IJCTA, 9(31), 2016, pp. 359-367 ISSN: 0974-5572
© International Science Press

Document Clustering Based on Semantically Enhanced Document Representation

N. Kannammal* and S. Vijayan**

ABSTRACT

Document clustering algorithms are important to categorize the growing web documents while pre processing for applications like recommendation system. Each documentis represented in a vector space model with frequency or occurrence of terms present in the document. The number of terms decides the dimension of document vector.The clustering algorithm which works on vector space model suffers from the problem of handling high dimensional space which increases computation complexity. The documents are clustered to relevant cluster only based on similarity between numerical value(i.e. term occurrences) of document vectors and without considering hidden semantic information in the documents.The inclusion of semantics at the document clustering level will result in qualitative recommendation very close to the query. The proposed work reduce thedimension on using semantic information and evaluated on two different clustering algorithm K-means andCover Coefficient (CC) concept based clustering methodology. The proposed method isstudied on comparing with standard K-means clustering algorithm without dimension reduction and semantic.The modified algorithms results in improved performance that is evaluated based on F-measure.

Keywords: Document clustering, Cover coefficient, dimensionality reduction, semantic similarity.

1. INTRODUCTION

Clustering process is applied to form meaningful categories of large pool of data's. Best cluster follow the hypotheses that the candidates of the cluster groups are more similar with candidates of the same cluster and more dissimilar with the candidates of different cluster. This clustering of information data's is very much essential for query- response, data information retrieval, to discover service,mining, information management system etc. It is in most of the system, clustering of documents is done as pre processing step or as post retrieval process [1].Text orDocument clustering is efficient technique used in all process which involves meaningful information usage. The clustering or grouping leads to fast retrieval of knowledge from documents. Document clustering methods group the documents based on their similarities. So like documents are partitioned under one cluster after the clustering process. The similarity between the documents is a measure of its contents based on occurrence of lexical or semantics or both.

1.1. Document representation

Traditional way to represent the document is by bag of words (terms) present in the documents. The terms are given as a vector. Each document corresponds to one vector. The content of this vector space model (VSM) may be binary or numeric. The presence or absence of a term is indicated by 1 or 0 in binary. The numeric value will be the term weight given for each term.As the role of clustering technique is grouping of documents, the categorizing is done by finding similarity between two document vectors.

* Assistant Professor, Dept. of Computer Science and Engineering, Surya Engineering College, Perundurai, Erode, India, *Email: n.kannammal@yahoo.in*

** Professor, Dept. of Electrical and Electronics Engineering, Surya Engineering College, Perundurai, Erode, India, *Email: svijayansurya@gmail.com*

Due to rapid growth of web world, the response for a single query may need to be worked with large number of documents. So the search space is huge consisting of thousands of documents. Each documentcontains thousands and thousands of terms which are considered as features or attributes of the document vector. The number of terms in document decides the dimension of vector [2]. Moreover, all the feature value need not be available for all the documents as they differ from each other which mean many entries are zero. This problem of the huge dimension of vector andsparsity of value deteriorates the efficiency of clustering process [3].

1.2. Clustering issues

Document clustering is important technique that result in grouping, pattern creation, summarization and navigate. The precision and recall, accuracy, purity, execution time, memory usage and complexity are measures to decideefficiency of a clustering algorithm.

In the existing growing big volume of data following are the notable issues that a clustering algorithm should handle[4].

(a) Big volume of data

(b) Curse of dimensionality

(c) Predefining number of clusters

(d) Not considering hidden semantics

(e) Clustering for dynamic environment

The proposed work is carried out with the aim of solving (b),(c) and (d) issues.

2. RELATED WORK

General human characteristic follow a divide and conquer method to sort out a problem i.e. we follow dividing a complex problem into many simple ones to reach a solution. The clustering technique highly coincide this idea by grouping similar instances under one class so that we can discard irrelevant classfrom problem solving.The cluster is to be balanced for dynamic environment without additional cost and complexity [5].

