Short-Text Clustering using Statistical Semantics

Sepideh Seifzadeh
University of Waterloo
Waterloo, Ontario, Canada.
N2L 3G1
sseifzad@uwaterloo.ca

Ahmed K. Farahat
University of Waterloo
Waterloo, Ontario, Canada.
N2L 3G1
afarahat@uwaterloo.ca

Mohamed S. Kamel
University of Waterloo
Waterloo, Ontario, Canada.
N2L 3G1
mkamel@uwaterloo.ca

Fakhri Karray
University of Waterloo
Waterloo, Ontario, Canada.
N2L 3G1
karray@uwaterloo.ca

ABSTRACT

Short documents are typically represented by very sparse vectors, in the space of terms. In this case, traditional techniques for calculating text similarity results in measures which are very close to zero, since documents even the very similar ones have a very few or mostly no terms in common. In order to alleviate this limitation, the representation of short-text segments should be enriched by incorporating information about correlation between terms. In other words, if two short segments do not have any common words, but terms from the first segment appear frequently with terms from the second segment in other documents, this means that these segments are semantically related, and their similarity measure should be high. Towards achieving this goal, we employ a method for enhancing document clustering using statistical semantics. However, the problem of high computation time arises when calculating correlation between all terms. In this work, we propose the selection of a few terms, and using these terms with the Nyström method to approximate the term-term correlation matrix. The selection of the terms for the Nyström method is performed by randomly sampling terms with probabilities proportional to the lengths of their vectors in the document space. This allows more important terms to have more influence on the approximation of the term-term correlation matrix and accordingly achieves better accuracy.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.7.2 [**Document and Text Processing**]: Document Preparation

Keywords

Short-term clustering; Semantic similarity; Nyström approximation.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742474.

1. INTRODUCTION

In social media, users usually post short texts. Twitter limits the length of each Tweet to 140 characters; therefore, developing data mining techniques to handle the large volume of short texts has become an important goal [1]. Text document clustering has been widely used to organize document databases and discover similarity and topics among documents. Short text clustering is more challenging than regular text clustering; due to the sparsity and noise, they provide very few contextual clues for applying traditional data mining techniques [2]; therefore, short documents require different or more adapted approaches. The representation of short-text segments needs to get enriched by incorporating information about correlation between terms.

The most common document representation model is vector space model (VSM) introduced by Salton et al. [3], which assumes that terms are independent and ignores the semantic relation among terms. The novelty of the VSM is to use frequencies in a corpus of text as a clue for discovering semantic information. The idea of the VSM is to represent each document in a collection as a point in a space (a vector in a vector space). Points that are closer in this space are semantically similar and points that are far apart are semantically distant.In Generalized Vector Space Model (GVSM) [4], the correlation between terms is estimated as inner product (un-normalized association matrix) or cosine similarities (normalized association matrix) of term vectors in dual space; Covariance matrix between terms, and the matrix of Pearson's correlation coefficients can also be used to estimate the Gram matrix of term vectors that encodes measures of statistical correlations between terms [5].

In this paper, the correlation between terms is used; however, in order to deal with the problem of high computation time arises when calculating correlation between all terms. We use Nyström approximation [6] using few terms to approximate the whole term-term correlation matrix. The selection of the terms for the Nyström method is performed by randomly sampling terms with probabilities proportional to the lengths of their vectors in the document space. This allows more important terms to have more influence on the approximation.

The rest of the paper is organized as follows. An overview of the related literature work and background is presented in Section 2. Section 3 provides the description of the pro-

posed method. Section 4 covers the experimental results and analysis. Finally, concluding remarks are presented in Section 5.

2. RELATED WORK

For mapping from the unclassified text to the given categories, short text classification requires sufficient number of training examples to achieve the high accuracy; therefore, using a totally unsupervised technique for short text grouping is more efficient. An overview of the approaches in the literature to tackle the problem of short text clustering is provided in this section.

