Investigation of Latent Semantic Analysis for Clustering of Czech News Articles

Michal Rott, Petr Cerva
Institute of Information Technology and Electronics
Technical University of Liberec
Studentska 2, 461 17, Liberec, Czech Republic,
https://www.ite.tul.cz/itee/
Email: michal.rott@tul.cz, petr.cerva@tul.cz

***Abstract*—This paper studies the use of Latent Semantic Analysis (LSA) for automatic clustering of Czech news articles. We show that LSA is capable of yielding good results in this task as it allows us to reduce the problem of synonymy. This is a very important factor particularly for Czech, which belongs to a group of highly inflective and morphologically-rich languages. The experimental evaluation of our clustering scheme and investigation of LSA is performed on query- and category-based test sets. The obtained results demonstrate that the automatic system yields values of the Rand index that are absolutely lower – by 20% – than the accuracy of human cluster annotations. We also show which similarity metric should be used for cluster merging and the effect of dimension reduction on clustering accuracy.**

I. Introduction

The task of automatic clustering of text documents has attracted a lot of attention recently. Since 1988, when one of the first studies on automatic clustering came out [1], various methods have been proposed in this field [2], [3]. These approaches can be divided into several categories [4], e.g., model-based, density-based, hierarchical or partitioning types.

The goal of this work is to find out an approach that would be suitable for clustering of documents in the highly inflective Czech language with a rich vocabulary. Thus, we take advantage of Latent Semantic Analysis (LSA) [5], which allows us to find a low-rank approximation of a term-document matrix, describing the number of occurrences of words (terms) in input documents [6]. In the case of inflective languages, rank lowering is important for two reasons. First, it should allow us to merge the dimensions associated with terms of similar meanings: the estimated term-document matrix is presumed overly sparse in comparison to the "true" term-document matrix, because it contains only the words actually seen in each document, whereas it should list a much larger set (due to synonymy) of all words related to each document. Second, rank lowering is expected to de-noise the original term-document matrix, which contains semantically unimportant (noisy) terms.

For these reasons, LSA has also been successfully applied in the task of automatic summarisation of documents and information retrieval. In the former case, it allows us to find a decomposition that describes an importance degree for each topic of the document in each sentence. The resulting summary of the input document is then created by choosing the most important sentences [7],[8]. In the latter task, given a query composed of several words (terms), LSA translates it into a low-dimensional space, and finds matching documents [9], [10].

The LSA-based clustering approach, which is adopted in this work combines hierarchical and model-based methods. First, LSA is used to create a vector representation of the input documents in the concept space. These vectors are then merged within an agglomerative hierarchical clustering approach. We also take advantage of a text-preprocessing module, which performs language-dependent operations such as substitution of synonyms and lemmatisation.

The rest of the paper is structured as follows: The next section describes the adopted clustering scheme. Section III then describes the measures used for assessment of document similarity and the metrics used for evaluation of clustering. Experimental evaluation of query-based and category-based test sets is then given in Section IV. In this section, we investigate various possibilities how LSA-based decomposition and clustering can be performed. Section V then concludes this paper.

II. Adopted Clustering Scheme

The clustering scheme used within this work is based on the use of LSA and it consists of three phases. In the first phase, a term-document matrix is constructed and decomposed to a concept space using LSA. Next, the dimensionality of the concept space is reduced and, after that, hierarchical clustering is performed in the third phase. All three of these steps are detailed in the following subsections.

A. The term-document matrix and its decomposition

At first, all the input documents are preprocessed and lemmatised. The resulting text then does not just contain lemmas, but also word forms, such as numbers or typing errors, which cannot be lemmatised. We refer to all of these items as terms.

Given the preprocessed set of input documents, the frequency of occurrence, i.e., the term frequency (TF), is calculated for every unique term from these documents, excluding terms in the stop list. After that, the frequency of each term is weighted by its inverse document frequency (IDF) [11], [12].

