Brill Academic Publishers
P.O. Box 9000, 2300 PA Leiden,
The Netherlands

Lecture Series on Computer and Computational Sciences
Volume 1, 2006, pp. 1-4

Density Based Text Clustering

E.K. Ikonomakis[a], **D.K. Tasoulis**[a,b], **M.N. Vrahatis**[a,1]

[a] Computational Intelligence Laboratory (CI Lab), Department of Mathematics, University of Patras Artificial Intelligence Research Center (UPAIRC), University of Patras, GR–26110 Patras, Greece.
e-mail: oikonem@master.math.upatras.gr, vrahatis@math.upatras.gr
[b] Institute for Mathematical Sciences, Imperial College London, South Kensington Campus, London SW7 2PG, United Kingdom
e-mail: d.tasoulis@imperial.ac.uk

Received 21 July, 2006; accepted in revised form 31 July, 2006

Abstract: As the discovery of information from text corpora becomes more and more important there is a necessity to develop clustering algorithms designed for such a task. One of the most, successful approach to clustering is the density based methods. However due to the very high dimensionality of the data, these algorithms are not directly applicable. In this paper we demonstrate the need to suitably exploit the already developed feature reduction techniques, in order to maximize the clustering performance of density based methods.

Keywords: Clustering, Text Mining, k-windows

Mathematics Subject Classification: 62H30, 68T30, 68T50, 68T10.

1 Introduction

The ever growing size and diversity of digital libraries, has rendered Text Mining in general important research area. Today, especially *Document Clustering* has become a necessity due to the daily increasing amount of digital available documents. Several methods and variations have been proposed over time that try to tackle this problem. However, the usual high associated computational cost, and the extreme dimensionality of the data retains the development of new methods and techniques an active research area.

Density based clustering methods constitute an important category of clustering algorithms [2, 4, 11], especially for data of low attribute dimensionality [3, 7]. In these methods, clusters are formed as regions of high density, in dataset objects, surrounded by regions of low density; proximity and density metrics need to be suitably defined to fully describe algorithms based on these techniques. However, in the high dimensionality of text data the effectiveness of most distance functions is severely compromised [1, 5].

One modern density based clustering method is the "Unsupervised k-Windows" (UkW) algorithm [15], that exploits hyperrectangles to detect clusters in the data space. Additionally using techniques from computational geometry allows a reduction in the number of objects examined at each step. Furthermore, the algorithm can detect clusters of arbitrary shapes and determine the cluster number without additional computational burden. Although it has already been applied

[1] Corresponding author: e-mail: vrahatis@math.upatras.gr, Phone: +30 2610 997374, Fax: +30 2610 992965

in numerous tasks that range from medical diagnosis [9, 14], to web personalization [12], it has not been used yet on text data. In this paper, through indicative results of the UkW algorithm, we demonstrate that the feature scoring metrics have to be suitably exploited, so that the cluster detection ability of UkW is optimized. Furthermore, we compare these results to the corresponding ones obtained by other density based methods. As a benchmark text dataset, we employ the well known and widely used RCV1 corpus [8].

As a text corpus is composed of word sequences, the *vector space model*, has been developed [16] as an algebraic model that allows the direct application of information retrieval algorithms on documents. This model, represents natural language documents in a formal manner, by the usage of vectors in a multi-dimensional space. The dimensionality d of the space is equal to the total number of words in the corpus. Each coordinate of this space is associated to a specific word in the set of all the words (vocabulary). In this way, a specific document x is represented as a vector of numbers $x = \{x_1, \ldots, x_d\}$, where each component x_i, $i = 1, \ldots, d$ of x, designates the number of occurrences of the ith word, in document x. The normalization of the document vector is also a common practice.

Performing feature reduction for an unsupervised procedure such as clustering, includes, feature scoring functions that do not need to take into account the class labels of a document. These functions, although very simple, they are reported to perform well [17]. Such functions are Document Frequency (DF) and Inverse Document Frequency (IDF) [6]. Furthermore, we examine and compare how common functions, that incorporate information of user determined document categories, can reduce the feature set size in such a way that cluster detection is optimized. Such functions are chi-square and OddsRatio [10].

The most challenging of problems associated with text clustering is the inability to efficiently measure distance due to the high dimensionality of the document vectors [1, 5]. In the case of density based clustering methods, this inability is additionally magnified, since these methods are based on suitably defined distance metrics. Therefore, a dimension reduction of the document vector space should be used, to select the features which may result in highly performing systems with very low computational cost. Even with the use of simple feature selection metrics such as *Document Frequency* the major part of the vocabulary can be discarded with no loss in accuracy. Document and *Inverted Document Frequency* are motivated by the very simple idea that the most common words are not of much use in a document clustering task, since their appearance in many documents has no distinctive information. To this end the most often appearing are usually considered as stopwords and therefore discarded. Furthermore, very rare words can not contribute much information to a clustering algorithm since their appearance could be random. Thus, discarding them results in negligible loss of information.