One possible approach is modifying the term weighting technique. As different terms have different importance in a given document, a term-weight is normally associated with every term. In order to enable an effective clustering process, the word frequencies need to be normalized in terms of their relative frequency of presence in the document and over the entire collection. These weights are often derived from the frequency of terms within a document or a set of documents. The idea of weighting is to give more weight to the terms of higher importance; the most popular way to formalize this idea for term-document matrices is the vector-space based *tf-idf* [7], [8] which is a simple and efficient representation. In *tf-idf* representation, the term frequency for each word is normalized by the inverse document frequency to reduce the weight of terms which occur more frequently in the collection. In addition, a sub-linear transformation function is often applied to the term frequencies in order to avoid the undesirable dominating effect of any single term that might be very frequent in a document.

A variety of methods have been proposed to extend term weighting techniques to work on short documents. First, Yan et al. [9] have proposed an alternative technique to calculate the term weighting in short documents. They have mentioned that for short texts using *tf-idf* is not very efficient since term frequency in all documents is not a good measure to capture the discriminative power of the data due to the sparsity. Despite from *tf-idf*, their method measures term discriminability by term level instead of document level. Weights are derived from well-known normalized cut (Ncut) method [10]. They first consider clustering the terms by applying Ncut to the graph model in which the nodes represent terms and the edges represent the correlation between terms which defines number of co-occurrence of two terms connected with that particular edge; the graph is then partitioned using Ncut method. For experiments they have only considered words with document frequency more than 6 and documents containing more than 4 words.

Another approach to handle short documents, is using explicit and external semantics by incorporating some external knowledge such as introducing external corpus which enables using external semantics. Ferragina et al. [11] proposed a simple and fast method for entity disambiguation (entity linking) for short texts using Wikipedia. Hu et al. [12] exploited features from Wikipedia for clustering of short texts. Using Wikipedia and ontology-based techniques is computationally complex; another approach to enhance the clustering is to use explicit internal semantics, such as term-term similarity [5]. One possible approach is to extend the features. Estimating the relation between terms takes advantage of co-occurrence information, which is based on the assumption that two terms are similar if they frequently co-occur in the same document. A term can then be represented by a term co-occurrence vector, rather than the document vector. Both co-occurrence and dependency of terms is considered. The weight of co-occurrence between each pair of terms is calculated, and then based on this measure two terms are decided to be in the same cluster or not.

Correlation matrix for short texts is very sparse (mostly zero) as a lot of words do not appear in each individual document and the respective element of the matrix is zero, and due to sparsity, the performance of clustering algorithms will decrease dramatically, as stated in [13]. In some works SVD is used as a way of simulating the missing text and as stated by [14] it is a way of compensating for the missing data, it also enables us to handle a large-scale data. In [15], it is also mentioned that a low rank approximation of the matrix inspired by [16]- [18] is used, in order to achieve the best rank-k approximation of the data matrix.

3. PROPOSED METHOD

Finding correlation among terms is important especially in short text clustering where limited knowledge is available. If using external semantics source, such as Wikipedia, is not considered, one should then focus on creating the semantic relation based on the internal semantics; e.g. statistical semantics, which is a measure of similarity between units of text (terms) and is evaluated based on the statistical analysis of term occurrence patterns [5], correlation between terms.

Short-text segments are typically represented by very sparse vectors, where each has non-zero weights for only a very few terms. In this case, traditional techniques for calculating text similarity results in measures which are very close to zeros. This is due to the fact that documents, even the very similar ones, do have a very few or mostly no terms in common. In order to alleviate this limitation, we need to enriched the representation of short-text segments by incorporating information about correlation between terms. In other words, if two short segments do not have any common words, but terms from the first segment appear frequently with terms from the second segment in other documents, this means that these segments are semantically related, and their similarity measure should be high. Towards achieving this goal, we employ a method for enhancing document clustering using statistical semantics proposed by Farahat and Kamel [5] and enhance it to handle large amounts of short documents.