These IDF values can be expressed as:

$$IDF(l) = log \frac{|D_b|}{|\{d_b \in D_b : l \in d_b\}|} \quad (1)$$

where $|D_b|$ is the total number of training documents in the background corpus and $|\{d_b \in D_b : l \in d_b\}|$ is the number of background documents containing the term l.

Given the weighted term frequency values, the term-document matrix $\boldsymbol{A}$ is constructed, where each column vector represents a weighted term frequency vector of one input document. Therefore, the size of $\boldsymbol{A}$ is $t \times d$, where t is the number of all unique terms in all input documents and d is the total number of input documents.

After that, the LSA is performed. This method employs the Singular Value Decomposition (SVD) to the matrix $\boldsymbol{A}$ as follows:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \quad (2)$$

where $\boldsymbol{U}$ is a $t \times m$ column-orthonormal matrix of left singular vectors, $\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2, ..., \sigma_m)$ is an $m \times m$ diagonal matrix whose diagonal elements represent non-negative singular values sorted in descending order, and $\boldsymbol{V}$ is an $m \times d$ orthonormal matrix of right singular vectors.

It has been shown [13] that the matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}^T$ represent a concept (semantic) space of the input documents: the matrix $\boldsymbol{U}$ describes a mapping of concepts to the space of terms, $\boldsymbol{V}^T$ captures how concepts are mapped to documents, and the singular values of $\boldsymbol{\Sigma}$ represent the significance of individual concepts. Note that a more detailed description of the SVD can be found in [9].

B. Concept space dimension reduction

In practice, the term-document matrix $\boldsymbol{A}$ is usually sparse because individual input documents can contain a) synonyms or b) partially or completely different words and word forms. Note that the latter problem occurs particularly for inflective languages such as Czech. The next issue is that $\boldsymbol{A}$ also contains noise that is represented by common and/or meaningless terms.

From a linguistics point of view, these issues can partially be eliminated by using:

- stop list of meaningless terms
- processing module for text normalisation and lemmatisation
- minimum occurrence threshold for terms

Within this work, we utilised the first two options to reduce the sparsity of $\boldsymbol{A}$ (see section IV-F for details and obtained results).

From the mathematical point of view, this problem can also be addressed by low-rank approximation that reduces the number of dimensions of the concept space from m to k (see Fig. 1). Unfortunately, the problem of finding the proper value of k is not trivial and its solution is usually based on a heuristic knowledge of the given task. In section IV-E we present our result obtained for values of k in the range from 10 % of sum of all singular values to 100 %.

C. Hierarchical clustering

The adopted hierarchical clustering approach is based on the assumption that the documents belonging to the same cluster should have similar concepts (topics). Therefore, after dimensionality reduction, clustering can be performed for vectors of $\boldsymbol{A}_K$ or for vectors of reduced matrix $\boldsymbol{V}^T$ which describes mapping of concepts to documents (see section for results IV-D).

In both cases, we perform clustering in an agglomerative way, where each document represents one cluster at the beginning. Then, pairs of the most similar clusters are merged in consecutive steps until the demanded number of clusters is reached. As a similarity measure, we employ various metrics described in section III-A). The outcome of clustering is a set of clusters where every cluster contains a list of documents that should have similar concept (topic).

III. Metrics

A. Similarity measures used for hierarchical clustering

The similarity metrics are used to find a pair of the closest documents (clusters) during the process of agglomerative hierarchical clustering. For this purpose, each pair of documents is represented by vectors $\vec{a}$ and $\vec{b}$ that correspond to columns of A or V^T. When two documents (or clusters) are merged to form a new cluster, this cluster is then represented by the average vector (centroid) calculated over all documents belonging to the cluster.

