Journal of Engineering and Applied Sciences 12 (21): 5554-5559, 2017
ISSN: 1816-949X
© Medwell Journals, 2017

# Canonical Data Model for Text Document Clustering

Siti Sakira Kamaruddin, Yuhanis Yusof, Farzana Kabir Ahmad and Mohammed Ahmed Taiye
School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

**Abstract:** The abundance of text data have been witnessed with the growth of web and other text repositories. There is an important need to provide improved mechanism to effectively represent and retrieve text data. This study advocates the construction of Canonical Data Models for mapping contents of multi-documents into a few general models that can represent the corpus. However, to construct Canonical Data Model for text, it involves non-trivial text mining techniques prior to the actual construction process. Furthermore, constructing Canonical Data Models for all terms in a set of documents will be costly and will not reduce the sparsity problem that are associated with text document processing. In order to solve this problem we propose a two tier dimensionality reduction step adopting commonly used feature extraction and feature selection methods. The reduced features are then used to construct a Canonical Data Model. A Canonical Data Model for text documents can be used as a general model that has potential to act as a reference model for text comparison in a wide variety of text mining tasks such as text clustering, text classification, text summarization and text deviation detection. Experimental result reveals that the proposed approach produces better results compared to methods without Canonical Data Model.

**Key words:** Canonical Data Model, text clustering, latent semantic analysis, text dimensionality reduction, summarization, multi-documents

## INTRODUCTION

Within the past decade, the growth of textual data has created research opportunity to effectively manage and retrieve relevant information from enormous repositories of text. With the rapid growth in internet technology, clustering has been the foundation of many knowledge discovery tasks. Organizing text documents into sensible groupings is acknowledged to contribute in learning and understanding of the content. Hence, the objective of clustering is to find a convenient and valid organization of the documents and this can be achieved by identifying structure of the text. However, the problem that exists in identifying clusters of text documents is to specify what proximity is and how to measure it.

In text document clustering, the challenges arise from the nature of text which is high in volume, contains data sparseness and involves semantics (Jing *et al*., 2005). The popular text representation method used in text clustering is the term frequency distribution and vector based representation as reported in Salton and Buckley (1988) which transforms text documents into vectors which are given weights. Usually, even a medium sized text file contains high number of words.

Hence, the problem of sparsity of text data could not be solved using term based vectors which may grow up to a few thousand dimensions. The sparsity of text is related to density since sparsity and density are terms used to describe the number of word in a text document either high or low, respectively. In real world textual data, it is common that different terms are used by different writers to denote the same meaning. Therefore, sparsity also depends on the semantics since, text documents commonly contain different terms to convey the same message. An alternative method to model the semantics of natural language is needed to overcome the sparsity problem. In this research, we are interested to reduce the sparsity and semantic problem by creating a Canonical Model for the text documents.

In text clustering tasks, the proximity calculation between documents is an issue. It involves exponential measures since each document need to be compared with each other. There is a need to formulate a less computationally demanding method to perform proximity calculation. The construction of Canonical Model for text can overcome the exponential complexity of proximity calculation in clustering tasks. The major challenge here is to work on methods to effectively construct a Canonical Model for text documents. In a group of topically-related study, the extent of redundancy of the contents are higher as each study is compelled to give similar overview of the subjectmatter. Hence, this circumstance is an advantage,

**Corresponding Author:** Siti Sakira Kamaruddin, School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

since similar text patterns can be extracted from multi documents in order to construct the text Canonical Model.

**Document clustering:** Document clustering is an important task in text mining because of the volumes of textual data that are generated today. It helps to uncover the underlying patterns in text documents. Document clustering is the process of identifying documents that can be grouped together based on their similarity. According to Han *et al*. (2011), existing research on clustering can be categorized into four main branches: partition, hierarchy, grid-based and density-based method. Hierarchical algorithms (Narang, 2015) produce a hierarchy of clusters while partitioning algorithms gives a flat partition of the set. Hierarchical clustering approaches are commonly used on static document collection and it does not require estimation of cluster number. Nevertheless, its competitive clustering quality requires high computational effort due to the clustering process on multiple hierarchies (which may be unnecessary in many applications) (Narang, 2015). The grid-based methods, quantize the object space into a finite number of cells that form a grid structure (Han *et al*., 2011). Such, an approach works well to cluster objects of different attributes (Berkhin, 2002). Density-based clustering (Luo *et al*., 2009) groups text documents according to the similarity of their neighborhood (features of their neighborhoods). In the research presented by Yang and Wang (2009) a clustering algorithm that uses domain ontology to calculate the similarity between neighbors of short messages was introduced. The undertaken research is proven to improve the scalability of the clustering.