Based on structure and pattern of clusters,Hierarchical and Partitioning are two major approach of clustering algorithm[6]. Both the algorithms differ in method followed to form clusters.Hierarchical clustering method follows combining two similar objects forming dendograms or tree either in top-down or bottom-up approach.On the other hand, partition algorithm, form k partition of document collection based on some criteria selecting one centroid as cluster head. The other documents are assigned to the nearest leader on finding similarity between the document and cluster centroid. K-Means and its variant K-mediods, Bisect K-Means are familiar partition algorithms.As the proposed work is a partition type of algorithm, the existing partition k-means algorithms are studied and the work is compared with standard k-means algorithm[7].

2.1. Dimensionality reduction

The dimension of the document is number of features (terms) preprocessed in the document. The feature vector is formed after pre processing step like stop word removal and stemming to the document. The clustering algorithm working with high dimensionality is a big challenge which results in computationally expensive, introduce noise and highly complex. The problem is technically stated as*Curse of Dimensionality* [8]**.** All the dimensions will not contribute to clustering and high dimension also lead to data sparsity as many entries are not filled. So the dimensionality is reduced to remove irrelevant features, to avoid redundancy of data and to increase the performance of clustering.

High dimensionality can be reduced by two methods:*Feature extraction and feature selection*[9].Feature extraction method estimates a low dimension result for a high dimensional vector input. LDA (Linear discriminate analysis, PCA (Principal component analysis) and SVD (Singular Value Decomposition) are common methods for feature extraction**.**As the proposed work is based on feature selection method, we restricted the content with feature selection.Feature selection is done by any of the follow models: filter, wrapper, embedded and hybrid. Feature selection is discriminant subset formation of original feature set reducing dimension. A method which needs a class label and training set is called *supervised learning* and another method is a data clustering of similar instances is called *unsupervised learning*. Wrapper is based on classifier and computational expensive but accuracy is good than filter method. So hybrid balances the computational complexity and accuracy of filter and wrapper. Feature selection results either in subset or weighted feature set.Many of the previous research have supported feature selection is better than feature extraction as cannot be traced back to original space once extracted. Whereas, feature selection maintains the original information in the new reduced space. Also it achieves high interpretability. Feature selection is deriving a part from larger vector with fewer dimensions and without loss of originality in information. IG (Information Gain), Chi square, Fisher score are some commonly used statistical methods of feature selection [10], [11], [12]. These methods lacks hidden meanings between document words, dependency between the terms, relevancy. So feature selection based on semantic information increase the performance of categorization.For automatic categorization, the work[13] has used the noun synonyms and sense of terms using wordnet ontology. The experiment was done on Reuters-21578 dataset and proved better than statistical methods.In [14], the work was contributed to the importance of polysemous and synonymous noun for document clustering. On disambiguating this nouns, core features are derived. 90% of feature reduction was achieved.In the work of[15], the authors have incorporated the background knowledge by phrase and word based semantic weight. The ontology with NLP procedure improved the performance of document clustering. The proposed work of [16] combined both statistical and semantic based feature selection method to select relevant features on background knowledge. They have justified on improving the clustering performance.The proposal of [17], uses statistical method for feature selection and semantic based clustering was used to improve clustering results.

3. MATERIALS AND METHODS

Vector Space Model (VSM) is commonly followed method to represent a document. Each vector represents one document. The values of the vector indicate the features of the document. It can be binary or weight value.As the document constitutes number of terms, the vector for the document grows which deteriorates the clustering efficiency and difficult to handle. Reducing dimension of vector is indispensable to improve the performance of clustering technique when applied for information retrieval or data mining[18]. In a simple manner, D is a collection of document to be clustered and denoted as {d1, d2……dm}. Each d is represented as vector {v1, v2…..vn}of frequency or weight. Document-term matrix of order m x n with 'm' documents and 'n' terms as dimension is taken for clustering. In clustering process, grouping is done by similarity measure taken between any two documents. More similar valued documents are clustered.Euclidean Distance, Cosinesimilarity,Jaccard Coefficient, Pearson Correlation are some similarity measures used for document clustering.