Within this work, the following similarity metrics were used:

Euclidean distance:

$$d_e\left(\vec{a}, \vec{b}\right) = \sqrt{\sum_{i=1}^{k} \left(a_i - b_i\right)^2} \quad (3)$$

Normalised Euclidean distance:

$$d_n\left(\vec{a}, \vec{b}\right) = \sqrt{\sum_{i=1}^{k} \frac{\left(a_i - b_i\right)^2}{\sigma_i^2}} \quad (4)$$

Cosine similarity:

$$d_c\left(\vec{a}, \vec{b}\right) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} \quad (5)$$

where k ($=< t$)is the number of dimensions of the reduced concept space and σ_i^2 is the variance of i^{th} element across the given space.

B. Clustering evaluation metrics

Silhouette: Silhouette is a clustering evaluation metric that describes how well documents are assigned to clusters [14]. Silhouette takes on values inside the interval $\langle -1, 1 \rangle$ and is defined as:

$$SI = \frac{1}{N} \sum_{i=1}^{N} \frac{n_i - c_i}{max\left\{c_i, n_i\right\}}, \quad (6)$$

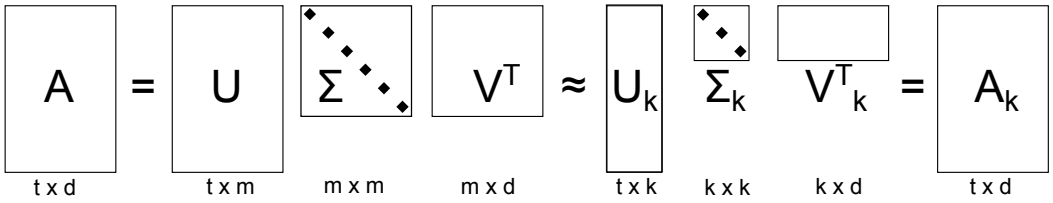

Fig. 1. Decomposition and concept reduction of the term-document matrix

where $c(i)$ is computed as an average similarity of the i^{th} document to all other documents in the same cluster and $n(i)$ is the minimum of average similarity of the i^{th} document to all other clusters. This means that $n(i)$ is the similarity to the closest (neighbour) cluster. N is the number of all documents.

Rand index: The Rand index [15] is an evaluation metric that compares automatically created clusters C with ideal reference clusters R. This metric takes on values inside the interval $\langle 0, 1\rangle$ and is expressed as:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where TP is the number of pairs of documents in the same cluster in R and in the same cluster in C; TN is the number of pairs of documents in different clusters in R and in different clusters in C;, FP is the number of pairs in different clusters in R and in the same cluster in C; and FN is the number of pairs in the same cluster in R and in different clusters in C.

IV. Experiments

A. Experimental setup

Two different test sets of documents were used for experimental evaluation.

To form the first set, a query "Česká národní banka" (Czech national bank) was first entered into a web search engine and after that, 20 documents containing this phrase were selected from search results. Therefore, we call this test set query-based. Ten people were then asked to cluster all these documents into five distinct clusters and as a result of this process, 10 × 5 reference clusters were obtained.

In contrast, the second test set is category-based. It is composed of 100 newspaper articles belonging to ten different topic categories such as culture, economics, etc. That means that, in this case, the documents were clustered just using information on their category. Note that each of the ten category-based clusters contains the same number of ten articles.

The statistics of both these test sets are summarised in Table I.

TABLE I. Test sets used for evaluation

test set	# of document	average # of characters	average # of words
query-based	20	2046	194
category-based	100	1480	151

B. The Preprocessing Module

For evaluation, we utilised a text preprocessing module developed originally for our summarisation engine[1]. The module converts the input text to its normalised form.

At first, every sentence is lemmatised using an external morphological tool HUNSPELL[2].

The goal of the next step is to substitute all synonyms of every lemma using one preferred form. The substitution is based on the use of a lemmatised dictionary of synonyms, which contains 7443 different groups of synonyms with a total of 22856 lemmas. These items are compiled from two sources. The first is the Czech version of the project Wiktionary[3]. The second is the Thesaurus project[4].

Note that the ALGLIB[5] mathematical library is used in our clustering system for matrix operations.

