Using Data Fusion for a Context Aware Document Clustering

P. Venkateshkumar
Master of Computer Applications
Anna University
Chennai, India

A. Subramani, Ph.D.
Department of Computer Applications
Professor & Head, KSR College of Engineering
Namakkal, India

ABSTRACT
The large volume of unstructured text data available at various sources such as digital libraries, news, internet, has given arise a need to organize the information as per the user's requirement. Search for relevant information is efficient when context of the selected word in the document is considered. Document Clustering aims to discover natural groupings, and present an overview of classes (topics) in a document collection. Thus, documents with similar contents are related to the same query. In this paper, a new method for clustering documents is proposed. In the proposed method, the term frequency of the document collection is computed and contexts based terms are fused. Agglomerative clustering and Bisecting K-Means are used to cluster the extracted features.

General Terms
Clustering, Algorithms

Keywords
Document clustering, term frequency, Bisecting K-means, Agglomerative clustering, Reuters dataset.

1. INTRODUCTION
Document clustering helps organize large volume of unstructured text data for easier organizing of information. When context of the selected word in the document is considered, information search for relevant information is efficient. Document clustering, a subset of data clustering, borrows concepts from among others information retrieval (IR), natural language processing (NLP), and machine learning (ML). Document clustering will hereafter be referred to as clustering. Clustering aims to discover natural groupings, and present an overview of classes (topics) in a document collection. Thus, documents with similar contents are related to the same query [1]. In artificial intelligence, this is called unsupervised machine learning. Clustering is not classification. In classification, the numbers of classes are known a priori, and documents are assigned to them. Conversely, in clustering, number, properties, nor membership of classes is known in advance. A good clustering is one that organizes a collection into groups so that documents in each group are similar and dissimilar to those in other groups [2, 3]. Clustering can produce disjoint/overlapping partitions. In an overlapping partition, a document appears in multiple clusters. As what constitutes good clustering is important to both formulation of clustering algorithms and their evaluation. Single link, complete link, group average, Ward's method, and weighted average are hierarchical clustering algorithms used in document clustering [4, 5].

The main challenge in a clustering problem is determining which document features are discriminatory. Most existing clustering approaches represent each document as a vector, reducing a document to a representation fit for traditional data clustering approaches. Document clustering are used to enhance search engine results, web crawling, document organizing and in information retrieval. The major clustering applications in information retrieval include search result clustering, collection clustering, cluster-based retrieval, Scatter-Gather clustering and language modeling. In search result clustering, search results are clustered to ensure that similar documents appear together. Effective information presentation for exploratory browsing is achieved through collection clustering. In Scatter-Gather clustering, user selected groups are merged and this is clustered again. Document clustering in language modeling leads to increased precision and/or recall. Clustering speeds up the retrieval process, as document matching the query contains similar documents in the same cluster.

Different methodologies for clustering exist; some popular used methods include hierarchical method, partitioning algorithms, graph-based, density and grid based clustering [6, 7]. Hierarchical algorithms and Partitional algorithms are the main groups of Clustering algorithms for document clustering. In Hierarchical method, documents are represented in a multi-level and tree-like structure [8]. These methods get better quality clustering results but it is impossible to reallocate documents [9] with time complexity being quadratic. In Partitional algorithms, documents are clustered in a single level. Partitional algorithms are applied to large document collection due to low computational requirements [10]. The clusters might overlap here. A popular partitional algorithm is the K-means algorithm that performs in linear time complexity.

2. RELATED WORKS
Singh et al [11] presented performance evaluation for clustering text documents based on K-means, heuristic K-means and fuzzy C-means algorithms. Different representations such as tf, tf.idf & Boolean was used for evaluation. Feature selection schemes such as with or without stop word removal & with or without stemming were considered. Implementations were run on some standard datasets and various performance measures for these algorithms were computed. The experimental results indicate that tf.idf representation with stemming achieves better

clustering. Fuzzy clustering is a more stable method, and achieves better results for almost all datasets.

