Clustering Narrow-Domain Short Texts Using K-Means, Linguistic Patterns and LSI

Svetlana Popova[1,2(✉)], Vera Danilova[3], and Artem Egorov[2]

[1] Saint-Petersburg State University, Saint Petersburg, Russia
svp@list.ru
http://spbu.ru
[2] ITMO University, Saint-Petersburg, Russia
http://www.ifmo.ru/
[3] Autonomous University of Barcelona, Barcelona, Spain
maolve@gmail.com
http://www.uab.cat/letras/

Abstract. In the present work we consider the problem of narrow-domain clustering of short texts, such as academic abstracts. Our main objective is to check whether it is possible to improve the quality of k-means algorithm expanding the feature space by adding a dictionary of word groups that were selected from texts on the basis of a fixed set of patterns. Also, we check the possibility to increase the quality of clustering by mapping the feature spaces to a semantic space with a lower dimensionality using Latent Semantic Indexing (LSI). The results allow us to assume that the aforementioned modifications are feasible in practical terms as compared to the use of k-means in the feature space defined only by the main dictionary of the corpus.

Keywords: Clustering · Short texts · Narrow domain texts · LSI · Linguistic patterns

1 Introduction

The task of short-text processing is important due to the increase of Internet content, such as news summaries, abstracts, forum messages, social networks, twitter etc. Clustering allows to obtain a structured representation of collections with automatic grouping of topically close documents.

We are interested in clustering academic papers. The solution to this task is required for structured representation of data within scientific domains, e.g. in academic search engines [1–4].

Abstracts are accessed from e-libraries. Many of these libraries usually provide free access to abstracts, while full papers require subscription or a payment.

D.I. Ignatov et al. (Eds.): AIST 2014, CCIS 436, pp. 183–189, 2014.
DOI: 10.1007/978-3-319-12580-0_18

However, abstracts are brief summaries of the contents and are deemed sufficient for clustering academic papers with adequate results [5].

The task of clustering abstracts is closely related to a number of problems, e.g., the task of identifying whether short texts and processed texts may belong to the same topic, which requires the use of an approach to narrow-domain short text clustering [6–10]. One of the main problems with this type of clustering is high data sparseness [8]. If the documents come from the same source it is probable that all of them are topically close. In this case there may be a significant overlap of common words, which complicates the clustering task even more [8]. The corpus for testing and analysis of algorithm performance includes a set of collections that are widely used in the field (CICling, SEPLN-CICling, EasyAbstracts) [5–10]. Experiments implemented within the framework of the present work are also based on these collections.

K-means has been chosen for testing. In the related work, k-means is considered the base or one of the base algorithms for comparison [6,7]. We have not found any papers describing the attempts to improve the quality of k-means performance in narrow-domain short-text clustering, researchers tend to use the base version. In [6,7] it was shown that considering document clustering as an optimization problem ensures high performance. These algorithms yield better results as compared to the base version of k-means. In [12] the advantage of combining LSI with the optimization algorithm is presented, which shows the benefit from the use of LSI in short text clustering (together with algorithms based on particle swarm optimization [6]). We study how to increase the performance of k-means by applying LSI and also by extending the feature space with word groups. The word groups extraction uses linguistic patterns (e.g., NN_NN, NN_NN_NN, JJ_NN, NNS_NN, NN_NNS, JJ_NN_NN, etc., where NN denotes Noun, singular o mass, NNS - Noun, plural, JJ - Adjective; examples of extracted phrases: cluster algorithm, binary vector, maximum entropy model method, etc. that were built on the basis of collections labeled for key phrase extraction tasks.

2 Objectives and Data

2.1 Research Tasks

For the purposes of the present study several tasks were formulated. Firstly, we check whether it is possible to improve the quality of narrow-domain short-text clustering using k-means together with LSI [12]. Secondly, we check the possibility to increase the quality of k-means algorithm performance by using patterns. Thirdly, whether it is possible to improve the clustering quality through the application of both patterns and LSI.

2.2 Testing Dataset

The test set was formed from three collections[1] that are often used for testing narrow-domain short text clustering algorithms. All of them are sets of abstracts

[1] http://sites.google.com/site/merrecalde/resources

divided between four clusters by the expert. For each collection the "gold standard" or the best variant of grouping is known. Each collection includes 48 documents. CICling 2002 is considered one of the most difficult collections for clustering [6–8].

2.3 Pattern Extraction

For the purposes of pattern extraction we employed one of the most widely used collections in the field of keyword extraction - SemEval 2010, which was used in the competition of algorithms for keyword extraction TREC 2010 [13]. The collection includes documents and keywords defined by the expert characterizing each document. We used 23 patterns based on the most frequent keywords determined by the expert. We relied on the assumption that, in this way, we will be able to select patterns that are specific to the keywords in texts and also word sequences that reflect textual semantics. We assumed that the use of such sequences would increase clustering quality.