3.1. Supporting concepts

K-means is a partitioning type algorithm.The number of initial clusters centres is randomly assumed by users. Basic steps of K- meanclustering is given below:

1. Randomly select k cluster centroids
2. Assign each instance to nearestcentroid

3. Update cluster centroid
4. Repeat step 2 and step 3 until centroids are stable

K-meansresults in local optima. It achieves linear complexity. High computational efficiency, easy implementation, quick convergence even for large dataset is some special characteristics of K-means clustering[19].Sequential k-means, k-mediods, bisecting k-means are some of its variants. The standard algorithm is also prone to some limitations:

a) Sensitive to random selection of initial centroids on which optimisation depends

b) Sensitive to sparsity or noisy data *(K-mediods overcome sparsity but cost is expensive)*

3.2. Cover Coefficient Clustering Methodology (C^3M) [20], [21]

C^3M is a single pass partitioning type of algorithm. Basic steps of *C^3M* algorithm is given below:

1. Find Number of cluster Nc
2. Find Seed power for each document
3. Identify first Nc seed document on sorting
4. Each non seed document is assigned to seed which covers it maximum
5. Uncovered documents are moved to ragbag cluster

As a pre processing step, documents are represented as *m x n* order document-term *D* matrix with $\{d_1, d_{,} ....d_m\}$ as rows and $T = \{t_1, t_2, ...t_n\}$ as column. Each entry d_{ij} (binary or weighted) is a occurrence or non occurrence or weight of the term in document.

A new document-by-document C matrix to find out each document coverage is computed by double stage probability. Each entry of C matrix, $c_{ij}(1 \le i, j \le m)_{is}$ computed as

$$c_{ij} = \alpha_i \times \sum_{k-1}^{n} d_{ik} \times \beta_k \times d_{jk}, \quad 1 \le i, \quad j \le m$$

Where

d_{ik} is probability that term t_k appear in d_i

d_{jk} is probability that a document d_j can be derived based on term $t_{k.}$

α_i is 1 / (sum of i^{th} row)

β_k is 1 / (sum of k^{th} column)

3.3. Inference from *C* matrix

3.3.1. Decoupling and Coupling Coefficient

Each entry in *C* matrix describes the amount of coverage between d_i and *dj* document which is taken as similarity evaluation between two documents.

The diagonal entry c_{ii} is coverage of document to itself which is maximum.

The diagonal entry c_{ii} gives a value which shows how much a document is dissimilar fromdocuments,otherwisecalled *decoupling coefficient* $\delta_{i.}$

The other versa is *coupling coefficient* ψ_i which is 1 - δ_i or i^{th} row off-diagonal entries sum.

3.3.2. Number of clusters

Many existing algorithm use some heuristic approach to predetermine the number of clusters for the collection. Number of clusters give clusters heads towards which the remaining members are attracted. The cluster number is maximum when the documents are more dissimilar and classification of term leads to classification of documents.

3.3.3. Seed Power

In order to derive cluster leaders, seed power for all the documents in initial collection is obtained by using the formula

$$P_i = \delta_i \times \psi_i \times \sum_{j=1}^{n} d_{ij}$$

Seed power is sorted. Documents with highest seed power form clusters.

3.3.4. Clustering

The entries in C matrix give the extent to which a document covers another document. The value is based on the probability to which a term is present in a document and the probability to which documents having the term can be derived. The first stage probability works for the former of above sentence and second stage probability is for latter one. For clustering, a non seed document whose C matrix entry is high against a seed document is covered maximally. The non seed document is assigned to that seed document.

3.3.5. Size of D matrix

The D matrix is a doc *X* term matrix which gives the probability of occurrence of a term in a document. When the number of terms in the D column increases, the computational space increases as the C entries depends on D matrix terms. In order to reduce the number of dimension and in turn the computational time, the proposed work follows a feature selection method.