Let X be an $m \times n$ matrix whose element X_{ij} represents the weight of term i inside document j. Farahat and Kamel [5] proposed the use of a document similarity kernel based on term-term correlation as:

$$K = X^T G X,$$

where G is $m \times m$ is a term-term correlation matrix represented that could be calculated either using association between terms (assc), normalized association between terms (asscn), covariance measures (cov) or Pearson's correlation coefficients (pcor). Using the data matrix X to calculate G results in the following matrices:

$$G_{\text{ASSC}} = XX^T$$

$$G_{\text{ASSC\_N}} = L_X^{-1/2} X X^T L_X^{-1/2}$$

$$G_{\text{COV}} = \frac{1}{n-1}\bar{X}\bar{X}^T$$

$$G_{\text{PCOR}} = \frac{1}{n-1}L_{\bar{X}}^{-1/2}\bar{X}\bar{X}^T L_{\bar{X}}^{-1/2}$$

where $\bar{X} = XH$ is the matrix that is obtained from X by centering its columns, $H = I - \frac{1}{n}ee^T$ is an $n \times n$ centering matrix, and e is the all-ones vector of size n. The matrices L_X and $L_{\bar{X}}$ are diagonal matrices whose diagonal elements are the lengths of the columns of X and $\bar{X}$ respectively. Farahat and Kamel [5] suggested the factorization of K into W^TW, where $W = X^TZ$ and Z are determined based on the term-term correlation matrix used such that $G = ZZ^T$. The vector-based clustering algorithms can then be performed on W instead of X and the similarity-based clustering clustering can be performed on the semantic kernel $K = W^TW$ instead of the kernel calculated based on X (e.g., $K = X^TX$ for linear kernels).

The main limitation of using these semantic kernels is that if all documents are used to estimate term-term correlations, the time and space complexities of the algorithm to calculate the semantic representation of documents is $O(n^2)$. In order to alleviate this problem, Farahat and Kamel suggested the use of a fewer documents to estimate these correlations. This solution considerably reduces the computational complexity and it works well in practice for long documents. In the case of short documents, the number of terms per document is very small and it will be required to select almost all documents in order to include all the terms in the corpus. In this case, it is required to develop a more elegant way to approximate the term-term correlation matrix.

We propose the selection of a few terms, rather than documents, and then use these terms with the Nyström method [6] to approximate the term-term correlation matrix X. The proposed method works as follow. Let G be the term-term correlation matrix, $\mathcal{S}$ be the set of selected terms, and $\mathcal{R}$ be the set of remaining terms. The Nyström method approximates the matrix G using the subset $\mathcal{S}$ as:

$$\tilde{G} = G_{:\mathcal{S}}G_{\mathcal{S}\mathcal{S}}^{-1}G_{:\mathcal{S}}^T$$

The approximation of G can be factorized as ZZ^T, where $Z = G_{:\mathcal{S}}G_{\mathcal{S}\mathcal{S}}^{-1/2}$. For different measures of term-term correlation, the semantic representation matrix W can be calculated as:

$$W_{\text{ASSC}} = \left(X_{\mathcal{S}:}X_{\mathcal{S}:}^T\right)^{-1/2}X_{\mathcal{S}:}X^TX,$$

$$W_{\text{ASSC\_N}} = \left(X_{\mathcal{S}:}X_{\mathcal{S}:}^T\right)^{-1/2}X_{\mathcal{S}:}X^TL_X^{-1/2}X,$$

$$W_{\text{COV}} = \frac{1}{\sqrt{n-1}}\left(X_{\mathcal{S}:}HX_{\mathcal{S}:}^T\right)^{-1/2}X_{\mathcal{S}:}HX^TX, \quad \text{and}$$

$$W_{\text{PCOR}} = \frac{1}{\sqrt{n-1}}\left(X_{\mathcal{S}:}HX_{\mathcal{S}:}^T\right)^{-1/2}X_{\mathcal{S}:}HX^TL_{\bar{X}}^{-1/2}X.$$

Note that $H = HH$ as H is a projection matrix.