The Document Frequency feature scoring measure, operates by retaining terms that occur in the largest number of documents. It can be easily computed as the sum of all the documents each term belongs to. This function seems to be trivial, nonetheless the features selected using DF allow classifiers to perform excellent while the computational cost can severely be reduced [17].

On the other hand, the motivation behind Inverted Document Frequency measure, is that commonly appearing features are not useful for discriminating relevant documents. This function defined by the logarithm of the size of the dataset divided by the number of documents containing the word.

2 A test experiment on The Reuters Corpus

The RCV1-v2 (Reuters Corpus Volume 1 - version 2) is a corrected version of RCV1-v1 as created by Lewis *et al.* [8]. The Reuters corpus is actually a collection of stories of the Reuters news agency. It consists of stories produced during 20 August 1996 and 19 August 1997. Although the

complete set of documents consists of over 800,000 documents, the RCV1-v2 is limited to 804,414 documents (more than 2,000 documents less than RCV1-v1), due to the removal of documents with no topic or region codes. RCV1-v2 is a multilabel text corpus. This means that each document may belong to more than just one category. The version of this corpus has already been tokenized and stemmed. Stopwords have also been removed [8].

In our experimental setting we used 1201 documents, from that collection. The documents in that part were assigned to 89 different categories. The total number of different words was 12955. Next we reduced the dimensionality of the data to 100, using a simple combination of Document Frequency (DF) and Inverted Document Frequency (IDF). In detail, we excluded the 50 words with the lowest IDF, and from the remaining ones we used the 100 with the highest DF value. Using this simple, completely heuristic approach without any justification for the exact number values, we examined the clustering ability of the UkW algorithm, as an indicative result. The UkW algorithm detected 20 clusters. One of those clusters contained 968 documents, the majority of the data. This fact exactly, demonstrates that although the data representation has a low dimensionality, the feature selection also results in loss of distinctive information. However, there are cases of clusters that clearly contain 20-50 documents from at most 2 or 3 categories. This designates that some of the descriptive power is still included, despite the brutality of the technique.

To evaluate the result of UkW algorithm, experiments were performed with the DBSCAN algorithm [13]. This algorithm is one of the most popular density based techniques. For various combinations of values for the Eps, and $MinPts$ parameters of this algorithm (see [13]), the algorithm resulted in a similar result. In detail, the algorithm recognized a series of clusters with 20-50 documents from 2 or 3 categories. However a very sparse cluster was detected that contained the majority of the documents.

3 Concluding Remarks

One of the most successful categories of clustering algorithms, Density Based methods, is hindered by the inability to calculate distances in the very high dimensionality of the involved data. Although, many feature selection schemes have been proposed so far, an extensive examination of their impact on the clustering quality of Density Based approaches has not yet been performed. In this contribution we demonstrate exactly this fact through indicative experiments. We intend to further investigate how such measure can be suitably exploited to optimize cluster detection.

References

[1] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proc. 8th Int'l Conf. Database Theory (ICDT)*, pages 420–434, London, 2001.

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, 1999.

[3] P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 25–72. Springer, Berlin, 2006.

[4] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int'l. Conf. on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[5] A. Hinneburg, C. Aggarwal, and D. Keim. What is the nearest neighbor in high dimensional spaces? In *The VLDB Journal*, pages 506–515, 2000.

[6] M. Ikonomakis, S. Kotsiantis, and V. Tampakas. Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8):966–974, 2005.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[8] D.D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[9] G.D. Magoulas, V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis. Tumor detection in colonoscopy using the unsupervised k-windows clustering algorithm and neural networks. In *Fourth European Symposium on "Biomedical Engineering"*, 2004.

[10] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In *16th International Conference on Machine Learning*, pages 258–267, 1999.

[11] C.M. Procopiuc, M. Jones, P.K. Agarwal, and T.M. Murali. A Monte Carlo algorithm for fast projective clustering. In *Proc. 2002 ACM SIGMOD*, pages 418–427, New York, NY, USA, 2002. ACM Press.

[12] M. Rigou, S. Sirmakessis, and A. Tsakalidis. A computational geometry approach to web personalization. In *IEEE International Conference on E-Commerce Technology (CEC'04)*, pages 377–380, San Diego, California, July 2004.

[13] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.

[14] D.K. Tasoulis, L. Vladutu, V.P. Plagianakos, A. Bezerianos, and M.N. Vrahatis. On-line neural network training for automatic ischemia episode detection. In Leszek Rutkowski, Jörg H. Siekmann, Ryszard Tadeusiewicz, and Lotfi A. Zadeh, editors, *Lecture Notes in Computer Science*, volume 2070, pages 1062–1068. Springer-Verlag, 2003.

[15] M. N. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides. The new k-windows algorithm for improving the k-means clustering algorithm. *Journal of Complexity*, 18:375–391, 2002.

[16] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector spaces model in information retrieval. In *SIGIR '85: Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–25, New York, NY, USA, 1985. ACM Press.

[17] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *14th International Conference on Machine Learning*, pages 412–420, 1997.