Babu et al [12] proposed a relevant document information clustering algorithm for web search machines. k-means partitioning algorithms and Hierarchical clustering algorithms used in clustering process have lot of disadvantages. They are usually slow and cannot be applied to large database. Thus, to overcome the shortcomings of the k-means algorithm, fast greedy k-means algorithm is used which is efficient and more accurate. The proposed algorithm computes the distortion for this algorithm in an efficient method. Experiments demonstrate that the proposed algorithm find the relevant documents more efficiently than by relevance ranking.

Chihli Hung et al., [13] proposed a document vector representation approach for extraction of relationships for document clustering. The proposed method merged statistical methods, competitive neural models, and semantic relationships from symbolic Word-Net. Reuter's corpus was used for evaluating the proposed method. The semantic lexicon was converted into its hypernym version word by word and topic by topic. Further to reduce the number of words in the dataset, the sibling words were replaced with different hypernyms. Experiments included a comparison of TFxIDF and ESV representation based on MLP and SOM using 10,000 full-text news articles. A 15 × 15 unit for each model was used. The experimental results showed that CA ranges from 75.25 and 81.72 percent for 10,000 full-text documents and AQE ranges from 2.538 to 3.511.

Smeaton, et al., [14] proposed mechanism for efficient document clustering and retrieval. Computational overhead is an issue faced by document clustering due to NXN similarity matrix required for creating clusters. Larger the documents collection, larger the similarity matrix required for information retrieval. The proposed mechanism implements the clustering process by applying several thresholds within the cluster generation process. Thus, leading to scalable clustering for large collection of documents. The proposed method used complete link clustering and was evaluated using the database of a newspaper.

Deng Cai et al., [15] proposed a novel document clustering method for clustering the documents into different semantic classes. The high dimensionality of document space was projected into a lower-dimensional semantic space using Locality Preserving Indexing (LPI). LPI relates documents of the same semantics close to each other. Both the geometric and discriminating structures of the document space were learnt. A modified LPI algorithm was proposed for document clustering. The proposed algorithm was applied on Reuters-21578 and TDT2 which performed better than the LSI-based clustering algorithm and close to the traditional spectral clustering algorithm.

Jones, et al., [16] proposed adapting genetic algorithm (GA) method for document clustering. The GA provides an efficient clustering when compared with the deterministic algorithm. The experiments used three different dataset and the effectiveness of the resulting clusters for searches was evaluated. E-measure was used as the performance measure; smaller the E values better the retrieval performance. Experimental results concluded that the value of E was smaller for the nearest-neighbor than the GA.

In this paper, a new method for clustering documents is proposed. In the proposed method, the term frequency of the document collection is computed and contexts based terms are fused. Agglomerative clustering and Bisecting K-Means are used to cluster the extracted features. Similarity measures like cosine function and correlation coefficient function are used during clustering. The paper is structured as follows: Section 1 gives a brief introduction about document clustering, its applications and Section 2 dealt with related works with respect to different types of clustering algorithm. Section 3 details the methodology, section 4 gives the results and discussion and section 5 concludes the paper.

3. METHODOLOGY

Term frequency is a statistical measure used to find the most crucial terms in documents. The number of times a term appears in a document is the term frequency. The relative importance of the term in the documents cannot be judged from the term frequency. Thus, to give weightage to the terms, a weighted term frequency is used and it is given as follows:

$$wf_{t,d} = \begin{cases} 0 & if \quad tf_{t,d} = 0 \\ 1 + \log tf_{t,d} & otherwise \end{cases}$$

The term frequency weight is inversely proportional to the word frequency. Higher the frequency of a word in a document, lower is its weights, thus, the measure is called term frequency-inverse document frequency (tf-idf) [17]. Most of the document clustering algorithm represent the documents and query in a vector space model. In vector space, each document d is considered to be a vector in the term-space. The it-idf is represented by the vector as:

$$d_{tfidf} or w_{i,d} = \left(tf_1 \log\left(n/df_1 \right), tf_2 \log\left(n/df_2 \right), ..., tf_m \log\left(n/df_m \right) \right)$$

where tf_i is the term frequency of the *i*th term , *n* is the total number of documents, and df_i is the document frequency. The length of each document vector is normalized to that of a unit length.