The aforementioned formulas for calculating the semantic representations can be implemented in an efficient manner. First, the $G_{\mathcal{S}\mathcal{S}}$ matrices are $\ell \times \ell$ matrices where $\ell \ll m$ is the number of selected terms. The square root of the inverse of these matrices $G_{\mathcal{S}\mathcal{S}}^{-1/2}$ can be easily calculated by computing the eigen decomposition of $G_{\mathcal{S}\mathcal{S}} = U\Lambda U^T$ and then taking the leading $k \leq \ell$ eigen values and vectors. By taking the inverse of the square root of the eigen values, $G_{\mathcal{S}\mathcal{S}}^{-1/2} = U\Lambda^{-1/2}U^T$ can be easily calculated. The other part of formulas can easily calculated and stored by either calculating $X_{\mathcal{S}:}X^T$ or $X_{\mathcal{S}:}HX^T$ which are both $l \times m$. The matrix multiplications with H can be further improved by distributing the multiplied matrices across I and $\frac{1}{n}ee^T$ (where $H = I - \frac{1}{n}ee^T$).

In the case when $k \leq \ell$ eigen-values and vectors are used, the semantic representation can be directly expressed in the space of the leading k eigen vectors, e.g., $W_{\text{ASSC}} = \Lambda^{-1/2}U^TX_{\mathcal{S}:}X^TX$. This is equivalent to using $G_{\mathcal{S}\mathcal{S}}^{-1/2}$ as W^TW in both cases will be equal. This reduces the dimension of the semantic space to k and makes the clustering algorithm more efficient.

The selection of the terms for the Nyström method is performed by randomly sampling terms with probabilities proportional to the lengths of their vectors in the document space. This allows more important terms to have more influence on the approximation of the term-term correlation matrix and accordingly achieves better accuracy.

4. EXPERIMENTS

4.1 Experimental Setup

Data Sets.

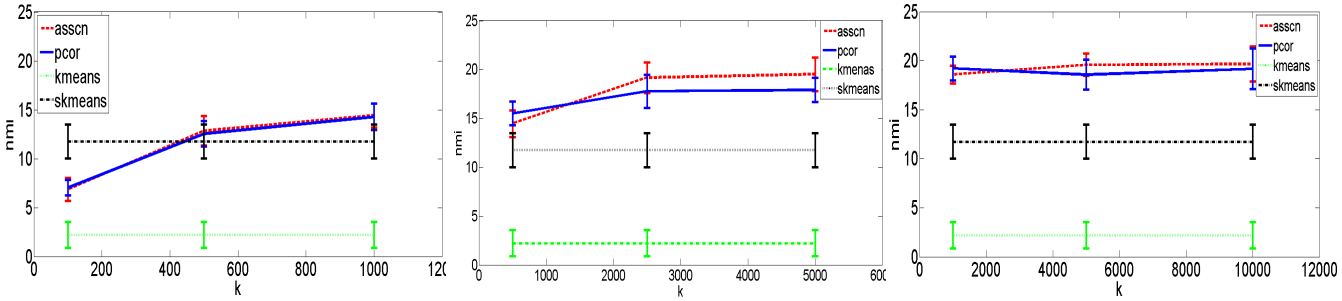
In order to evaluate the proposed approach, we used a dataset of a large number of labeled tweets which was collected and labeled by Zubiaga *et. al* [19]. They have collected tweets that contain a URL to a web page, and each tweet was automatically labeled using the content of the page which the URL refers to, they have used the categories of the Open Directory Project (ODP) as their label set. This dataset contains 10 different categories (classes) and totally around 360k labeled tweets. The data used for the experiments are unbalanced random samples selected from this datasets with different sizes.

