Vector Representation of Context Networks of Latent Topics

Ondrej Hava, Miroslav Skrbek, Pavel Kordik

***Abstract*—Transforming of text documents to real vectors is an essential step for text mining tasks such as classification, clustering and information retrieval. The extracted vectors serve as inputs for data mining models. Large vocabularies of natural languages imply a high dimensionality of input vectors; hence a substantial dimensionality reduction has to be made.**

We propose a new approach to a vector representation of text documents. Our representation takes into account an order of latent topics that generate observed words; an extracted document vector includes information about the adjacency of words in a document. We experimentally proved that the proposed representation enables to build document classifiers of higher accuracy using shorter document vectors. Short but informative document vectors enable to save memory for storing data, to use simpler models that learn faster and to significantly reduce an overfit effect.

***Index Terms*—text mining, document representation, latent Dirichlet allocation, document classification, transition matrix, context window, network centrality measures**

I. Introduction

The process of transformation of text documents to structured vector representation involves the substantial reduction of information. In the bag-of-words approach [1] a document is modeled as a container of vocabulary words where an order of words does not matter. The adjusted frequencies of words are used as features to describe documents [2]. The dimensionality reduction is performed either before the frequency vectors are derived or after that. Natural Language Processing (NLP) methods can help to reduce the size of vocabulary e.g. by stemming or lemmatization [3] of documents before they are transformed to vectors. The methods of linear algebra such as Singular Value Decomposition (SVD) that reduce the dimensionality of matrices can be used to simplify document representation [4]. Even more sophisticated processes of dimensionality reduction are offered by generative models such as Probabilistic Latent Sematic Indexing (pLSI) [5] or Latent Dirichlet Allocation (LDA) [6]. These models assume that words in documents are generated by latent topics. Each document can be described as a mixture of latent topics that generate observed words.

The order of words or the order of latent generative topics is also important in natural languages. If the order of words or topics is projected into a document representation together with their frequencies the consequent modeling methods can take advantage of both. The word adjacency is usually denoted by transition matrices that are well known from language modeling [7]. Unfortunately word or topic transition matrices derived for each document further increase the dimensionality when they are used as a document representation. Such matrices can be considered as networks of words or topics with the weighted connections. The weights are proportional to joint co-occurrence of two words or topics within a short part of a text.

The matrix representation better describes the content of a document [8] but matrix processing is rather time consuming. Moreover data mining models are not usually arranged to accept matrices as their inputs. Hence the adjacency or context matrices have to be downgraded to vectors. If a matrix is considered as a representation of a network any centrality statistics of network vertices [9] can be used as a document representation which comprises both word frequencies and word adjacencies in one vector.

In the second section of the presented paper we briefly summarize the well-known bag-of words document representation. The third section describes the initial reduction of the dimensionality. The words are mapped to latent topics by the means of LDA. The description of our proposed document representation starts in the fourth section. We introduce context windows used to record an order of topics and we propose a matrix representation of a document. In the fifth section we describe an additional dimensionality reduction based on transformation of matrices to vectors using network centrality measures.

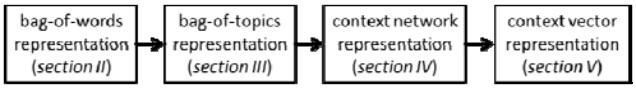

Figure 1 The structure of sections in the paper

Sections six and seven describe a classification experiments with Czech documents and our experimental results. The last section include a discussion about the appropriateness of the proposed document representation.

II. Initial Document Representation

Let us have a training collection C of N text documents D_n, n=1..N.

$$C = \{D_1, D_2, \ldots, D_N\} \quad (1)$$

Each document D_n is represented by a row real vector $\mathbf{d}_n$.

$$\mathbf{d}_n = (d_{n1}, d_{n2}, \ldots, d_{nM}) \quad (2)$$

A vector element d_{nm}, m=1..M, is proportional to a frequency of a term W_m in a document D_n. A set of M terms W_m composes a vocabulary V.

Ondrej Hava is with Faculty of Electrical Engineering, Czech Technical University in Prague.

Miroslav Skrbek and Pavel Kordik are with Faculty of Information Technology, Czech Technical University in Prague.

(Email: hava@acrea.cz, skrbek@fit.cvut.cz, pavel.kordik@fit.cvut.cz)

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$V = \{W_1, W_2, \ldots, W_M\} \quad (3)$$

The terms can be words, phrases, n-grams or some other properties derived from a text. The vocabulary terms are either known in advance or they are derived from the training collection of documents. The whole training collection is then described by a matrix **D** where the documents are organized as row vectors.

