International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 5 (2013) pp. 509-515
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Supervised and Unsupervised Text Classification via Generic Summarization

Dmitry Tsarev[1], Mikhail Petrovskiy[2] and Igor Mashechkin[3]

[1] Computer Science Department, Lomonosov Moscow State University,
Moscow, Russia
tsarev@mlab.cs.msu.su

[2] Computer Science Department, Lomonosov Moscow State University,
Moscow, Russia
michael@cs.msu.su

[3] Computer Science Department, Lomonosov Moscow State University,
Moscow, Russia
mash@cs.msu.su

***Abstract*: This paper presents a new generic text summarization method using Non-negative Matrix Factorization (NMF) to estimate sentence relevance. Proposed sentence relevance estimation is based on normalization of NMF topic space and further weighting of each topic using sentences representation in topic space. The proposed method shows better summarization quality and performance than state of the art methods on DUC 2002 standard dataset. In addition, we study how this method can improve the performance of supervised and unsupervised text classification tasks. In our experiments with Reuters-21578 and 20 Newsgroups benchmark datasets we apply developed text summarization method as a preprocessing step for further multi-label classification and clustering. As a result, the quality of classification and clustering has been significantly improved.**

Keywords: generic text summarization, latent semantic analysis, non-negative matrix factorization, multi-label classification, clustering.

I. Introduction

Automatic classification of documents has become an important research issue since the overload of electronic text information. There are mainly two machine learning approaches to resolve this task: supervised approach, where predefined category labels are provided for training set of documents, and unsupervised document classification (also known as document clustering), where the classification must be done entirely without reference to external information.

In this paper, we consider both of these approaches. As a supervised classification we have chosen a multi-label classification, which is a further generalization of traditional multi-class learning task. In multi-label case the classes are not mutually exclusive and any sample may belong to several classes in the same time. As an unsupervised classification we use two clustering methods, both flat and hierarchical. Flat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other. Hierarchical clustering creates a hierarchy of clusters.

Another text mining task we consider is automatic text summarization. It becomes very important recently because of upraising information overload. Text summaries can be either query-based summaries or generic summaries. A query-based summary presents the contents of the document that are closely related to the initial user query. As opposed to that, a generic summary is aimed at a broad community of readers and should contain all main topics of the text [1], [21].

This paper presents a new text summarization method, which constructs generic summaries in extracts form. These are the summaries completely consisting of fragments taken from the original text. Phrases, sentences or paragraphs can be used as the text fragments. A sentence is usually used to express content in summarization. We will consider text sentences as basic fragments below. However, for longer documents content can be represented by a set of paragraphs as basic fragments. The developed method has been experimentally verified on DUC 2002 benchmark dataset [2], [3] with state of the art methods.

Generic summary of the document contains the fragments (sentences), which describe all main topics of the text. Therefore in this paper we also study the applicability of documents summaries instead of original texts in multi-label classification and clustering tasks. Since the document may have more than one topic the multi-label classification task has been chosen as a more general approach in comparison to traditional multi-class classification. In multi-label case each document can belong to several classes, i.e. may have several topics. In addition to supervised classification, we experimented with flat and hierarchical clustering, to study how our method can improve classification of unlabeled documents.

The remainder of this paper is organized as follows: Section II presents a proposed generic text summarization method and its experimental comparison with state of the art methods. Section III is devoted to experimental investigation of our approach, where each full text document is replaced with its summary, in multi-label classification and clustering tasks. Finally, in Section IV, we conclude the paper.

II. Generic Text Summarization Methods

Nowadays the most state of the art methods of automatic text summarization which build generic summaries in the extracts form are based on Latent Semantic Analysis (LSA) [1], [4]-[6]. In these methods the original text is represented in the form of a numerical matrix. Matrix columns correspond to text sentences (or other fragments), and each sentence is represented in the form of a vector in the text term space. Further, LSA is applied to the received matrix to construct sentences representation in the text topic space. The dimensionality of the topic space is much less than dimensionality of the initial term space. The choice of the most important sentences is carried out on the basis of sentences representation in the topic space. The number of important sentences is defined by the length of the demanded summary (the length is usually measured in the number of words).