Lack of Semantics

According to the CC concept, the coverage between two documents is based on two stage probability of occurrence of terms in the two documents. Even if the term is irrelevantto a document it will contribute to the coverage value and this can be omitted. When a term is not semantically important to a document, the term is not that much important to a document. The proposed work takes into account the semantic relation between the terms to decide the coverage between the seed and non seed document.

3.4. Proposed Work

Based on the feature selection method, the number of terms is reduced in the proposed work. The following are the ideas considered for double level dimensionality reduction of term vectors.

1. The terms which are related to more number of othertermsin many rela tions (merynym, hyponym, hypernym..) are considered to be more important to the documents.
2. The term frequency in each document decides the importance of that term to that document. Also the term which occurs very frequent and very rare gives less contribution.
3. The term which occurs more times in all documents will not be a good discriminator of a document. (tf - idf value is used).

3.4.1. First level reduction

The terms in document term vector is checked for its semantic relation with other term in the vector using WordNet synsets which consists of meaning of a term (only synonyms and hypernyms are taken for the work). Mutual relation between two terms of document is checked and counted.

Let *Dc* be document collection and $d_1, d_2 .... d_n$ be documents belong to Dc

Let $t_1, t_2 ......t_m$ be the terms of document.

Let D be a *doc* x *term* in the order *n*x *m*

Let *d(i, j)* be the count for each term with zero initially. Mutual semantic relation count for each candidate term t_j using Wordnetsynset is given as

For each t_{j1}< *is in wordnet_synset* > t_{j2}

$i = 1 .... n$

$j1 = 1 ....m, j\ 2 = 1 .... m, j1$□□ $\neq j2$

d(i, j) = d(i, j) +1 when a term is present in synset of candidate term and

d(i, j) =d(i, j) otherwise

The resultant count of each term weighs the number of semantic link that a term have.

The *total semantic_ link_ weight*

$W_{sem—link(i,j)} = d(i, j)$

The term with maximum count is semantically very important term for a document to extract semantic information of the document. The term with *d(i, j)* > λ is filtered for next level reduction, where λ is threshold and λ = *1* for our work

3.4.2. Second level reduction

The term which is more related to other terms in a document and also occurs in many documents can bring out semantic of the document but cannot be an individualdiscriminator to document. The frequency of occurrence of a term in a document and in a collection is evaluated through *tfidf value.tf(term) gives term frequency of a term .*

tf(i, j) be the frequency of term t_j in document d_i

df(j) gives the number of documents that have t_j. The value *tfidf* (term frequency/inverse document frequency) is given as;

*tfidf(i, j)=tf(i,j)*log(N/df(j))* where *i* = 1 … *N* documents

The *total tfidf_weight*of each term

$W_{tfidf}(t_j) = \sum_{i=1}^{N} tfidf\left(i, j\right)$

In the first level reduction, the pruning of the term is decided by semantic correlation information. Similar to this level, second level cannot be decided based on tfidf value as more semantically related terms may be discarded due to minimum tfidf. This leads to loss of information. The dimensionality reduction with maximum pruning of terms is not encouraged as it leads to loss of information which affects cluster accuracy.

Therefore, after second level reduction an aggregation of semantic link count and tfidf frequency is done to arrive total weight of a term.

Total weight W(*tj*) = $W_{sem\_link\,+}W_{tfidf}(t_j)$

The value of W(*tj*) falls in any one of the following cases:

Case 1: terms which have maximum mutual semantic count and maximum*tfidf* weight is very important to document

Case 2: terms which have maximum semantic count and minimum *tfidfweight* can contribute semantic information to document similarity.

Case 3: terms which have minimum semantic count (i.e. term which is not correlated with many terms) and maximum *tfidfweight (i.e. frequency in particular document)*. More occurring terms are important to document.

