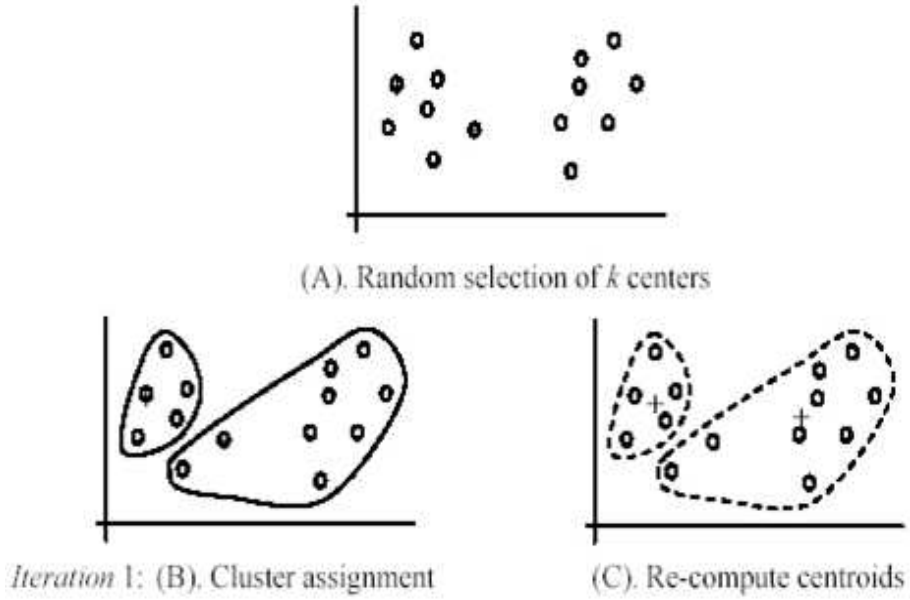

Figure 2. K means Algorithm Example

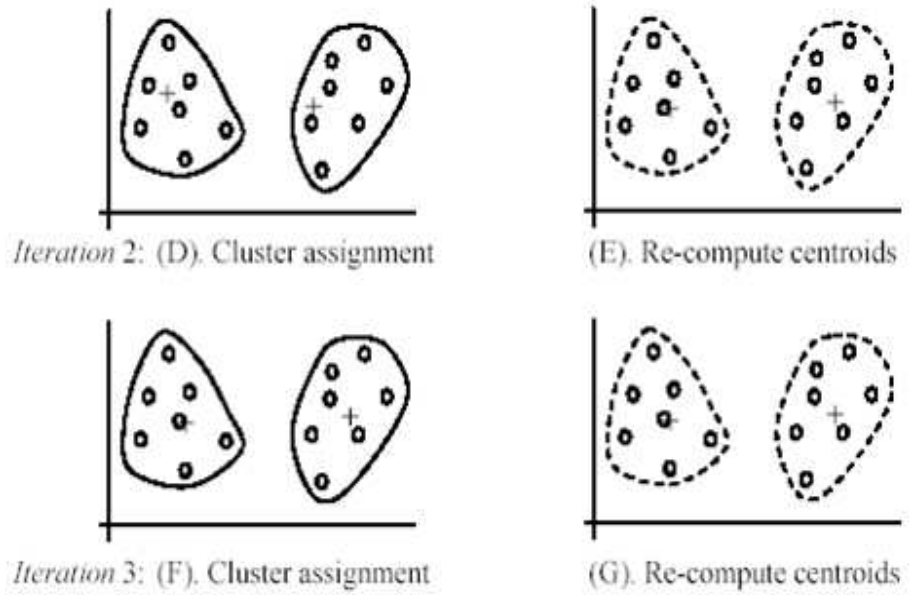

Figure 3. K means Algorithm Example

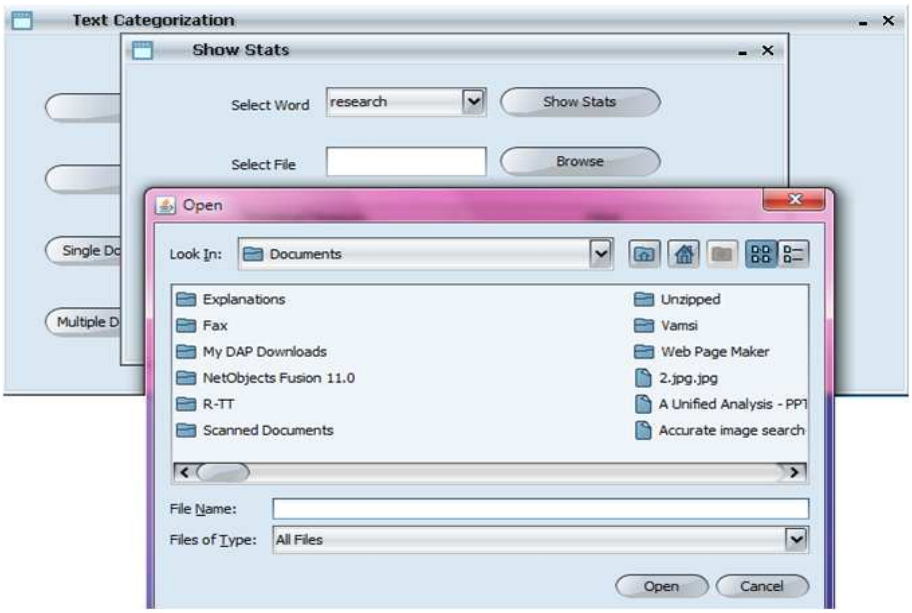

Figure 4. Selection of Document from the Folder

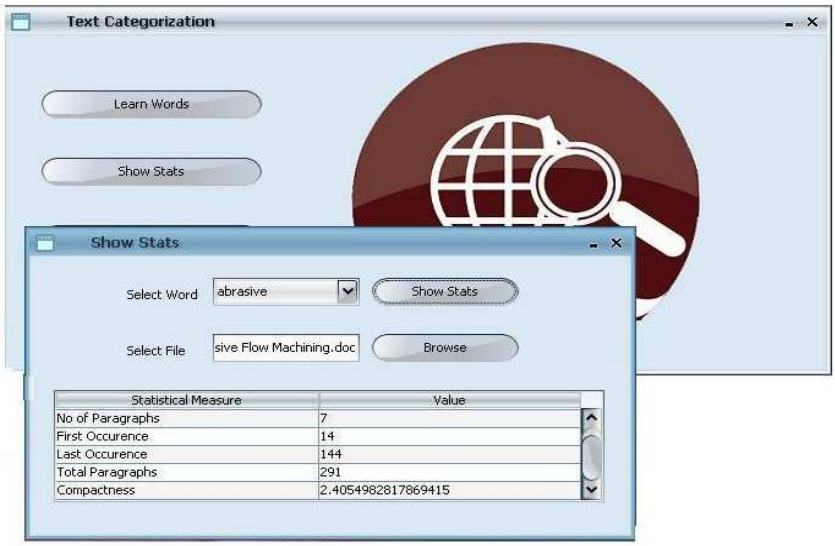

Figure 5. Show Stats

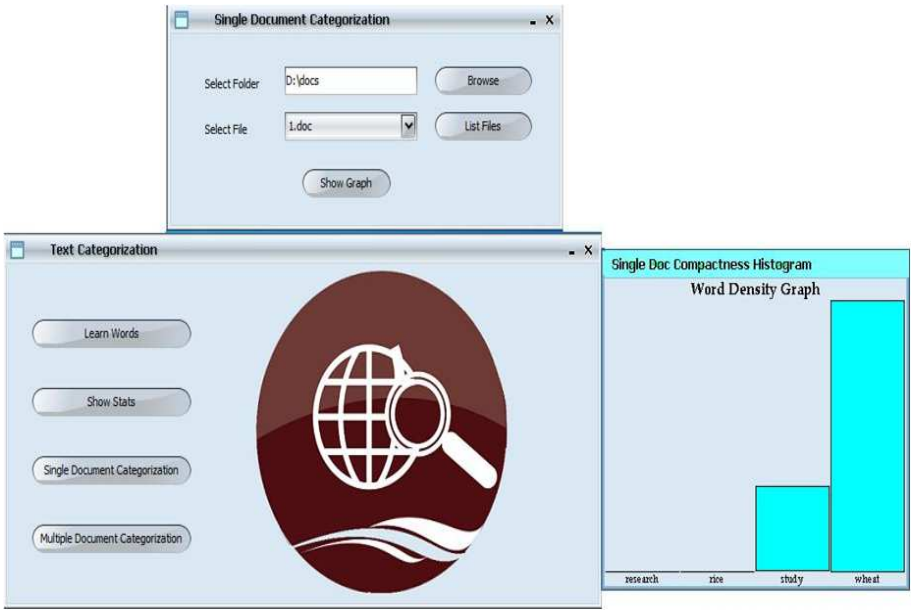

Figure 6. Results of Single Document Compactness Histogram

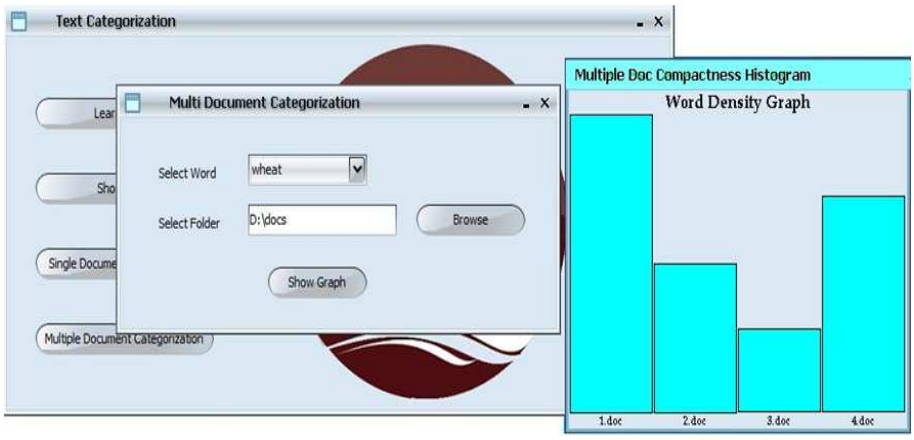

Figure 7. Results of Multiple Document Compactness Histogram

5. F Li and Y Yang. A Los Function Analysis for Classification Methods in Text Categorization, " *in Proceedings of 20th International Conf. Machine Learning (ICML 03)*, pages 472- 479, 2003.
6. Li Baoli, Lu Qin and Yu Shiwen. An Improved k-Nearest Neighbor Algorithm for Text Categorization, "*in Proceedings of 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China*, pages 1-7, 2003.