LSA performs one of the matrix decomposition algorithms on the original text matrix to construct sentences representation in text topic space, thereby bringing out the semantic connectedness present among the sentences [7]. Singular Value Decomposition (SVD) is the traditional matrix decomposition algorithm used for LSA, wherein lower dimensional components from the decomposition are truncated. On truncation, the linguistic noise present in the vector representation is removed, and the semantic connectedness is made visible. One of the disadvantages of using SVD is that the truncated matrix will have negative components, which is not natural for interpreting the textual representation. Nonnegative Matrix Factorization (NMF) addresses this issue by constructing non-negative parts-based representation as the matrix decomposition algorithm for performing LSA [6]-[8].

Further we describe the proposed generic text summarization method using NMF to estimate sentence relevance. And also we adduce its experimental comparison with state of the art methods using SVD and NMF.

A. Proposed Generic Text Summarization Method

The first step is the creation of a term by sentences matrix $A = [A_1, A_2, \ldots, A_n]$, where each column A_i represents the weighted term frequency vector of sentence i in the document under consideration. The sentence vector $A_j = [a_{1j}, a_{2j}, \ldots, a_{mj}]^{\mathrm{T}}$ is defined as: $a_{ij} = L(t_{ij})\cdot G_i$, where t_{ij} denotes the frequency with which term i occurs in sentence j, $L(t_{ij})$ is the local weight for term i in sentence j, and G_i is the global weight for term i in the whole document. We have experimented with various weighting schemes [9] and received the best results by using binary local weight and entropy global weight:

- *Binary local weight*: $L(t_{ij}) = 1$, if term i appears at least once in sentence j; $L(t_{ij}) = 0$, otherwise.
- *Entropy global weight* (1):

$$G_i = 1 - \sum_{j=1}^{N} \left(\frac{p_{ij} \log p_{ij}}{\log N} \right), \quad (1)$$

where $p_{ij} = t_{ij} / F_i$, F_i is the total number of times that term i occurs in the whole document, N is the number of sentences in the document.

If there are m terms and n sentences in the document we obtain an $m \times n$ matrix A for the document. The next step is to apply NMF to matrix A: $A \approx W\cdot H$. Matrix W derives a mapping between the m dimensional term space and the r dimensional topic space. Each column of H represents corresponding text sentence as an additive combination of the basis topics.

Then we normalize the topic space: $A_k = W\cdot H = NormW\cdot NormH$, where $NormW = W\cdot Norm^{-1}$, $NormH = Norm\cdot H$, $Norm = diag(\|W^1\|, \ldots, \|W^r\|)$. We use Euclidean norm for columns of matrix W (2):

$$\left\| W^k \right\| = \sqrt{\sum_{i=1}^{m} W_{ik}^2}\ , \ 1 \le k \le r. \quad (2)$$

Columns of matrix $NormH$ correspond to n sentences in the normalized topic space. The k-th row $NormH_k = [normh_{k1}, \ldots, normh_{kn}]$ indicates weights of k-th topic in all n sentences. The greater norm of $NormH$ rows, the greater weights of corresponding topics in all text. Proceeding from it, we calculate topic weights as norms of rows of matrix $NormH$. Weight of k-th topic is (3), (4):

$$\left\| NormH_k \right\| = \sqrt{\sum_{j=1}^{n} NormH_{kj}^2}\ , \quad (3)$$

$$\left\| NormH_k \right\| = \left\| W^k \right\| \sqrt{\sum_{j=1}^{n} H_{kj}^2} = \left\| W^k \right\| \cdot \left\| H_k \right\|, \ 1 \le k \le r. \quad (4)$$

The weighted sentences representation in the normalized topic space is matrix $WeightedH$ (5), (6):

$$WeightedH = diag\left(\left\| W^1 \right\| \cdot \left\| H_1 \right\|, \ldots, \left\| W^r \right\| \cdot \left\| H_r \right\| \right) \cdot NormH\ , \quad (5)$$

$$WeightedH = diag\left(\left\| W^1 \right\|^2 \cdot \left\| H_1 \right\|, \ldots, \left\| W^r \right\|^2 \cdot \left\| H_r \right\| \right) \cdot H\ . \quad (6)$$

Sentences for the summary are selected according to their sum of topic weights. Relevance of j-th sentence is (7), (8):

$$R_j(WeightedH) = \sum_{i=1}^{r} (WeightedH_{ij})\ , \quad (7)$$

$$R_j(WeightedH) = \sum_{i=1}^{r} \left(\left\| W^i \right\|^2 \cdot \left\| H_i \right\| \cdot H_{ij} \right)\ . \quad (8)$$

Finally required number of sentences with the highest relevance values is selected for the summary [22], [23].