Case 4: terms which have minimum semantic link count and minimum tfidf weight are not important.

The terms with the W(*tj*) > threshold(σ) form term vectors for a document. The threshold is fixed to satisfy first three cases.

Algorithm:

```
Onto_Dim_Reduction ($Dc$, Tc, $\lambda$, $\sigma$){
    1. Input:
       $D^{nxm}$: doc x term matrix with n documents and m terms
       $\lambda$: minimum number of semlinkthreshold
       $\sigma$: minimum frequency occurrence threshold
    2. Output:
       $D'^{n\,xm}$: doc x term matrix with reduced dimension
    3. For each $t_j$ of $d(i, j) \in D, i = 1 \ldots n, j = 1 \ldots m,$
       Compare with $t' \in (d(i, j) - t_j)$ terms
    4. If a term is present in synset of another term then they are semantically related and $d(i,j)$ of term $t_j$
       is incremented.
    5. For each $t_j$ of $d(i, j)$ $D, i = 1 \ldots n, j = 1 \ldots m,$
       If $(d\ (i, j) > \lambda)$ then
       $tfidf(i, j) = tf(i, j) * log(df(j))$
       $d(i, j) = d(i, j) + tfidf(i, j)$
    6. For each $d_i \in D$ retain the terms with
       $d(i, j) > \sigma$as final dimension
```

4. EVALUATION METRIC

Our proposed system is evaluated by standard metrics of relevance such as precision, recall and F-measure in the field of information retrieval for quality assessment.

1. Precision = TP /(TP+FP)
2. Recall =Tp /(TP+FN) Where

TP: No. of items are relevant, recalled

FP: No. of irrelevant items recalled

FN: No. of relevant items that are not recalled

3. F-measureis a trade off between precision and recall. It is used to prove the effectiveness of retrieval and is given as

$$F_{ij} = \frac{2.precision(i,j).recall(i,j)}{precision(i,j) + .recall(i,j)}$$

Larger F-measure denotes higher quality of clustering.

5. EXPERIMENT AND EVALUATION

The above proposed work is worked on IMDB dataset with the standard Kmeans and C^{3}M concept based new clustering method.The document collection to be clustered is pre-processed by stop word removal and porter stemming process. The evaluation metric precision, recall and F-measure are used to study the proposed work. The following table is the details of metrics for 640 documents

Table 1
Precision, recall and measure for existing and proposed.

Algorithm	*Precision*	*Recall*	*F-measure*
K-Means	0.662	1.0	0.699
K-Means + Onto VSM	0.684	1.0	0.724
C^{3}M + Onto VSM	0.833	0.814	0.757

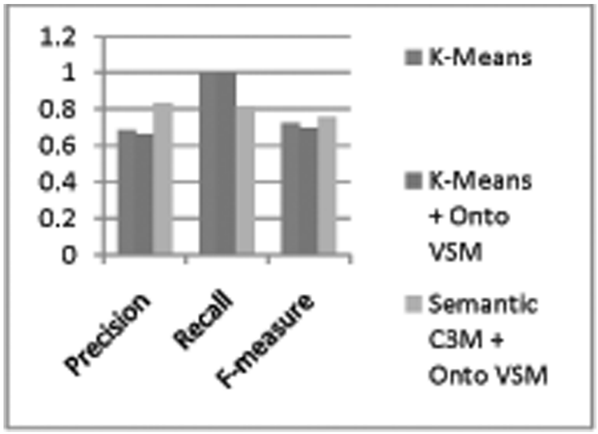

Figure 1: Comparison of precision, recall and F-measure between K means and proposed system.

The evaluation exhibit that the when basic Kmeans clustering method and C3M based clustering method are improved semantically with wordnet ontology , the dimensionality reduction based on semantic knowledge of each term preserve the information without loss and also show an improvement in performance than standard Kmeans.