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1M} \\ d_{21} & d_{22} & \ldots & d_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ d_{N1} & d_{N2} & \ldots & d_{NM} \end{pmatrix} \quad (4)$$

The process of extraction of features from training documents must be applicable to any new document. The new document is then represented using the same vocabulary terms.

III. Initial dimensionality reduction

The dimensionality of matrix **D** is usually large due the richness of written natural languages. Even when documents are preprocessed by stemming or lemmatization the number of dictionary terms is in the rank of thousands or higher. Latent Dirichlet Allocation (LDA) [6] is a generative statistical model that assumes that observed terms are generated by latent topics T_p, p=1..K. The number of topics K is significantly smaller than the number of dictionary terms (K<<M). The topics form a topic dictionary *Q*.

$$Q = \{T_1, T_2, \ldots, T_K\} \quad (5)$$

The popular EM algorithm [10] enables to derive mapping among terms and topics using the training collection *C*. LDA then enables to estimate a generative topic for each term of a text. The model is applicable to new documents.

Let us have a document *D* (from training collection or the new one) containing the sequence of L terms from the vocabulary *V*. The terms not included in *V* are omitted from the sequence.

$$D = w_{(1)} w_{(2)} \ldots w_{(L)} \quad (6)$$

After the application of LDA the document *D* can be described by a vector of topic probabilities **p**

$$\mathbf{p} = (p_1, p_2, \ldots, p_K), \quad (7)$$

and also by a sequence of distinct topics that generated observed vocabulary terms

$$D = t_{(1)} t_{(2)} \ldots t_{(L)}. \quad (8)$$

The vector of topic probabilities **p** (7) can substitute vector **d** (2), but both vectors do not reflect the order of terms or topics. We will denote the vector **p** as the bag-of-topic representation of a document..

IV. Context mapping

To improve the document representation described in the previous sections we propose to enhance it by information about the order of topics or terms. To record the adjacency of terms we define a context window $R_{(i)}$ for each term $w_{(i)}$ of the document *D*. The context window of term $w_{(i)}$ is a set of terms that follow the term $w_{(i)}$. The context window consists up to S subsequent vocabulary terms.

$$R_{(i)} = \{w_{(i+1)}, w_{(i+2)}, \ldots w_{(i+S)}\} \quad (9)$$

The context window can be shorter than S if there is smaller number of terms after $w_{n(i)}$. It happens at the end of text unity which is considered contextually independent. The text unity can be a sentence, a paragraph or the whole document.

Let us substitute the terms in the context window $R_{(i)}$ by the topics revealed by LDA.

$$R_{(i)} = \{t_{(i+1)}, t_{(i+2)}, \ldots t_{(i+S)}\} \quad (10)$$

The set of all context windows $R_{(i)}$ of the document *D* implies an integer matrix **A** of the size KxK. The element a_{kl} is equal to the number of topics t_l that are included in all context windows for the topic t_k.

$$a_{kl} = \left| \{ t_{(i)} : t_{(i)} = t_l, t_{(i)} \in R_{(j)}, t_{(j)} = t_k \} \right| \quad (11)$$

The matrix **A** represents a document *D* including the information about the order of topics.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1K} \\ a_{21} & a_{22} & \ldots & a_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ a_{K1} & a_{K2} & \ldots & a_{KK} \end{pmatrix} \quad (12)$$

V. Network simplification

The matrix **A** itself represents the document *D* but it is rather difficult to use the matrix as an input for data mining models. The models usually expect a vector profile instead of the matrix profile. Additionally the number of items a_{kl} is KxK which is still high for the processing of larger number of documents fluently. Hence we propose to derive contextually dependent topic properties from the matrix **A**.

The matrix **A** can be viewed as a representation of an oriented weighted network. The topics T_p from topic dictionary *Q* (5) construct vertices and the values a_{kl} are the weights of connections between pairs of vertices. Hence for each document *D* we define its context network *H*=(*Q*, **A**) as a pair of vertices and weighted edges. Context networks of all documents share the same set of vertices *Q* while the weighted edges characterize the content of an individual document. Then the topic centrality measures derived from weight matrix **A** can be used as a KxK-dimensional representation of a document *D*. Such vector representation includes information about the neighborhood of latent topics.

A large number of centrality measures can be derived from a context network. Such measures can be adopted from social network analysis (SNA). They quantify an importance or prominence of vertices in a network using different criteria. Let us briefly name the widely used centrality measures that could be used as scores of topics in the context networks. The exact definitions and the formulas can be found in [9].

- InDegree, OutDegree and Degree are the sums

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

of weights of ingoing, outgoing and all edges of a vertex.