On the other hand, existing partition based clustering such as K-means (Balabantaray *et al*., 2013) is dependent on the number of initial centroids, however, it is the most popular clustering algorithm. In this study, K-means was used to highlight how the proposed method enhances the performance of the K-means algorithm.

## MATERIALS AND METHODS

**Cononical Data Model:** In categorical dataset a Canonical Data Model is often created to enable communications between different data formats (Gonzalez *et al*., 2011). Canonical Data Modeling has been successfully used for categorized data as shown in the by Gonzalez *et al*. (2011) and Dietrich and Lemcke (2011). Gonzalez *et al*. (2011) utilizes the Canonical Model to visualize the simulation interoperation in a collaborative system for disaster response simulation. The Canonical Model, consists of entities shows an abstract topological model of the physical layer and depict the interaction among simulators at an abstract and general level that allows exchanging information between the simulators every time step. From this study, it can be concluded that the Canonical Model is worthwhile in visualizing the complexity of the disaster response simulator design. In Dietrich and Lemcke (2011) a Canonical Data Model was constructed for input schemas. Here, the Canonical Data Model was created to capture the semantic equivalences of the input schemas. This also helped in differentiating the input schema where any deviations from the equivalent part was identified and detected.

There is a couple of research done on developing Canonical Models for text documents for example, in Iria and Ciravegna (2005) a canonical graph-based data model was developed to be used by all algorithms implemented within a text based relation extraction framework. Bloechle *et al*. (2006) a reverse engineering was performed on pdf documents to create canonical representation. However, their research focuses on the structure of the documents and not on the contents.

**Canonical Data Modeling approach for text clustering:** The proposed Canonical Data Modeling approach for text clustering consists of the following 6 phases.

**Text pre-processing**. The textual data need to be pre-processed prior to modeling. This process includes pre-processing steps such as word tokenization and stop word removal. Tokenization is the process of breaking a series of text into individual terms which are called tokens. Stop word removal is the process of identifying and removing words that will not contribute to the clustering task such as determiners and prepositons, e.g., “a”, “an”, “am”, “the”,“is”, etc.

**Feature extraction:** The next phase after text pre processing is the feature extraction phase. For feature extraction we implement the Vector Space Model (VSM). VSM is one of the classic way to represent a set of documents first introduced by Salton and Buckley (1988). It has been intensively used in the field of information retrieval. This method is sometimes referred to as “bag of words”. The popularity of this text representation scheme is due to its simplicity. Using VSM, each dimension corresponds to a weighted term in the document. The normal weighting schemes are term frequency, inverse document frequency or most popularly a combination of both, i.e., term frequency and inverse document frequency referred to as tf-idf. Term frequency (tf) computes the weight (importance) of a term in a document. It is based

on the idea that a term is important if it occurs more frequently in a document. Therefore, in this method each term is related to term frequency (tf) which is the number of occurrence of that particular word in the entire document. Document frequency is the total number of documents a word $w_i$ occurs. The fundamental concept of this text representation scheme is in the vector representation of documents where the vector of weights for a document $d_j$ is $d_j = (w_{1,j}, w_{2,j}, \ldots, w_{m,j})$ that is weighted term $w_{i,j} \times w_{i,j} = 1$ if term $t_i$ appear in document $d_j$ and equals to 0 otherwise. The weight of the term in a document is calculated using Eq. 1:

$$tf_{i,j} = \frac{f_{i,j}}{\max_h f_{h,j}} \quad (1)$$

Where:
$f_{i,j}$ = The frequency of the terms
$t_i$ = The document $d_j$

The denominator is the maximum number of occurrence of all terms in document $d_j$. The Inverse Document Frequency (idf) is to measure the discriminatory power of term $t_i$ and is computed using Eq. 2:

$$idf_i = \log_2 \frac{C}{n_i} + 1 \quad (2)$$

Where:
$n_i$ = The number of document that contains $t_i$
C = The size of the document