Cosine function [18] are used to compute the similarity between two documents d_i and d_j. The cosine of the angle between the two documents gives the distance between the documents. The vectors of the documents are normalized by dividing each of its components by its length as follows:

$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$

This maps the vectors in unit sphere:

$$\left| \vec{d}_j \right| = \sqrt{\sum_{i=1}^{n} w_{i,j}} = 1$$

The similarity is given as:

$$sim\left(d_j, d_k \right) = \frac{\vec{d}_j . \vec{d}_k}{\left| \vec{d}_j \right| \left| \vec{d}_k \right|}$$

The cosine is computed as the dot product for normalized vectors as follows:

$$\cos\left(\vec{d}_j, \vec{d}_k \right) = \vec{d}_j . \vec{d}_k$$

Pearson's correlation coefficient is another popular similarity measure used to measure the extent of relation between two vectors [19]. The measure ranges from +1 to -1. For a term set $T = \{t_1, .., t_m\}$, the coefficient is given by

$$PCC(\vec{t}_a,\vec{t}_b)=\frac{m\sum_{t=1}^{m} w_{t,a}xw_{t,b}-TF_a xTF_b}{\sqrt{\left[m\sum_{t=1}^{m} w_{t,a}^2-TF_a^2\right]\left[m\sum_{t=1}^{m} w_{t,b}^2-TF_b^2\right]}}$$

where $TF_i=\sum_{t=1}^{m} w_{t,a}$

The correlation coefficient selects terms that are highly indicative of membership in a category. Correlation coefficient is given by:

$$Co(T,C)=\frac{(TP*TN-FN*FP)}{\sqrt{(TP+FN)(FP+TN)(TP+FP)(FN+TN)}}$$

where TP is true positive, FP is false positive, FN is false negative and TN is true negative.

A bottom-up clustering method is the agglomerative hierarchical clustering [20]; where the process starts with a single entity/small cluster and agglomerates/merges to form the next higher level clusters leading to clusters having sub-clusters and they in turn have the same. Agglomerative hierarchical cluster algorithm agglomerates the closest pair of clusters in successive iteration by satisfying a similarity criterion, till all data is in one cluster. These steps are repeated till the required clusters number is obtained or distance between two closest clusters is above a specific threshold distance. Cluster similarity is computed using any distance measurement like Euclidean, Euclidean squared, Pearson correlation, or Spearman method. The disadvantage in this procedure is that an entity cannot be relocated, and using different distance metrics could lead to differing results [21].

The Agglomerative clustering process is as follows:

- Each object is given to a separate cluster.
- Intra cluster distances are measured.
- Construction of a distance matrix using distance values.
- Selection of cluster pair with shortest distance selected.
- Selected pair merged and removed from matrix.
- Distances from new cluster to other clusters calculated and updated in the matrix.
- Repeat till distance matrix is reduced to a single cluster.

Basic K-means Algorithm, a centroid-based approach, generates both k number disjoint and flat clusters [22]. The K-means method is unsupervised and iterative. The K-mean has the following clustering process:

- Selects K points randomly as initial centroids.
- Assigns all entities to closet centroid.
- Cluster centroid recalculated.
- Repeats above steps till the centroids remain the same.

The drawback of K-means method is that the clusters quality is difficult to compare and cluster size varies widely. These problems are addressed by the Bisecting K-means method which starts with a single cluster of all documents and works in the following manner:

- Picks a cluster to split.
- Basic K-means algorithm used to find 2 sub-clusters.
- Repeats above bisecting step a specific number of times and takes the split which produces clustering with highest overall similarity.
- Repeats above steps till desired cluster number is reached.

Using the concept of Hypernym words represented as a two level tree are fused and the TDF computed.

4. RESULTS

The transcribed Reuters dataset is used for evaluation of the proposed method. The dataset consists of 10 class labels. The obtained feature set was reduced based on the importance of the word with respect to the class label. Nearest neighbor and Agglomerative clustering are used to classify the clusters. The classification accuracy of the clusters using different methods is shown in Table 1 and Figure 1.

Table 1: Classification Accuracy

Number of clusters	K Nearest Neighbor	Agglomerative clustering
5	73.47	77.55
10	**76.53**	78.57
15	75.51	**82.65**
20	75.51	80.61

From Table 1 it is seen that the classification accuracy obtained is comparable with other techniques available in the literature. Figure 1 shows the plots obtained.