In the first category of experiments, two 10k datasets are sampled and a comparative study has been conducted to compare the Normalized Mutual Information (NMI) [20] and computation time when using the different approaches. In the second category of experiments, five datasets with different sizes, 10k, 25k, 50k, 75k and 100k are sampled to evaluate the performance of different methods with regards to different dataset sizes. The term-document matrix is built using *tf-idf* after stemming and stop words removal of the raw data.

Baseline Methods.

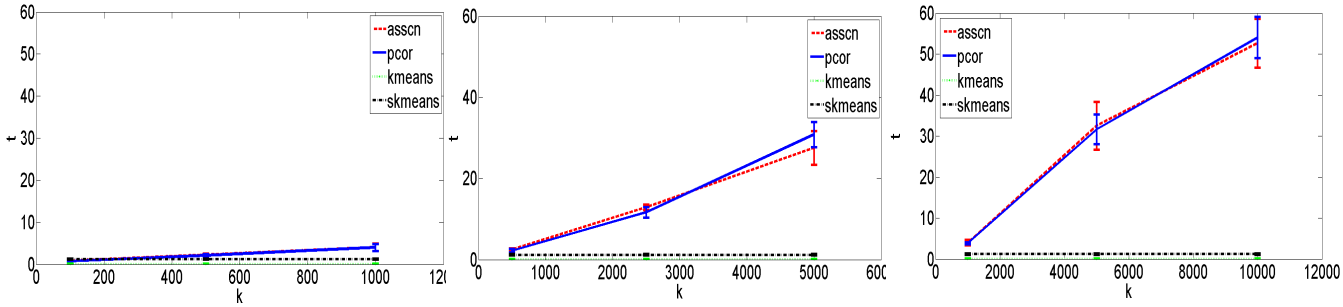
The baseline methods that are employed for the comparison are as follows:

- K-means with terms weighted using *tf-idf*
- Spherical K-means with *tf-idf* [21, 22], which uses the cosine similarity between documents, that considers the angle between them, not the length of the vectors; therefore, it is independent of the document's length.
- Non-negative Matrix Factorization (NMF) with terms weighted using *tf-idf* and Ncut approach (Ncut+NMF) [9]