The relative frequency is computed using Eq. 3:

$$idf_i = \frac{\log C}{n_i} \quad (3)$$

The weight of term $t_i$ in a document $d_j$ is obtained by multiplying $tf_{ij}$ and $idf_i$ as shown in Eq. 4:

$$W_{i,j} = tf_{i,j} \times idf_i \quad (4)$$

The value of $w_{i,j}$ will be high if the term $t_i$ is frequent in the whole document $d_j$ and rare in the other documents. The concept of $tf_{ij}$ can be explained using the following example. If $d_1$ contains ten occurrences of the term "profit" is compared with another document $d_2$ which contains five occurrences of the term "profit". And if the term "profit" is the only term exists in the user query, $d_1$ will be ranked higher than $d_2$. Some of the advantages of this method is it is easy to implement and well-studied and it has shown effective results empirically. Using Eq. 4, each document will be converted into a sparse matrix of $tf_i$ df features.

**First level text clustering:** The next step is to perform text clustering on the tfidf features. Any text clustering algorithm can be used for this purpose. The aim is to obtain a first level clusters to create the Canonical Text Model which will then be used to perform the second level clustering. We advocate that these double clustering can improve the accuracy of the results.

**Selecting cluster features:** This phase involves the processes to identify relevant features that can accurately represent each clusters produced in the previous step. Cluster feature selection is a vital step in order to reduce the dimensionality problem associated with text document collection. This step will also reduce the text sparsity problem in any text representation schemes. In this research, we employ Latent Semantic Analysis (LSA). LSA is able to analyze how the terms in a set of documents are related. It produces a set of related concepts. The assumption made by LSA is if a word is semantically related it will occur in similar document. In LSA a factorization called Singular Value Decomposition (SVD) is used to decompose the high dimensional term document matrix into smaller dimensions by taking into consideration the associations between words in text documents (Song and Park, 2007). The dimensionality of data can be reduced using SVD by projecting data points onto the space spanned by the left singular vectors. Using this technique a term document matrix is decomposed into the product of three matrices as shown in Eq. 5:

$$X = U \times S \times V^T \quad (5)$$

Where:
U = A (u×n) matrix
S = (n×n) matrix
$V^T$ = (v×n) matrix with n "latent semantic" dimensions

The matrices U and V which represents terms and documents in the new space contains the left and right singular values of X, respectively and the diagonal of S contains the singular values of X in decreasing order.

A reduction of the original matrix into k dimensions can be performed by taking the k largest singular values (Hachey *et al.*, 2006). k is a user defined threshold that can be adjusted to determine the number of terms or features that will be extracted and identified as important. These terms will then be used in the next phase which is the text Canonical Model construction.

**Text Canonical Model construction:** The selected terms are then represented as Canonical Models. The output from the previous phase is a term by concept matrix. It shows how important a term is to the concept. The matrix

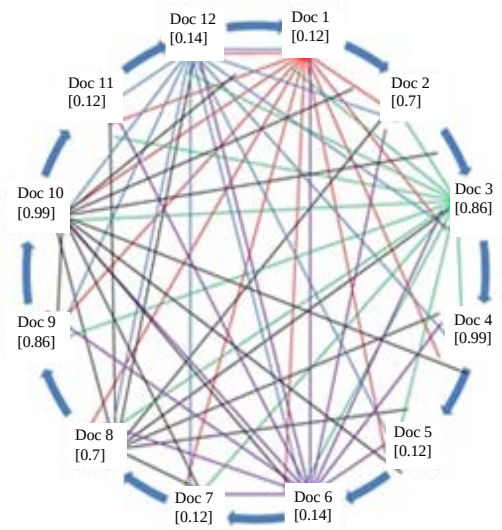


Fig. 1: Similarity measurement between documents

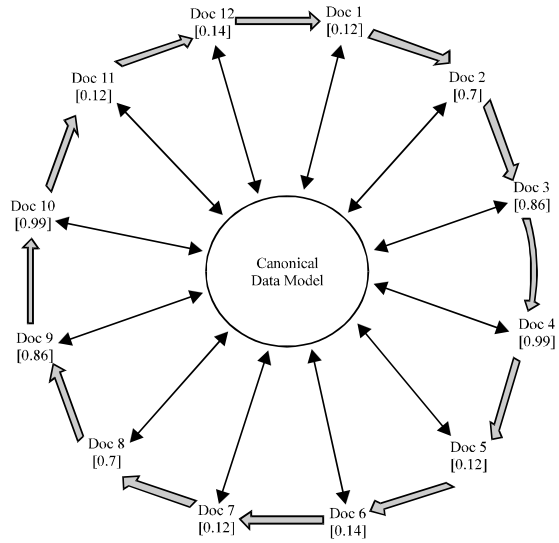


Fig. 2: Similarity measurement using Canonical Data Model

is used to extract the terms that are important to a certain concept. All the extracted terms will be used to construct a Canonical Model which is represented as sets as shown in Eq. 6:

$$CM_x = \{t^1, t^2, t^3 ... t^k\} \quad (6)$$

Where:

CM = The Text Canonical Model
x = The index for the CM
t = The terms
k = The threshold explained in the previous phase

**Clustering with Canonical Text Model:** In this 2nd level of text clustering, the clustering involves using an appropriate similarity measure to group related documents into the same clusters (Mihalcea, 2004). Text similarity measure will be used to calculate the proximity between the dataset and the text represented in text Canonical Models. Without text Canonical Model, the proximity calculation are exponential since each document need to be compared with each other to find the similarityindex. This can be shown graphically as depicted in Fig. 1.

Figure 1 represents the number of comparison and computation of similarity scores will increase exponential with the increase in the number of documents. The construction of Canonical Data Model in our proposed method will contribute to the text clustering task by alleviating the complexity of the proximity calculation as shown in Fig. 2. With the construction of canonical text model for the documents, the exponential complexity of the proximity calculation between the documents can be the reduced to linear.

## RESULTS AND DISCUSSION

An experiment was performed in order to demonstrate the applicability of our proposed approach. Any unstructured text data can be used in this resaerch for the purpose of assessing the feasibility of the proposed approach. However, to ensure future comparison with other methods, we propose to use dataset that are commonly used in previous research. We evaluated our approach on news articles, i.e., 20 Newsgroup dataset. The data which comprise of a collection of news articles have pre-defined clusters (e.g., comp.graphics, rec.autos, rec.motorcycles, sci. crypt, misc. forsale, talk. politics. mideast etc.,) for evaluation purposes.

After undergone the pre-processing steps, the feature extraction is performed. The result is a sparse matrix of weighted terms. These matrix were used in first round of the text clustering using the chosen clustering algorithm. In this study, we used K-means to cluster the produced term weight features.

We then implemented the LSA on the resulting clusters to identify relevant features that can accurately represent each clusters. Then we take the k largest singular value. In our research, we set k = 10. As mentioned earlier, k represents the number of terms that are recognized as relevant. The selected k terms are then represented as Canonical Text Models using Eq. 6. The produced canonical text models are then used in the 2nd level of text clustering where each of the text documents is compared to each of the Canonical Text Models using cosine similarity measures. We measure the performance of the clustering using F-measures. The F-measure$_1$

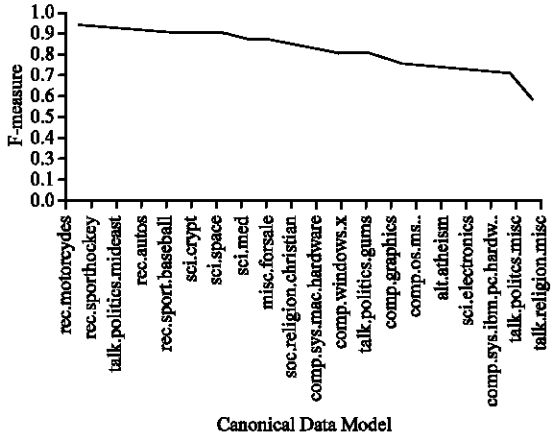


Fig. 3: Similarity measurement using Canonical Data Model

Table 1: Performance of the proposed approach

| Method | F-measure |
|---|---|
| Proposed approach | 0.82 |
| K-means+LSA | 0.75 |

score can be interpreted as a weighted average of the precision and recall where its best value is 1 and its worst value is 0. Figure 3 presents the F-measure for each category of the 20 Newsgroup dataset.

The rec.motorcycles category produces the highest F-measure score of 0.94 while the talk.religion.misc category produces the lowest F-measue score of 0.55. Table 1 presents the average F-measrues for the proposed approach. The produced average F-measure is compared to the F-measure score produced using K-means and LSA without Canonical Text Model creation as reported in Song and Park (2007).

Table 1 shows the proposed approach produces F-measure of 0.82 which is better then the compared method (K-means and LSA without Canonical Text Model).