B. Experiments

We use the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) package to evaluate the proposed method [10], [3]. It includes measures to automatically determine the quality of a summary by comparing it to other model (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. ROUGE measures ROUGE-2, ROUGE-L, ROUGE-S, ROUGE-W is recommended to use for evaluation single-document summarization methods on datasets DUC 2001 and DUC 2002 [3]. In the last editions of Document Understanding Conferences (DUC), ROUGE was used as an automatic evaluation method. As experimental data, we use the DUC 2002 standard dataset. This dataset consists of 533 documents and 925 model summaries.

We evaluated state of the art summarization methods such as the *SVD-Classic* (Gong and Liu approach [4]), the *SVD-Square* (Steinberger and Ježek approach [5], [11]), the *NMF-Generic* (Lee, Park, Ahn, Kim approach [6]), and our proposed method based on NMF.

Also we consider method which extracts sentences with the highest word count, i.e. sentences for the summary are selected according to their number of words, except stop words. And finally we consider random sentence extraction method. We denote these methods as *Word Count* and *Random*, respectively.

The number of topics for LSA is selected much smaller than dimensionalities of the text matrix, i.e. $r << \min(m, n)$. For DUC 2002 dataset average number of document matrix rows is 239, and average number of columns is 37. The diagram (Fig. 1) shows change of the ROUGE-2 *f*-measure depending on number of topics, where number of topics varies from 1 to 20. ROUGE-2 *f*-measure has been selected because the results of using various ROUGE measures are similar.

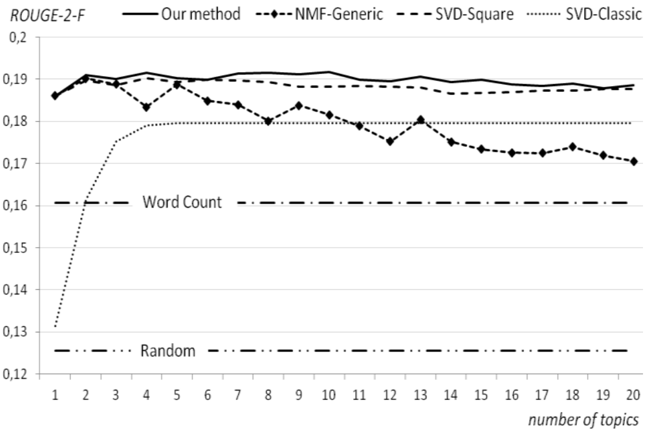

Figure 1. ROUGE-2 *f*-measure / number of topics

When we perform one of the matrix decomposition on an $m \times n$ text matrix, we can view the new dimensions as some sort of pseudo sentences: linear combinations of the original terms (left singular vectors in SVD case, and matrix W in NMF case). From a summarization point of view, the number of extracted sentences is dependent on the summary ratio (a ratio of the summary length with respect to the length of the original text). We know what percentage of the full text the summary should be: part of the input to the summarizer is that a p% summary is needed. If the pseudo sentences were real sentences that a reader could interpret, we could simply extract the top r pseudo sentences, where $r = (p/100)\cdot n$. However, because the linear combinations of terms are not really readable sentences, we use four methods mentioned above to extract the actual sentences that 'overlap the most' in terms of vector length with top r pseudo sentences [11].

We built summaries consisting of 100 words since model summaries of DUC 2002 dataset consist of 100 words. For each method we calculate number of topics as $r = (100/nw)\cdot n$, where nw is the total number of terms in the text. Table I shows the ROUGE *f*-measure values of four methods using ROUGE evaluation. In addition we show the ROUGE scores for *Random* and *Word Count* methods in Table II.