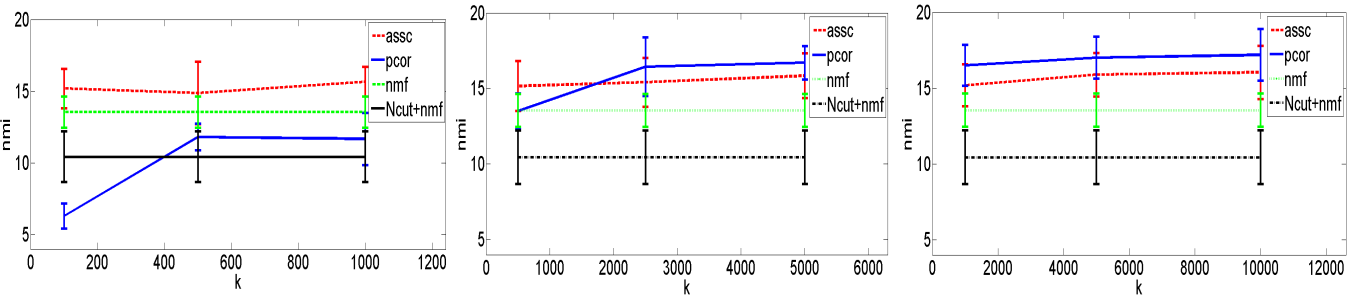
(a) ℓ=1000 (b) ℓ=5000 (c) ℓ=10000

Figure 1. Comparison of NMI using Spherical K-means



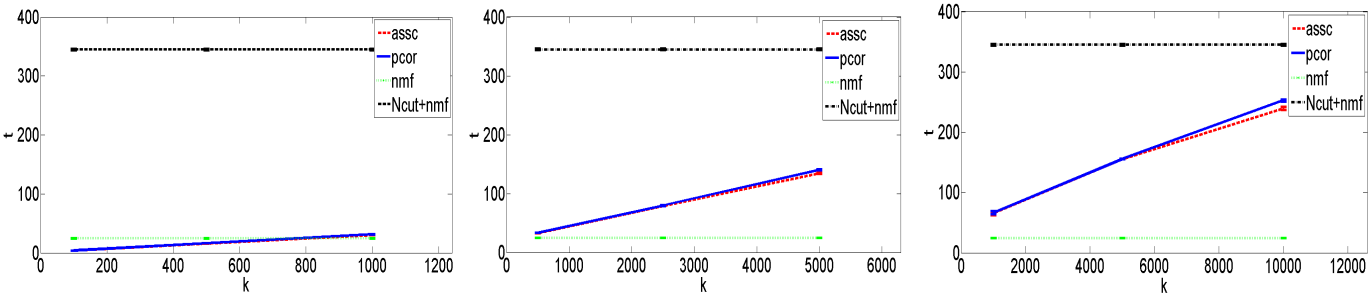
(a) ℓ=1000 (b) ℓ=5000 (c) ℓ=10000

Figure 2. Comparison of computation time using Spherical K-means



(a) ℓ=1000 (b) ℓ=5000 (c) ℓ=10000

Figure 3. Comparison of NMI using NMF



(a) ℓ=1000 (b) ℓ=5000 (c) ℓ=10000

Figure 4. Comparison of computation time using NMF

4.2 Results and Analysis

Category I.

In this category of experiments, we have conducted experiments for a 10k dataset and averaged over 10 trials for different methods. In our approach, term-term correlation matrix could be calculated either using association between terms (assc), normalized association between terms (asscn), covariance measures (cov) or Pearson's correlation coefficients (pcor), which all encode measures of statistical correlations between terms described in details in [5]. In order to demonstrate the results and compare our approach with baseline methods, we have selected the best two methods for calculating the term-term correlation for our approach shown in Figure 1 to Figure 4.

In order to compare the proposed approach, K-means, spherical K-means (Spherical K-means) and Non-negative Matrix Factorization (NMF) are the clustering methods. When using NMF two different weighting methods are used, *tf-idf* and Ncut which is proposed by Yan *et al.* [9] and is using term-term correlation for short-text clustering. In this set of experiments NMI and computational time are compared. As mentioned by Yan *et al.*, in the experiment section of their paper, there is a pre-processing for selecting the documents with more than a specific number of terms and the terms that have frequency more than a specific threshold. However, we have not applied this pre-processing, we used all documents and terms for all methods in order to be able to do the comparison.

Our approach consists of two steps as discussed in the previous section; after selecting a subset of terms (ℓ), the term-term correlation matrix is evaluated and then rank-k approximation of the matrix is used for clustering. In this category of experiments, we study the effect of using different number of terms and different ranks on NMI and computation time.

Figure 1 shows the comparison of different approaches, K-means, Spherical K-means, and two cases of the proposed approach in which pcor and asscn are used to measure the term-term similarity. As shown in the figures using low rank approximation of the original term-term matrix, our proposed approach can achieve higher NMI using very low number of terms and low rank approximation of the similarity matrix. As shown in this figure, using 5k terms and rank-5, the achieved NMI is the same as using the full rank matrix of term-term correlation and this results is approximately 7% higher than when using Spherical K-means and 18% higher than when using K-means with almost the same computation time as shown in Figure 2.

Non-negative Matrix Factorization is also used as the clustering approach and for pre-processing of the data *tf-idf* and Ncut are used as term weighting approaches. As can be seen from the results presented in Figures 3 and 4, which show the comparison in terms of NMI and computation time for the different methods, using our proposed approach with full rank matrix and Ncut weighting approach both require high computation time, since they both calculate the similarity measure between all terms. Ncut calculates the similarity based on inner product of terms for all terms. Using NMF with either *tf-idf* and Ncut results in almost same NMI. The effectiveness of our method is that by using lower rank and fewer terms we can achieve the best NMI with much lower computation time in comparison with Ncut+NMF and *tf-idf*+NMF.