6. CONCLUSIONAND FUTURE WORK

The document representation merely using terms will reflect on performance of any clustering algorithm. Most of the existing techniques use the total term vector as such for finding document similarity process of clustering algorithm which have to face issues like noisy, sparsity, difficult to store and more computation time. The numbers of occurrence alone will not enough to decide importance of term. Moreover, important word may occur minimum time in a document to convey the information. Such terms are treated less important in frequency method. This issue can be solved only by hidden semantics of the terms that are revealed through ontology. Our proposed work concentrates on dimensionality reducing based on semantic knowledge. On our experiment, we studied that K-means as usual performs good in computing speed than C3M. Introducing semantic reduces dimension and improves F measure than basic algorithm. The seed selection in C3M algorithm overcome the random selection of number of cluster in K means giving good F measure which reflect in accuracy of algorithms. The future work is to include semantics for cluster labelling and response a query on purely semantic based search by using the above two clustering methodology in a recommendation system.

REFERENCES

[1] D.R. Cutting, D.R. Karger, J.O. Pedersen and J.W. Tukey, Scatter = Gather: a cluster-based approach to browsing large document collections, in: Proc. of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), pp. 318–329, 1992.

[2] L Jing, L Zhou, MK Ng, JZ Huang-Ontology-based distance measure for text clustering- Proc. of SIAM SDM 2006.

[3] G. Chandrashekar, F Sahin, A survey on feature selection methods - Computers & Electrical Engineering, 2014 – Elsevier

[4] Issues, Challenges and Tools of Clustering Algorithms, ParulAgarwal,M. AfsharAlam,Ranjit Biswas, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011 .

[5] Fazli CAN , Esen AOZKARAHAN, A dynamic cluster maintenance system for information retrieval, Proceedings of the 10th annual international, ACM,1987

[6] Lior Rokach, Oded Maimon, Clustering Methods, Data Mining and Knowledge Discovery Handbook, pp. 321-352.

[7] P. Berkhin, A Survey of Clustering Data Mining Techniques, Grouping multidimensional data, 2006 – Springer

[8] T. Hastie, R. Tibshirani, and J. Friedman.The Elements of Statistical Learning.Springer, 2001.

[9] Vipin Kumar and SonajhariaMinz, Feature Selection: A literature Review, Smart Computing Review, vol. 4, no. 3, June 2014.

[10] M. Dash, H. Liu, Feature selection for clustering, Pacific-Asia Conference on Knowledge Discovery and …, 2000 – Springer

[11] X Wang, K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, Pattern recognition, 2003 – Elsevier

[12] S. Alelyani, J Tang, H Liu, Feature Selection for Clustering: A Review Data Clustering: Algorithms and Applications, 2013.

[13] Stephanie Chua, Narayanan Kulathuramaiyer, Semantic Feature Selection Using Wordnet, WI '04 Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence.

[14] S. Fodeh, B. Punch, P.N. Tan, On ontology-driven document clustering using core semantic features, Knowledge and information systems, 2011 – Springer

[15] B. Drakshayani and E. V. Prasad. 2012. Text Document Clustering based on Semantics.International Journal of Computer Applications

[16] A. Benghabrit, B. Ouhbi, H. Behja and B. Frikh, 2013.Text clustering using statistical and semantic data. Proceedings of the IEEE World Congress on Computer and Information Technology

[17] Thangamani.M and P.Thangaraj, Semantic clustering with feature selection for text documents, International J. of Engg. Research &Indu.Appls, Vol. 3, No. II (May 2010).

[18] G. Salton, and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Inc., 1983.

[19] T. Kanungo, D.M. Mount, N.S. Netanyahu, An efficient k-means clustering algorithm: Analysis and implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 24, Issue: 7, Jul 2002 .

[20] Can, F.: Incremental clustering for dynamic information processing. ACM Transactions on Information Systems 11(2), 143–164 (1993)

[21] F. Can, IS Altingövde, E Demir, Efficiency and Effectiveness of Query Processing in Cluster-Based Retrieval, Information Systems, 2004 – Elsevier.