7. Li Youwen, Xia Shixiong and Zhou Yong. An Improved KNN Text Classification Algorithm Based on Clustering, *Journal of Computers*, 4(30):230-237, 2009.
8. Manu Konchady. Text Mining Application Programming, *Cengage Learning* , 1st edition, pages 209–233, 2008.
9. N Tishby, R Bekkerman, R El-Yaniv and Y Winter. Distributional Word Clusters versus Words for Text Categorization, *Journal on Machine Learning Research*, 3:1182-1208, 2003.
10. X B Xue and Z H Zhou. Distributional Features for Text Categorization, *in Proceedings 17th European Conference on Machine Learning (ICML 06)*, pages 497-508, 2006.
11. N Tishby, R Bekkerman, R El-Yaniv and Y Winter. Distributional Word Clusters versus Words for Text Categorization, *Journal on Machine Learning Research*, 3:1182-1208, 2003.
12. Li Youwen, Xia Shixiong and Zhou Yong. An Improved KNN Text Classification Algorithm Based on Clustering, *Journal of Computers*, 4(3):230-237, 2009.

Srikanth Bethu is currently the Assistant Professor, Holy Mary Institute of Technology and Science, JNTU Hyderabad, Hyderabad. He obtained his Bachelor of Engineering from JNTU Hyderabad. He received his Masters degree in Computer Science and Engineering from Osmania University, Hyderabad.

G Charless Babu is a Professor, Holy Mary Institute of Technology and Science, JNTU Hyderabad, Hyderabad. He was a Professor since 2010 with the Department of Computer Science and Engineering, HITS college, JNTU Hyderabad. During the past 10 years of his service at various institutions he has over 30 research publications in refereed International Journals and Conference Proceedings.

J Vinoda is currently the Assistant Professor, Holy Mary Institute of Technology and Science, JNTU Hyderabad, Hyderabad.

E Priyadarshini is currently the Assistant Professor, Holy Mary Institute of Technology and Science, JNTU Hyderabad, Hyderabad.

M Raghavendra rao is currently the Assistant Professor, Holy Mary Institute of Technology and Science, JNTU Hyderabad, Hyderabad.