## CONCLUSION

In this study, we present a text clustering approach by developing Canonical Text Model. We show how the popular weighting measures such as tfidf and LSA can be used as dimensionality reduction techniques in order to extract only important features from text so that the construction of the Canonical Data Model will be less computationally complex. Besides that, we also show how the complexity of proximity calculation between documents can be alleviated using the constructed Canonical Data Model.

Experiments was performed using the 20 newsgroup dataset to evaluate the effectiveness of the proposed text clustering using Canonical Data Models by measuring the F-measures of the produced clusters. The result reveals that the approach produces better scores compared to clustering without the construction of Canonical Text Model.

## RECOMMENDATION

Future research will concentrate on testing the proposed approach on different algorithms and different datasets.

## ACKNOWLEDGEMENT

We would like to thank the Malaysian Ministry of Higher Education for providing the funding for this research through the Exploratory Research Grant Scheme (ERGS).

## REFERENCES

Balabantaray, R.C., C. Sarma and M. Jha, 2013. Document clustering using K-means and K-medoids. Int. J. Knowledge Based Comput. Syst., 1: 7-13.

Berkhin, P., 2002. Survey of Clustering Data Mining Techniques. Accrue Software Inc, San Jose, California, Pages: 56.

Bloechle, J.L., M. Rigamonti, K. Hadjar, D. Lalanne and R. Ingold, 2006. XCDF: A canonical and structured document format. Proceedings of the 7th International Workshop on Document Analysis Systems, February 13-15, 2006, Springer, Nelson, New Zealand, pp: 141-152.

Dietrich, M . and J. L emcke, 2011. A refined Canonical Data Model for multi-schema integration and mapping. Proceedings of the 2011 IEEE 8th International Conference on E-Business Engineering (ICEBE), October 19-21, 2011, IEEE, Beijing, China, ISBN:978-1-4577-1404-7, pp: 105-110.

Gonzalez, M.A., J.R. Marti and P. Kruchten, 2011. A canonical data model for simulator interoperation in a collaborative system for disaster response simulation. Proceedings of the 2011 24th International Canadian Conference on Electrical and Computer Engineering (CCECE), May 8-11, 2011, IEEE, Niagara Falls, Ontario, Canada, ISBN:978-1-4244-9788-1, pp: 001519-001522.

Hachey, B., G. Murray and D. Reitter, 2006. Dimensionality reduction aids term co-occurrence based multi-document summarization. Proceedings of the Workshop on Task-Focused Summarization and Question Answering, July 23-23, 2006, Association for Computational Linguistics, Sydney, Australia, ISBN:1-932432-79-5, pp: 1-7.

Han, J., J. Pei and M. Kamber, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier, Amsterdam, Netherlands, Pages: 702.

Iria, M. and P. Ciravegna, 2005. Relation extraction for mining the Semantic Web. Dagstuhl, Wadern, Germany.

Jing, L., M.K. Ng, J. Xu and J.Z. Huang, 2005. Subspace clustering of text documents with feature weighting K-means algorithm. Proceedings of the 9th International Conference on Pacific-Asia Advances in Knowledge Discovery and Data Mining, May 18-20, 2005, Springer, Hanoi, Vietnam, pp: 802-812.

Luo, C., Y. Li and S.M. Chung, 2009. Text document clustering based on neighbors. Data Knowl. Eng., 68: 1271-1288.

Mihalcea, R., 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Proceedings of the 2004 ACL Conference on Interactive poster and demonstration sessions, July 21-26, 2004, Association for Computational Linguistics, Barcelona, Spain, pp: 181-184.

Narang, T., 2015. Hiearchical clustering of documents: A brief study and implementation in Matlab. Intl. J. Innovations Adv. Comput. Sci. IJIACS., 4: 153-159.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage., 24: 513-523.

Song, W. and S.C. Park, 2007. A novel document clustering model based on latent semantic analysis. Proceedings of the 3rd International Conference on Semantics, Knowledge and Grid, October 29-31, 2007, IEEE, Shan Xi, China, ISBN: 0-7695-3007-9, pp: 539-542.

Yang, S. and Y. Wang, 2009. Density-based clustering of massive short messages using domain ontology. Proceedings of the Conference on Information Processing Asia-Pacific APCIP Vol. 2, July 18-19, 2009, IEEE, Shenzhen, China, ISBN:978-0-7695-3699-6, pp: 505-508.