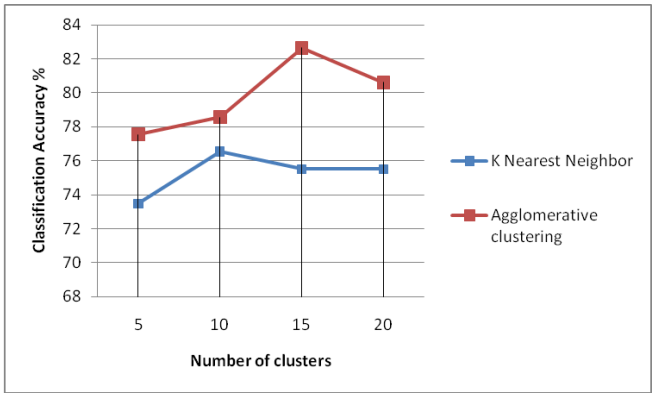

Figure 1: The classification accuracy for different number of clusters.

5. CONCLUSION

Document clustering helps organize large volume of unstructured text data for easier organizing of information. When context of the selected word in the document is considered, information search for relevant information is efficient. In this paper, it was proposed to investigate a novel method of word fusion based on word context using hypernym. The transcribed Reuters dataset was used to evaluate the hypothesis. The results obtained are promising with classification accuracy of up to 82.65%.

6. REFERENCES

[1] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

[2] Patel, D., & Zaveri, M. (2011). A Review on Web Pages Clustering Techniques. Trends in Network and Communications, 700-710.

[3] Everitt, B.S. Cluster Analysis. London: Edward Arnold, 1993.

[4] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

[5] Oikonomakou, N., & Vazirgiannis, M. (2010). A Review of Web Document Clustering Approaches. Data Mining and Knowledge Discovery Handbook, 931-948.

[6] Grira N, Crucianu M, Boujemaa N (2005) Unsupervised and semi-supervised clustering: a brief survey. In: 7th ACM SIGMM international workshop on multimedia information retrieval, pp 9–16.

[7] Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 264–323.

[8] Mahdavi, M., Chehreghani, M. H., Abolhassani, H., & Forsati, R. (2008). Novel meta-heuristic algorithms for clustering web documents. Applied Mathematics and Computation, 201(1), 441-451.

[9] Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 264–323.

[10] Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. KDD'2000. Technical report of University of Minnesota

[11] Singh, V. K., Tiwari, N., & Garg, S. (2011, October). Document Clustering using K-means, Heuristic K-means and Fuzzy C-means. In Computational Intelligence and Communication Networks (CICN), 2011 International Conference on (pp. 297-301). IEEE.

[12] SureshBabu, Y., Mutyalu, K. V., & Prasad, Y. S. (2012). A Relevant Document Information Clustering Algorithm for Web Search Engine. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(8), pp-16.

[13] Chihli Hung ,Stefan Wermter and Peter Smith, " Hybrid Neural Document Clustering Using Guided Self-Organization and WordNet", IEEE Intelligent Systems Volume 19 Issue 2, March 2004 .

[14] A. Smeaton, M. Burnett, F. Crimmins, and G. Quinn. An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts. In BCS-IRSG Annual Colloquium on IR Research, Workshops in Computing, 1998.

[15] Deng Cai, Xiaofei He, and Jiawei Han, "Document Clustering Using Locality Preserving Indexing", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 12, DECEMBER 2005, pg(1624-1637).

[16] Jones, Gareth, Robertson, Alexander M., Santimetvirul, Chawchat and Willett, Peter (1995) "Non-hierarchic document clustering using a genetic algorithm". Information Research, 1(1).

[17] Raghavan VV, Birchand K (1979) A clustering strategy based on a formalism of the reproductive process in a natural system. In: Proceedings of the second international conference on information storage and retrieval, pp 10–22

[18] Salton G (1989) Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley

[19] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. Noise reduction in speech processing, 1-4.

[20] El-Hamdouchi, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. The Computer Journal, 32(3), 220-227.

[21] Rokach, L., & Maimon, O. (2005). Clustering methods. Data mining and knowledge discovery handbook, 321-352.

[22] McQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings f the fifth berkeley symposium on mathematical statistics and probability, pp 281–297