2.4 Clustering

`Data Pre-processing.` Word stems were used for text representation during clustering. Stemming was performed with Porter Stemmer[2]. Stop-words were removed using a standard list. PoS tags were assigned to each word for the pattern-based extraction of phrases using the Stanford PoS tagger.[3] `Clustering Algorithm.` K-means algorithm was chosen for testing in the present work. Each document is represented as a vector in the feature space defined by the main dictionary or within the feature space formed by the extended dictionary, which includes word groups extracted using patterns. The word order was taken into account during the retrieval of word groups. Upon the extraction each word group was transformed into another group (a set of words). The calculation of term/word group weight was performed using TF-IDF, the distance was measured by the cosine similarity between feature vectors. The number of clusters was the same as in the gold standard, which means that there were 4 clusters in each case. LSI maps the feature space to a semantic space of a lower dimensionality using singular value decomposition of the text-attributes matrix. As a result, each document has a vector representation in the semantic space. To perform LSI we need to indicate the dimensionality of the space where the data are mapped.

[2] http://tartarus.org/martin/PorterStemmer/
[3] http://nlp.stanford.edu/software/tagger.shtml

2.5 Evaluation

Clustering quality evaluation was performed in a classic way using the combined information on Precision and Recall of the resulting clusters [14,15]:

$$F = \sum_i \frac{G_i}{|D|} \max_j F1_{ij}, where\ F1_{ij} = 2 \times \frac{Precision_{ij} Recall_{ij}}{Precision_{ij} + Recall_{ij}},$$

$$Precision_{ij} = \frac{|G_i \bigcap C_j|}{|G_i|}, Recall_{ij} = \frac{|G_i \bigcap C_j|}{|C_j|}$$

$G = \{G_i\}_{i=\overline{1,m}}$ - clusters generated by the algorithm, $C = \{C_j\}_{j=\overline{1,n}}$ - clusters identified by the experts, D - number of documents in the collection.

3 Experiments and Results

In the course of the experiments two feature space settings were compared: the first was based on the main dictionary of the collection (main dict.) and the second - on the combination of the main dictionary and that of word groups obtained through pattern-based extraction from texts (ext.dict.). For each of the variants we compared the performance of k-means before and after applying the LSI. For each experiment there were 500 iterations on the basis of which the best (max), the worst (min) and the average (avg) results were selected. The results are presented in the Table 1. The values of semantic space dimensionality ("num") that yield the best results are shown.

The analysis of Table 1 allows us to put forward the following assumptions. Firstly, in case of deploying k-means algorithm based on the extended dictionary, no improvement (or a small one) in clustering quality is observed.

Table 1. K-means performance before and after applying LSI ("num" stands for semantic space dimensionality value; "main. dict." indicates all cases where the feature space defined by the main dictionary of the collection was used; "ext. dict." indicates cases where an extended feature space was applied)

	CICling			SEPLN-CICling			EasyAbstracts		
	avg	min	max	avg	min	max	avg	min	max
	with LSI								
Main dict.	0.48	0.35	0.65	0.58	0.36	0.79	0.58	0.36	0.81
Ext. dict	0.49	0.35	**0.66**	0.57	0.35	0.81	**0.59**	0.35	0.86
	with LSI								
Main dict.	**0.53**	**0.48**	0.56	**0.75**	**0.58**	**0.84**	0.51	0.42	0.60
Num (main dict.)	6			3			6		
Ext. dict.	0.49	0.42	0.53	0.74	0.53	0.80	0.49	**0.43**	0.61
Num (Ext. dict.)	7			4			12		

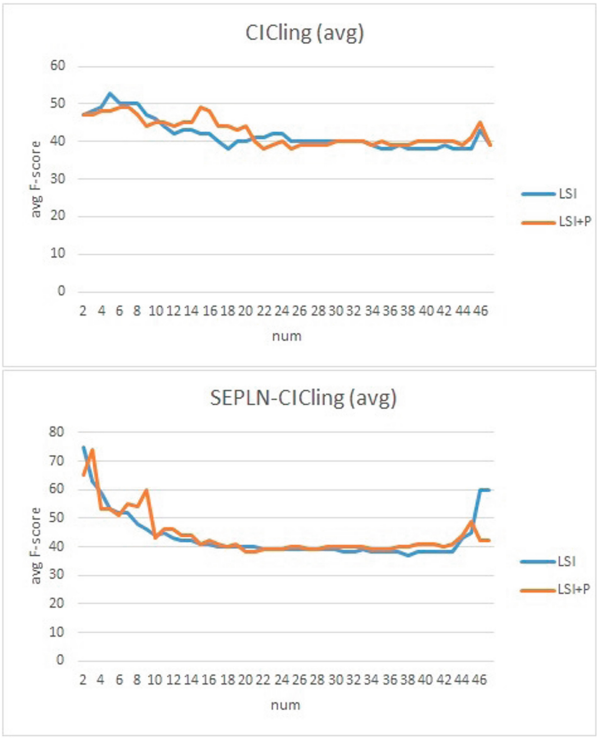

Fig. 1. Dependency of k-means performance on the dimensionality of the semantic space created using LSI (CICling and SEPLN-CICling)

The application of LSI to this feature space does not produce better results than when we use the main dictionary alone. It turns out that in case word groups are extracted using patterns, the results contain much noise (common phrases, such as "experiment results").