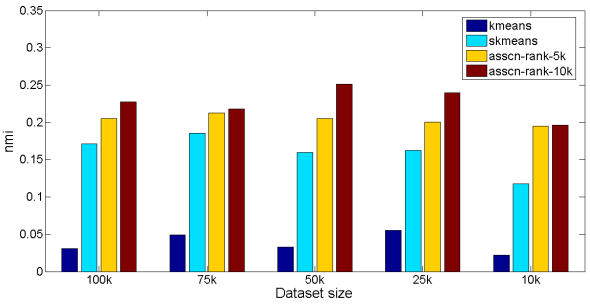


Figure 5. Comparison of NMI w.r.t different dataset sizes

In our approach, as the rank and the number of selected terms decrease the computation time decreases drastically; however, NMI keeps improving even when the lower rank approximation is used. We have shown that using lower rank and less number of terms, the clustering results achieved are competitive with other approaches and the computation time increases as the number of rank increase. However same accuracy can be achieved using low rank approximation which requires less computation time.

Category II.

In this category of experiments NMI is evaluated using different methods and for datasets with different sizes 10k, 25k, 50k, 75k and 100k. As shown in Figure 5, our method outperforms other methods even in the case in which dataset size is 100k; our method achieves higher NMI using only 10% of terms ($\ell = 10$). The baseline methods are K-means and Spherical K-means and the clustering method used for our approach is Spherical K-means. When using NMF+Ncut as the clustering method, it requires a huge computation time and is not feasible since it computes the similarity measure for all terms; therefore, when using very large datasets we could not perform the comparison. When using huge dataset it is not feasible to calculate the whole term-term correlation matrix; since it takes a huge amount of time. Therefore in this category of experiments we only focus on K-means and Spherical K-means as baselines. As for our proposed approach we calculate up to rank 10k for huge dataset, if the dataset is very large, for example 100k, better results could be achieved by using higher rank for the approximation; however, using low rank approximation which helps to reduce the computational time extensively, results in acceptably higher NMI than the baseline methods.

The results achieved using our method, highly depends on the terms selected; as shown in Figure 5, NMI for rank-5k and rank-10k for dataset of size 10k are almost the same; and in both cases our method outperforms the baseline approaches. When using rank-5k, it means that 5000 words are selected (ℓ=5000) and the rank used for the approximation is also 5000 (k=5000).

5. CONCLUSION

In this work, the effectiveness of short-text document clustering algorithms has been improved. As shown in the experimental results, our proposed approach outperforms the baseline methods by incorporating information about similarity measure based on statistical correlations between terms. We have proposed a technique to select a subset of terms

and then using the selected terms along with the Nyström approximation, it can obtain a low rank approximation of the term-term correlation matrix in order to alleviate the problem of high computation time.

Different techniques for calculating the semantic similarity measure are evaluated on term-term correlations. These methods are discussed and their performance with three different clustering algorithms is evaluated. A low-dimension representation for documents is calculated based on random sampling of terms. Experiments show that using this random sampling and low rank approximation of the matrix, considerably reduces the run-time while maintaining much of the improvement achieved by the semantic similarity models in the document clustering task.

6. ACKNOWLEDGMENTS

This publication was made possible by a grant from the Qatar National Research Fund through National Priority Research Program (NPRP) No. 06-1220-1-233. Its contents are solely the responsibility of the authors.

7. REFERENCES

[1] K. Verma, M. K. Jadon, and A. K. Pujari, "Clustering Short-Text Using Non-negative Matrix Factorization of Hadamard Product of Similarities," *Information Retrieval Technology Lecture Notes in Computer Science*, Volume 8281, pages 145–155, 2013.

[2] Z. Faguo, Z. Fan, Y. Bingru, Y. Xingang, "Research on Short Text Classification Algorithm Based on Statistics and Rules," *In proceedings of third International Symposium on Electronic Commerce and Security*, pages 3–7, 2010.