Measure	SVD-Classic	SVD-Square	NMF-Generic	Our Method
ROUGE-2	0.17922	0.18933	0.18385	**0.19251**
ROUGE-L	0.35351	0.36799	0.36467	**0.37230**
ROUGE-S4	0.14009	0.15091	0.14636	**0.15358**
ROUGE-W	0.19893	0.20874	0.20405	**0.21066**

TABLE I. ROUGE F-MEASURE VALUES

Measure	Random	Word Count
ROUGE-2	0.12364	0.14727
ROUGE-L	0.29483	0.31329
ROUGE-S4	0.09503	0.11441
ROUGE-W	0.16227	0.17510

TABLE II. ROUGE F-MEASURE VALUES

We use MATLAB function *svds()* for implementation SVD. NMF also is implemented on MATLAB. The number of iterations in SVD and NMF algorithms is restricted 300. We have received, that NMF works faster SVD for 20 percent.

The experiments demonstrate better summarization quality and performance of our proposed method in comparison with other methods.

III. Using developed text summarization to improve documents classification tasks

We use our NMF-based method and SVD-Square method, as shown the best results in previous section, to replace the full document text by its summary as a preprocessor step for further classification.

The length of the summary is defined from a percentage of initial text information amounts. We use topic weights to estimate an information amount. In SVD-Square method weights are defined as a square of corresponding singular values [20], [5], [11]. In order to the summary contained pi% information of the initial text, we perform the full decomposition of $m \times n$ text matrix for $ri = \min(m, n)$ dimensions. All ri topics produce 100% information and their contribution corresponds to their weights. The number of summary topics k is selected proceeding from ratio of the sum of k maximum weights with respect to the sum of all weights and this ratio should equal to $pi/100$ (9):

$$\frac{\sum_{i=1}^{k} sort(weight,'descend')_i}{\sum_{j=1}^{ri} weight_j} \approx \frac{pi}{100}, \quad (9)$$

where *weight* is a sequence of topic weights, *sort(weight, 'descend')* is a sequence of topic weights which elements are sorted in the descending order. Further we use our NMF-based or SVD-Square methods, corresponding to NMF or SVD decompositions, to extract *k* most relevant sentences.

In this section we adduce experiments with replacement the full document text by its summary for multi-label classification and clustering tasks. A traditional vector space representation is used in all classification methods considered in this paper [9]. Therefore the normalized weighting scheme *tf·idf* was used for vector representation, both documents, and their summaries [24].

A. Multi-label Classification

The naive approach to multi-label learning is based on *one-against-all binary decomposition (1 vs. all).* For each class separate binary classification subproblem is formulated. In this subproblem all samples from the multi-label training set are divided into two disjoint subsets: "positive" samples, whose belong to this class, and "negative" samples that do not belong. Then traditional binary learning algorithm is applied to each binary subproblem. As a result, a set of independent binary classifiers are trained. Each classifier is associated with class and predicts whether a given sample belongs to this particular class.

In this paper we also use multi-label classification method based on paired comparisons approach (i.e. *one-against-one binary decomposition, 1 vs. 1*), which is described in [12]. In this method each pair of possibly overlapping classes is separated by two probabilistic binary classifiers, which isolate the overlapping and non-overlapping areas. Then individual probabilities generated by binary classifiers are combined together to estimate final class probabilities fitting extended Bradley-Terry model with ties.

As the binary classifier in these multi-label classification methods we use a Support Vector Machine (SVM) [13]. In addition to these methods we apply linear threshold function defined on the class relevancies vector space [14].

Results of multi-label classification experiments we evaluate by Hamming Loss criterion. Hamming loss measures average symmetric set difference (Δ) between predicted and relevant sets of classes for test documents (10):

$$HammingLoss = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|Y|}\left|F(x_i)\,\Delta\,y_i\right|, \quad (10)$$

where *k* is number of test documents; *Y* is the set of document classes; x_i is a test document; $F(x_i)$ is the set of predicted classes for document x_i; y_i is the set of relevant classes for document x_i.

We evaluated multi-label classification of full texts and their summaries on Reuters-21578 dataset [15]. This is one of the most popular benchmark datasets for multi-label classification. Reuters-21578 documents are presented in SGML format. Therefore the texts we have selected by following criteria: the TOPICS node contains one or more elements; attribute TYPE of TEXT node possess value NORM. We have divided the obtained dataset on training and test using values of attribute LEWISSPLIT. The value TEST of attribute LEWISSPLIT indicates the document was used for testing, the other values of this attribute indicate the document was used for training. In addition to this dataset we also use its short version, where documents are not less than 512 bytes. Table III shows characteristics of the obtained datasets.