C. Comparison of similarity measures

The goal of the first experiment was to compare the performance of individual similarity measures. The obtained results shown in Table II are presented in terms of silhouette (SI) and Rand index (RI). For the query-based test set, the output from our clustering scheme was compared to all ten reference clusters. Thus, the mean values of SI and RI are presented in this case.

In this experiment, dimension reduction was also performed to different levels. For example in the case of the query-based test set, reduction to two dimensions means that just the two largest singular values were retained that together represent 20% of the sum of all singular values. Note that the vectors of matrix $\boldsymbol{V}_K^T$ were cluster.

TABLE II. Comparison of different similarity measures

test set	query-based		category-based	
	RI	SI	RI	SI
reduction 80 %	2 dimensions		10 dimensions	
euclidean	0.565	0.691	0.235	-0.438
norm. euclidean	0.543	-0.473	0.235	-0.438
cosine	0.500	-0.645	0.733	-0.350
reduction 50 %	7 dimensions		33 dimensions	
euclidean	0.429	0.764	0.233	-0.102
norm. euclidean	0.431	-0.311	0.233	-0.216
cosine	0.651	-0.230	0.752	-0.273
reduction 20 %	13 dimensions		64 dimensions	
euclidean	0.446	0.449	0.235	0.115
norm. euclidean	0.431	0.081	0.237	0.039
cosine	0.635	-0.224	0.714	-0.163

[1] http://summec.ite.tul.cz
[2] http://hunspell.sourceforge.net/
[3] http://cs.wiktionary.org
[4] http://packages.debian.org/sid/myspell-cs
[5] http://www.alglib.net

The obtained values of SI show that a high compactness of clusters (i.e., a high value of SI) does not mean that the clusters correspond to reference (i.e., that the clusters have a high value of RI). That is the reason why just the Rand index is used as the main evaluation metric in the following experiments. It is also evident that the cosine similarity metric yielded the highest values of RI in both cases. Therefore, this metric alone is employed for clustering in all other experiments.

D. Comparison of the use of $\boldsymbol{A}_K$ and $\boldsymbol{V}_K^T$ for clustering

As mentioned in section II-C, two possibilities exist how clustering can be performed within the LSA concept. The first approach corresponds to the previous experiment, where the matrix $\boldsymbol{V}_K^T$ was used. The second method is based on the use of $\boldsymbol{A}_K$, which is created by multiplication of reduced matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}_K^T$. In Tab. III, we present a comparison of results yielded by using both of these options.

TABLE III. VALUES OF RI AFTER CLUSTERING OF VECTORS FROM $\boldsymbol{A}_K$ AND $\boldsymbol{V}_K^T$

test set	$\boldsymbol{A}_K$	$\boldsymbol{V}^T$
query-based	0.621	0.635
category-based	0.624	0.714

The obtained results show that higher values of RI can be reached when vectors $\boldsymbol{V}_K^T$ are clustered. Another advantage of using $\boldsymbol{V}_K^T$ is that the clustering process takes less computational time.

E. Experiments with dimension reduction

Given all the previous results, the next experiment is aimed at finding the optimal reduction coefficient for our evaluation sets. The obtained results are illustrated in Fig. 2, where also the mean values of results on both test sets are depicted.

They show that the number of dimensions can be reduced to a level where 50% of the total sum of singular values is retained. In this case, the value of RI for the query- and category-based test set was 0.651 and 0.784, respectively. The next conclusion is that dimension reduction should always be performed. Without this step, RI went below 0.4.

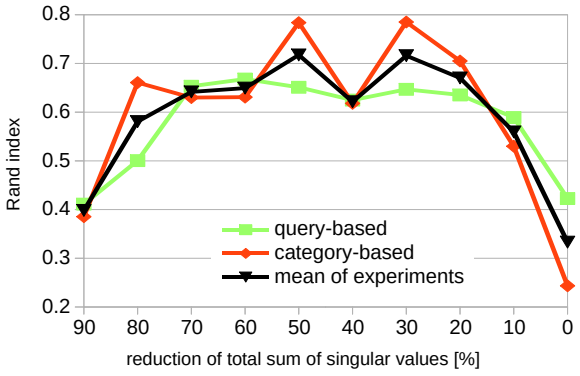

Fig. 2. Rand index in dependency on dimension reduction

F. Investigation of the preprocessing module influence

The next experiment (see Tab. IV) shows the RI values reached with and without the use of the preprocessing module. We can see that this module improves RI for both test sets as it reduces dimensions of the term-document matrix prior to application of LSA. However, this improvement is only slight: by 0.037 on average.