- Farness of a vertex is the sum of the shortest distances to every other vertex. Closeness is the reciprocal of Farness.
- Betweenness measures how often a vertex occurs on shortest paths between all others pairs of vertices.
- Hub and Authority assign scores to vertices based on the concept that connections to high-scoring vertices contribute more to the score than connections to low-scoring vertices. A good hub vertex is one that points to many good authorities and vice versa.
- PageRank is proportional to the number of incoming links and to PageRank of vertices where the incoming links start.

The centrality measures listed above are applicable to directed weighted networks. To use these measures as a final document representation the following transformations should be considered.

Some centrality measures (e.g. Betweenness) rely on path lengths between vertices. A higher weight implies a shorter distance between a connected pair of vertices. The weights have to be transformed to distances before distance based measures are calculated.

The weights in the context network (12) are influenced by the length L of a document *D* because frequencies of adjacent topic pairs are summed over all context windows $R_{(i)}$ (11). To eliminate the document length dependency from the proposed representation the normalized versions of centrality measures should be used. The normalized centrality measures fall within the range <0;1>. The normalized formulas can be found again in [9].

VI. Experimental setup

We tested the proposed document representation for a classification task. We used a collection of 7669 short press releases written in Czech. They were downloaded through RSS feeds from the server ceskenoviny.cz in February and March 2013. The typical length of a document is 1kB. The documents were partitioned to training and test sets randomly by ratio 50:50. The dictionary was derived from the train documents including the words that occurred at least in two documents. The members of stop-word-list and non-linguistic entities such as numbers were filtered out from the dictionary. The final dictionary size is 10692 words.

All documents from the collection were preprocessed by LDA. The words had been substituted by topics before the context networks were derived. We also extracted the topic probabilities for each document to receive the bag-of-topics representation which was used as a reference representation in subsequent comparisons. We tested solutions with 5, 10, 20, 50 and 100 extracted topics in our presented experiments.

The documents were also parsed to sentences. Each sentence was considered as a text unity for a construction of context windows. The partial context window of a term cannot exceed a sentence boundary. We constructed context networks for each document using window lengths of 3, 5 and 10 adjacency vocabulary words.

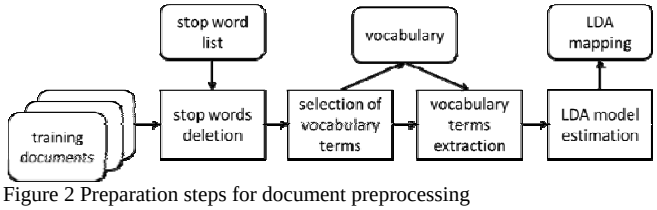

Figure 2 Preparation steps for document preprocessing

Together with the reference bag-of-topic representation we derived document profiles from context networks using the following standardized centrality measures of the topics: Degree, InDegree, OutDegree, Closeness, Betweenness, Authority, Hub and PageRank. The centralities enable to represent the documents by the same number of dimensions as bag-of-topic approach. Additionally we also tried to represent the documents in one-dimensional space using global network statistics: Assortativity and Shortest Path Length.

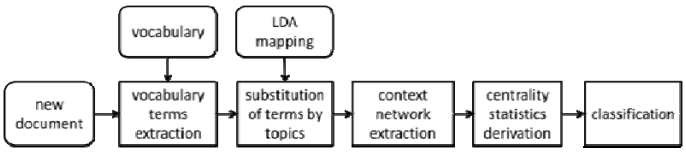

Figure 3 The document processing pipeline

From the total number of 24 categories included in our downloaded collection we selected 6 categories for final classification as depicted in

Table 1. The results presented in the next section were obtained by the Bayesian classifier. The reported measure of quality is F_1-measure (harmonic mean of precision and recall) [11] averaged over all six categories.

Table 1 Document frequencies in selected categories

Category	Count
Culture	292
Weather	72
Domestic news	906
International news	679
Finance	465
Sport	521
Total	**2935**

VII. Experimental results

To evaluate the relevance of the proposed document representation we will compare the averaged F_1 accuracy of any classification to a reference. The Bayesian classifiers using a bag-of-topic representation serve as our reference (Table 2).

Table 2 The reference classification results on the test set. Input vectors include the probabilities of latent topics in documents. The mean of F_1 measure over all categories is used as the evaluation measure.

number of topics	reference F1
5	49%
10	53%
20	52%
50	56%
100	53%

Firstly we tried to compare the classification outcomes if the whole context network is used as document representation. In such representation each connection in a network forms one dimension. The number of possible connections in a network is the number of its vertices squared. Hence the input dimensionality grows rapidly with

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the number of extracted topics. We do not present results for 50 and 100 topics in Table 3 because the classifiers were not able to finish their learning in reasonable time using 2500 and 10000 inputs respectively.