Dataset	Training documents	Test documents	Number of classes(topics)
complete	7068	2745	120
reduced	2295	800	120

TABLE III. MULTI-LABEL BENCHMARK DATASETS

Diagrams in Fig. 2 are showing changes of Hamming Loss criterion for *1 vs. all* multi-label classification depending on percentage of the information amount selected for summaries in our NMF-based summarization method. The diagram *"100 words"* corresponds to a case of using additional summary length limitation — the minimum summary length should be 100 words; the diagram *"0 words"* corresponds to a case without this limitation; the line *"full text"* is a case of full document text classification. From this data follows the best results have been received with conditions: 30% threshold of the information amount selected for summaries and *"100 words"* limitation. Similar results have been received for *1 vs 1* multi-label classification. We will use these two conditions further in classification tasks.

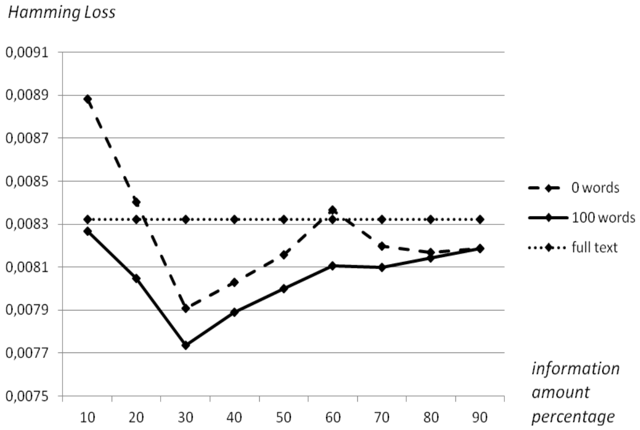

Figure 2. Hamming Loss / information amount percentage

Experimental results with multi-label classification of Reuters-21578 datasets and its summaries are presented in the Table IV and the Table V, corresponding to our NMF-based and SVD-Square summarization methods. Our summarization method shows better results, than SVD-Square method. In addition, we have received the text preprocessing by NMF decomposition with our weights calculation works faster SVD for approximately 10 percent.

A comparison of the dataset sizes (in word count and in megabytes) and dimensionality of feature space (number of different terms) with its summarized versions by proposed NMF-based method is resulted in Table VI and Table VII.

Method	**Dataset**	**Normal Hamming loss**	**Summarized Hamming loss**	**Improving**
1 vs. 1	reduced	0,0092917 ±2,6e-05	0,0087813 ±3,1e-05	5,5%
1 vs. 1	complete	0,0083677 ±1,8e-05	0,0042653 ±2,7e-05	48,8%
1 vs. all	reduced	0,0083229 ±3,6e-05	0,0077361 ±5,2e-05	7%
1 vs. all	complete	0,0038859 ±1,5e-05	0,003769 ±1,5e-06	3%

TABLE IV. MULTI-LABEL CLASSIFICATION RESULTS (NMF CASE)

Method	**Dataset**	**Normal Hamming loss**	**Summarized Hamming loss**	**Improving**
1 vs. 1	reduced	0,0092917 ±2,6e-05	0,009257 ±2,6e-05	0,4%
1 vs. 1	complete	0,0083677 ±1,8e-05	0,0045598 ±9,1e-06	45,5%
1 vs. all	reduced	0,0083229 ±3,6e-05	0,0083125 ±1,0e-05	0,12%
1 vs. all	complete	0,0038859 ±1,5e-05	0,0039208 ±1,5e-06	-0,9%

TABLE V. MULTI-LABEL CLASSIFICATION RESULTS (SVD CASE)

Dataset	**Initial (words/Mb)**	**Summarized (words/Mb)**	**Improving**
complete	803330 4,81Mb	361823 3,9Mb	55% 19%
reduced	518412 3Mb	198536 2,2Mb	61,7% 26,7%

TABLE VI. REUTERS-21578 SIZE REDUCTION

Dataset	**Training documents**	**Test documents**	**Number of classes(topics)**
complete	21211	19534	7,9%
reduced	14803	12769	13,7%

TABLE VII. REUTERS-21578 FEATURE SPACE REDUCTION

From the obtained experimental results follows the using summaries instead of full texts improves quality of multi-label classification. Therefore, it is possible to draw a conclusion, that the text summarization methods well defines main topics of documents and on their basis selects sentences, which in the best way describe them. But our presented NMF-based method shows better classification quality and performance than SVD analogue.