Note that when dimension reduction is not performed after SVD decomposition within LSA, a big drop in RI was observed (by 0.35 absolutely) – see Fig. 2. That means that the mechanism of dimension reduction built-in within LSA is much more efficient and important than the reduction effect of the preprocessing module.

TABLE IV. EFFECT OF THE PREPROCESSING MODULE ON VALUES OF THE RAND INDEX

test set	query-based (mean of RI)	category-based (RI)
original input text	0.610	0.752
preprocessed text	0.651	0.785

G. Accuracy of human cluster annotations

The previous results on the query-based test set showed that the adopted automatic clustering approach yields RI around 0.65. In the last experiment, a cross-validation of human reference annotations was performed to compare this value with the accuracy of clustering by humans. Each time, one of the ten human annotations was selected as the reference and the remaining nine annotations were evaluated against this one. As a result of this process, ten values of RI were obtained. Their minimal, maximal and average values are presented in Tab V.

TABLE V. ACCURACY OF HUMAN CLUSTER ANNOTATIONS

test set	min RI	mean RI	max RI
query-based	0.816	0.871	0.908

These results show that the accuracy of human clustering in terms of RI is around 0.87. This means that our human annotators were capable of producing clusters with the accuracy that is absolutely by 20% higher than the accuracy yielded by our automatic clustering scheme.

V. CONCLUSIONS

In this paper, we presented our document clustering scheme and evaluated its performance on two different test sets containing Czech newspaper articles. This evaluation demonstrated that the cosine distance should be used as a similarity measure within clustering and that the vectors of matrix $\boldsymbol{V}^T$ should be merged. Our results also showed that it is important to perform dimension reduction: only the largest singular values corresponding to 50% of the total sum of all singular values can be retained. Using these settings, the presented clustering scheme yielded RI that was absolutely by 20% worse than RI of human cluster annotations.

ACKNOWLEDGMENT

This paper was supported by the Technology Agency of the Czech Republic (Project No. TA01011204) and by the Student Grant Scheme (SGS) at the Technical University of Liberec.

REFERENCES

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[2] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[3] N. O. Andrews and E. A. Fox, "Recent developments in document clustering," *Computer Science, Virginia Tech, Tech Rep*, 2007.

[4] L. Rokach, "A survey of clustering algorithms," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2010, pp. 269–298. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-09823-4_14

[5] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01638539809545028

[6] C.-P. Wei, C. C. Yang, and C.-M. Lin, "A latent semantic indexing-based approach to multilingual document clustering," *Decis. Support Syst.*, vol. 45, no. 3, pp. 606–620, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2007.07.008

[7] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[8] M. Rott and P. Cerva, "Summec: A summarization engine for czech." in *TSD*, ser. Lecture Notes in Computer Science, I. Habernal and V. Matousek, Eds., vol. 8082. Springer, 2013, pp. 527–535. [Online]. Available: http://dblp.uni-trier.de/db/conf/tsd/tsd2013.html#RottC13

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, no. 6, pp. 391–407, 1990.

[10] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57. [Online]. Available: http://doi.acm.org/10.1145/312624.312649

[11] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Mit Press, 1999. [Online]. Available: http://books.google.cz/books?id=YiFDxbEX3SUC

[12] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *Journal of Documentation*, vol. 60, p. 2004, 2004.

[13] K. Baker, "Singular value decomposition tutorial," 2005.

[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53 – 65, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257

[15] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971. [Online]. Available: http://www.jstor.org/stable/2284239