Secondly, if the feature space is not expanded with word groups, its mapping to the semantic space with the optimal dimensionality value may lead to some increase in clustering quality. In order to check the feasibility of searching the necessary dimensionality value we modeled the dynamics of conditional dependency between the quality of k-means performance and the dimensionality value.

Figure 1 presents the dependency of the average quality of k-means clustering, according to the results of 500 iterations, on the dimensionality of the semantic space created using LSI (for CICling and SEPLN-CICling). Notation: `LSI`, if the feature space is based on the main dictionary alone; `LSI+P`, if the dictionary of word groups is added. According to the diagrams, applying LSI representation, the best result on the narrow-domain short texts can be obtained if the dimensionality of the resulting semantic space is less than 10 ($num < 10$). Also, the diagrams show that when applying LSI most results are in the range from

0.40 to 0.50 (F). For CICling collection the maximum value of this range causes only insignificant increase in quality value obtained before using LSI. The real effectiveness of LSI application can be observed only for pre-defined values of space dimensionality (CICling - 6, SEPLN-CICling - 3). If these values cannot be identified a priori, the use of LSI is not feasible.

4 Conclusion

The present paper considers the problem of clustering narrow-domain short texts such as abstracts to academic papers. The purpose is to check the possibility of improving k-means performance on such collections using a feature space expanded with the dictionary of pattern-extracted word groups. In addition, we examined the possibility to increase clustering quality by applying LSI to project the feature spaces onto a semantic space of lower dimensionality. The results allow us to assume that the indicated modifications cannot be deemed feasible in practical terms (except when the optimal dimensionality value can be determined) as compared to the use of the simple k-means algorithm and the feature space defined by the main dictionary of the corpus.

Acknowledgement. This work was partially financially supported by the Government of Russian Federation, Grant 074-U01.

References

1. Bernardini, A., Carpineto, C.: Full-subtopic retrieval with keyphrase-based search results clustering. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, vol. 1 (2009)
2. Zhang, D., Dong, Y.: Semantic, hierarchical, online clustering of Web search results. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 69–78. Springer, Heidelberg (2004)
3. Zeng, HJ., He, QC., Chen, Zh., Ma, WY., Ma, J.: Learning to cluster web search results. In: Proceeding SIGIR '04 Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 210–217 (2004)
4. Popova, S., Khodyrev, I., Egorov, A., Logvin, S., Gulyaev, S., Karpova, M., Mouromtsev, D.: Sci-search: academic search and analysis system based on keyphrases. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2013. CCIS, vol. 394, pp. 281–288. Springer, Heidelberg (2013)
5. Alexandrov, M., Gelbukh, A., Rosso, P.: An approach to clustering abstracts. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 275–285. Springer, Heidelberg (2005)
6. Cagnina, L., Errecalde, M., Ingaramo, D., Rosso, P.: A discrete particle swarm optimizer for clustering short text corpora. In: BIOMA08, p. 93103 (2008)
7. Errecalde, M., Ingaramo, D., Rosso, P.: ITSA: an effective iterative method for short-text clustering tasks. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) IEA/AIE 2010. LNCS, vol. 6096, pp. 550–559. Springer, Heidelberg (2010)

8. Pinto, D.: Analysis of narrow-domain short texts clustering. In: Research report for Diploma de Estudios Avanzados (DEA), Department of Information Systems and Computation, UPV (2007)
9. Pinto, D., Rosso, P., Jiménez, H.: A self-enriching methodology for clustering narrow domain short texts. Comput. J. **54**(7), 1148–1165 (2011)
10. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 536–546. Springer, Heidelberg (2006)
11. Hasanzadeh, E., Poyan, M., Rokny, H.: Text clustering on latent semantic indexing with particle swarm optimization (PSO) algorithm. Int. J. Phys. Sci. **7**(1), 116–120 (2012)
12. Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009)
13. Kim, S.N., Medelyan, O., Kan, M.Y., Baldwin, T.: Automatic keyphrase extraction from scientific articles. Lang. Resour. Eval. **47**(3), 723–742 (2012)
14. Eissen, S.M., Stein, B.: Analysis of clustering algorithms for Web-based search. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2002. LNCS (LNAI), vol. 2569, pp. 168–178. Springer, Heidelberg (2002)
15. Stein, B., Meyer zu Eissen, S., Wißbrock, F.: On cluster validity and the information need of users. In: Hanza, MH. (ed.) 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03), Benalmádena, Spain, pp. 216–221, ISBN 0-88986-390-3. ACTA Press, IASTED (2003)