[3] G. Salton, A. Wong, C. S. Yang, "A vector space model for automatic indexing," *Magazine Communications of the ACM*, Volume 18, Issue 11, pages 613–620, Nov. 1975.

[4] S. Wong, w. Ziark, P. Wong, "Generalized vector spaces model in information retrieval," *In Proceedings of the eighth annual international ACM SIGIR conference on research and development in information retrieval. ACM,* New York, pages 18âĂŞ-25, 1985.

[5] Ahmed K. Farahat, Mohamed S. Kamel, "Statistical semantics for enhancing document clustering," *Knowledge and Information Systems*, Volume 28, Issue 2, pages 365–393, 2010.

[6] Kumar, Sanjiv, Mehryar Mohri, and Ameet Talwalkar, "Sampling techniques for the Nyström method," *In International Conference on Artificial Intelligence and Statistics*, pages 304–311. 2009.

[7] G. Salton, "An Introduction to Modern Information Retrieval," *Mc Graw Hill*, 1983.

[8] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pages 11–21, 1972.

[9] X. Yan, J. Guo, Sh. Liu, X. Cheng, Y. Wang, "Clustering Short Text Using Ncut weighted Non-negative Matrix Factorization," *CIKM 12 Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2259–2262, 2012.

[10] J. Shi, J. Malik, "Normalized cuts and image segmentation," *IEEE Trans PAMI*, 22(8), pages 888–905, 2000.

[11] P. Ferragina, U. Scaiella, "Fast and Accurate Annotation of Short Texts with Wikipedia Pages," *Software, IEEE*, Volume:29, Issue: 1, 2011.

[12] X. Hu, N. Sun, C. Zhang, T. Chua, " Exploiting internal and external semantics for the clustering of Short texts using world knowledge," In Proc. CIKM Hong Kong, China, pages 919–928, Nov. 2009.

[13] P. Lin, Z. Lin, B. Kuang, P. Huang, "A Short Chinese Text Incremental Clustering Algorithm Based on Weighted Semantics and Naive Bayes," *Journal of Computational Information Systems*, 8(10), pages 4257-âĂŞ4268, 2012.

[14] E. Vozalis, K. Margaritis, "Analysis of recommender systems algorithms," *In Proceedings of the 6th Hellenic European Conference on Computer Mathematics and its Applications*, Athens, Greece, 2003.

[15] Sindhwani, V., Thomas J., Yorktown Heights, Ghoting, Ting, Lawrence, "Extracting insights from social media with large-scale matrix approximations," *IBM Journal of Research and Development*, Vol. 55 , Issue: 5, 2011.

[16] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, "Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation," *John Wiley & Sons Ltd: Chichester*, UK, 2009.

[17] N. D. Ho, P. V. Dooren, and V. D. Blondel, "Descent methods for non-negative matrix factorization, in Numerical Linear Algebra in Signals," *Systems and Control*, 2007.

[18] N. Gillis, "Nonnegative matrix factorization: Complexity, algorithms and applications," *M.S. thesis, Univ. Catholique de Louvain, Louvain-la-Neuve*, Belgium, 2011.

[19] Zubiaga, Arkaitz, and Heng Ji, "Harnessing web page directories for largescale classification of tweets," *Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee*, pages 225–226, 2013.

[20] Nathan D. Cahill, "Normalized measures of mutual information with general definitions of entropy for multimodal image registration," *In Proceedings of the 4th international conference on Biomedical image registration*, pages 258-âĂŞ268, Berlin, Heidelberg, 2010.

[21] S. Zhong, "Efficient Online Sphercal K-means Clustering," *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, Montreal, Canada, 2005.

[22] G. Pant, K. Tsioutsiouliklis, J. Johnson, C. L. Giles, "Panorama: extending digital libraries with topical crawlers", *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 142–150, 2004.