From the obtained experimental results follows the using summaries instead of full texts improves quality of multi-label classification. Therefore, it is possible to draw a conclusion, that the text summarization methods well defines main topics of documents and on their basis selects sentences, which in the best way describe them. But our presented NMF-based method shows better classification quality and performance than SVD analogue.

B. Clustering

We consider two most popular clustering algorithms based on matrix decomposition, such as SVD, NMF. The first is the Principal Direction Divisive Partitioning (PDDP) algorithm separates the entire set of documents into two partitions by using principle directions, which are obtained after SVD decompositions term-document matrix. Each of two partitions will be separated into two sub-partitions using the same process recursively. The result is hierarchical of partitions arranged into a binary tree. Thereby for specified k we can receive $n \in [k, 2^k]$ clusters [16].

The second is a document clustering method based on the NMF of the term-document matrix of the given document corpus. In the latent semantic space derived by the NMF, each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value [17].

We use two external criteria of clustering quality: Rand index and F-measure [18]. The Rand index (RI) measures the percentage of clustering algorithm decisions that are correct. There are $N \cdot (N-1)/2$ decisions, one for each of the pairs of documents in the collection, where N is the number of documents. We want to assign two documents to the same cluster if and only if they are similar. A true positive (TP) decision assigns two similar documents to the same cluster; a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit. A false positive (FP) decision assigns two dissimilar documents to the same cluster. A false negative (FN) decision assigns two similar documents to different clusters. The Rand index is simply accuracy (11):

$$RI = \frac{TP + TN}{TP + FP + FN + TN}\ . \quad (11)$$

The Rand index gives equal weight to false positives and false negatives. Separating similar documents is sometimes worse than putting pairs of dissimilar documents in the same cluster. We use the F-measure to penalize false negatives more strongly than false positives by selecting a value $\beta > 1$, thus giving more weight to recall (12):

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN},\ F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}\ . \quad (12)$$

We evaluated PDDP and NMF clustering of full texts and their summaries by proposed NMF-based method on 20 Newsgroups dataset [19]. This is one of the most popular benchmark datasets for clustering. We have chosen the documents which size isn't less than 512 bytes and remove duplicates. As a result we have obtained 15800 documents distributed on 20 predefined clusters.

Experimental results with PDDP and NMF clustering of 20 Newsgroups dataset and its summaries are presented in Table VIII and Table IX. It is worth noting we chose dimensions equal to 5 and 6 for PDDP algorithm which has constructed partitions into 16 and 32 clusters, as a full binary tree of depth 4 and 6, respectively. It is worth noting that PDDP algorithm has constructed partition into 16 clusters, as a full binary tree of depth 4. Differently from PDDP in NMF the required number of clusters is specified as an input parameter. Reductions of the dataset size and the feature space are resulted in the table X.

From the obtained experimental results follows that using summaries instead of full texts slightly improves quality of NMF and PDDP clustering, but significant reduces the size of the processed data.

Method	Normal RI	Summarized RI	Improving RI
PDDP (16 clusters)	0,897759	0,905403	0,8%
PDDP (32 clusters)	0,922439	0,923483	0,1%
NMF (20 clusters)	0,931904	0,937078	0,6%

TABLE VIII. RI CLUSTERING RESULTS

Method	Normal F_2	Summarized F_2	Improving F_2
PDDP (16 clusters)	0,333349195	0,345335028	3,5%
PDDP (32 clusters)	0,282020695	0,283759706	0,6%
NMF (20 clusters)	0,490048293	0,505714559	3,1%

TABLE IX. F_2 CLUSTERING RESULTS

	Initial	Summarized	Improving
Size (words/Mb)	2677083 15,9Mb	1712574 10,3Mb	36% 35,2%
Feature space	80171	62684	21,8%

TABLE X. 20 NEWSGROUPS SIZE AND FEATURE SPACE REDUCTIONS

IV. Conclusion

This paper presents a new generic text summarization method using NMF to estimate sentence relevance. Proposed sentence relevance estimation is based on normalization of NMF topic space (or feature space) and further weighting of each topic using sentences representation in topic space. NMF has the advantage over SVD that it produces a natural “additive parts-based” representation of data, owing to its non-negativity which can be helpful in interpretation of semantic features (topics). The proposed method shows better summarization quality and performance than state of the art methods on DUC 2002 standard dataset.

In addition, we use this text summarization method to replace full text documents by its summary in supervised and unsupervised text classification tasks. Our experiments show applicability of this approach and even improvement of the classification quality of multi-label classification and clustering on benchmark datasets. Therefore, it is possible to draw the following conclusions. The presented method of text summarization defines main topics of documents well. It removes noise and improves classification performance. It is worth to use this method as a preprocessor step in real text mining systems, because the summaries which it produces, are easier to store and process and very informative at once.

As an example of real text mining systems it is possible to adduce system for relevance assessment of research publications in educational research, which was realized within the European Educational Research Quality Indicators (EERQI) project as part of the European Seventh Framework Programme [25]. In EERQI project methods of automatic semantic analysis for the detection of key sentences in a text are used. One of results of their research is that highlighting of key sentences makes it possible to rapidly filter out bad quality: processing the highlighted texts took 4 times shorter time [26].

Also in the EERQI project tested the role of key sentence in relevance ranking. In the EERQI search and query engine, the basic ranking algorithm of the publicly available Lucene search engine was used. They compared the results of this relevance ranking with the list of documents in which the query word(s) occur(s) in key sentences. Lucene uses term frequencies and inverse document frequencies for ranking the retrieved documents. The results show that the top ranked relevant articles returned by Lucene and those selected by their tool are disjoint, which indicates that the two approaches are complementary. Since their tool returns a considerable number of relevant articles that would appear late in Lucene’s ranked list, they consider that this approach is promising and that the integration of the two tools is beneficial for the user [26].

Thus text summarization methods and approaches to their application in classification, clustering and information retrieval tasks are actual research issues.

Acknowledgment

The production of this publication has been made possible through the financial support of the Ministry of Education and Science of the Russian Federation (the state contract #14.514.11.4016) and of the Russian Foundation for Basic Research (the projects #11-07-00616, #12-07-00585).

References

[1] Karel Ježek; Josef Steinberger. “Automatic Text Summarization (The state of the art 2007 and new challenges)”. *In Proceedings of Znalosti 2008*, Bratislava, Slovakia, pp. 1–12, February 2008, ISBN 978-80-227-2827-0.

[2] Document Understanding Conferences, http://duc.nist.gov.

[3] Chin-Yew Lin. “Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough?”. *In: Proc. of NTCIR 2004*, Tokyo, Japan, pp. 1765–1776, 2004.

[4] Yihong Gong, Xin Liu. “Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis”. *In SIGIR-2001*, 2001.

[5] Josef Steinberger, Karel Ježek. “Text Summarization and Singular Value Decomposition”. *In Lecture Notes for Computer Science vol. 2457*, Springer-Verlag, pp. 245-254, 2004.

[6] Ju-Hong Lee, Sun Park, Chan-Min Ahn, Daeho Kim. "Automatic generic document summarization based on non-negative matrix factorization". *Information Processing and Management: an International Journal*, Pages: 20-34, 2009.

[7] Rakesh Peter, Shivapratap G, Divya G, Soman KP. "Evaluation of SVD and NMF Methods for Latent Semantic Analysis". *International Journal of Recent Trends in Engineering*, Vol. 1, No. 3, May 2009.

[8] Daniel Lee, Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". *Nature, 401*, pp. 788-791, 1999.

[9] Susan Dumais. "Improving the retrieval of information from external sources". *In Behavior Research Methods, Instruments & Computers, 23(2)*, pp. 229–236, 1991.

[10] Recall-Oriented Understudy for Gisting Evaluation, http://berouge.com.

[11] Josef Steinberger. "Text Summarization within the LSA Framework". *Doctoral Thesis*, Pilsen, 2007.

[12] Mikhail Petrovskiy. "Paired Comparisons Method for Solving Multi-label Learning Problem". *Proceedings of International Conference on Hybrid Intelligent Systems, Neuro-Computing and Evolving Intelligence*, New Zealand, IEEE Press, 6 pages, 2006.

[13] J. Platt. "Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods". *Adv. in Large Margin Classifiers,* MIT Press, pp. 61–74, 1999.

[14] Mikhail Petrovskiy, Valentina Glazkova. "Linear Methods for Reduction from Ranking to Multilabel Classification". In Lecture Notes for Computer Science *vol. 4304*, Springer-Verlag, pp. 1152-1156, 2006.

[15] Reuters-21578 Text Categorization Collection, http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[16] D.L. Boley. "Principal direction divisive partitioning". *Data Mining and Knowledge Discovery, 2(4)*, pp. 325–344, 1998.

[17] Wei Xu , Xin Liu , Yihong Gong. "Document clustering based on non-negative matrix factorization". *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, July 28-August 01, 2003.

[18] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. "Introduction to Information Retrieval". *Cambridge University Press*, 2008.

[19] The 20 Newsgroups data set, http://people.csail.mit.edu/jrennie/20Newsgroups/.

[20] Chris H. Q. Ding. "A probabilistic model for latent semantic indexing". *In Journal of the American Society for Information Science and Technology, 56(6)*, pp. 597–608, 2005.

[21] I. Mani. and M.T. Maybury. "Advances in Automatic Text Summarization". *Cambridge. MA: The MIT Press*, 442 pp., 1999.

[22] I.V. Mashechkin, M.I. Petrovskiy, D. S. Popov and D.V. Tsarev. "Automatic text summarization using latent semantic analysis". *Programming and Computer Software*, pp. 299-305, 2011.

[23] D.V. Tsarev, I.V. Mashechkev, M.I. Petrovskiy. "Text Summarization Method Based on Normalized Non-Negative Matrix Factorization". *International Conference on Mechanical and Electrical Technology, 3rd, (ICMET-China 2011), Volumes 1–3*, pp.563-567, 2011.

[24] D.V. Tsarev, M.I. Petrovskiy, I.V. Mashechkin. "Using NMF-based text summarization to improve supervised and unsupervised classification". *11th International Conference on Hybrid Intelligent Systems (HIS)*, Malacca, MALAYSIA, pp. 185-189, 2011.

[25] European Educational Research Quality Indicators (EERQI) Project, www.eerqi.eu.

[26] European Educational Research Quality Indicators (EERQI) Project Final Report, http://eerqi.eu/sites/default/files/Final_Report.pdf.

Author Biographies

D. V. Tsarev is a post-graduate student of Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University (MSU). He received his Diploma (2007) in Applied Mathematics and Informatics from Faculty of Computational Mathematics and Cybernetics, MSU, Russian Federation. His primary research areas are: text mining, including text summarization, multi-label classification, cluster analysis.

M. I. Petrovskiy is an associate professor of Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University (MSU). He received his Diploma (1997) and PhD (2003) in Applied Mathematics from Faculty of Computational Mathematics and Cybernetics, MSU, Russian Federation. Since 1999 he has been working at MSU as teacher assistant, assistant professor and currently as associate professor (since 2006). His primary research areas are: statistical learning theory, including kernel methods, fuzzy methods, ensemble learning, robustness; text and data mining, including multi-label classification, ranking, information extraction; and data mining applications, including intelligent intrusion detection, user behavior modeling, keystroke dynamics, mining structured data.

I. V. Mashechkin is a professor of Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University (MSU). He received his Diploma (1978), PhD (1981) and Doctor of Computer Science (1998) in Applied from Faculty of Computational Mathematics and Cybernetics, MSU, Russian Federation. Since 1978 he has been working at MSU as engineer, assistant professor, associate professor and currently as full professor (since 1999). 30 years experience in participation and project management in research and development of IT technologies. Development of CRAB time-sharing system for the BESM-6 Soviet high-performance computer. Development of resource quota and planning system for time-sharing system of the BESM-6 computer. Development of multi-functional high-level languages cross-programming system, based on machine-independent intermediate representation. Currently his primary research areas are: system programming and development of intelligent and data mining systems.