Table 3 The comparison of reference classifiers with classifiers that use network weights as inputs

number of topics	network train	network test	topics train	topic test
5	81%	56%	59%	49%
10	89%	45%	57%	53%
20	95%	35%	65%	52%

The large number of dimensions also implies an overfit effect. There is a significant difference between the results for test and training sets. The higher number of dimensions the difference is larger. The smaller overfit effect was also detected for the reference document representation in Table 3. Unfortunately the overfit effect was observed for the proposed representations as well. An example is depicted in Table 4.

Table 4 The overfit effect for Degree centrality

number of topics	degree train	degree test
5	64%	62%
10	69%	54%
20	84%	58%
50	91%	54%
100	91%	52%

Hence if we need to suspend the overfit we should consider a relatively smaller number of topics. As our experiments confirmed the proposed representations performs well for small number of topics. The comparison of the proposed document representation using different centrality measures is depicted in Table 5.

Table 5 The performance of different centrality measures on test documents

number of topics	degree	indegree	outdegree	closeness
5	62%	60%	62%	28%
10	54%	55%	54%	41%
20	58%	59%	58%	43%
50	54%	53%	52%	45%
100	52%	50%	50%	41%

number of topics	betweenness	authority	hub	pagerank
5	39%	62%	53%	57%
10	41%	57%	56%	54%
20	54%	62%	60%	59%
50	49%	61%	59%	56%
100	45%	56%	56%	53%

Comparing the results on test set with the reference results from Table 2, we can conclude that some centrality measures outperformed the reference solution while some of them degrade the solution. The Authority centrality provided the best results; we recommend to use it together with context networks. Especially in combination with a small number of extracted topics it offers high F_1 together with negligible overfit. On the contrary we don't recommend using context networks together with Closeness and Betweenness centralities. Their performance is significantly worse than the reference.

The tested one-dimensional document representations (Assortativity and Shortest Path Length) mitigate the overfit effect. However as one may expect they do not perform well; the one- dimensional representation of a document is far from sufficient (Table 6). We don't recommend using one-dimensional global statistics derived from context networks.

Table 6 The evaluation of one-dimensional classifiers

number of topics	shortest path train	shortest path test	assortativity train	assortativity test
5	8%	9%	8%	9%
10	19%	20%	8%	7%
20	18%	17%	8%	8%
50	19%	18%	8%	9%
100	8%	8%	8%	8%

All previous comparisons were presented for the length of context window of five words. We performed all test also for different lengths but we conclude that the length of context window has not a significant impact on the classification results. The situation is illustrated in Table 7 where the results for best performing Authority are shown.

Table 7 The length of context window has not significant effect on the accuracy of classifiers. The Authority centrality measure is used in this example.

number of topics	length of context window		
	3	5	10
5	64%	62%	62%
10	57%	57%	56%
20	63%	62%	62%
50	60%	61%	60%
100	56%	56%	56%

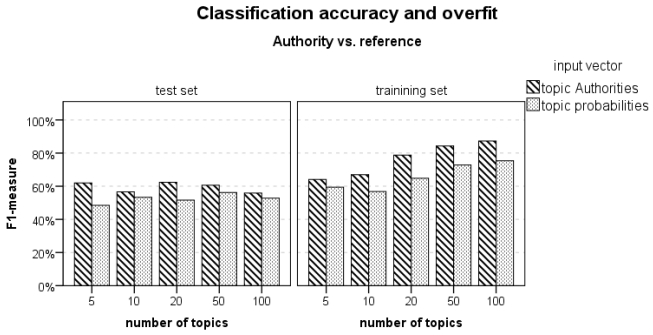

Figure 4 The best centrality measure for transformation of a context network to an input vector is Authority. The performance of Authority is always better than the reference mixture of topics. If it is used with a small number of topics the quality of a classification is high enough and the overfit effect is negligible.

VIII. Conclusions

In the presented paper we proposed a vector representation of documents that contains information about the adjacency of words. Our approach also includes the reduction of dimensionality of input vectors. Instead of words we use latent topics revealed by LDA.

Even though the network representation of documents is more suitable for recording of the adjacency of topic, a representation of documents by vector profiles is the appropriate input for consequent data mining models. To transform the networks into vectors we proposed using several centrality measures that characterize the connectivity of topics in the context networks.

In our classification experiments with different centrality measures we found out that Closeness and Betweenness did not perform well. Other tested centrality measures outperformed the reference bag-of-topic representation at least for the small number of extracted topics. The best performing centrality measure is Authority.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

We also found out that the length of the context window has a minor effect on the performance of tested classifiers regardless of the centrality measure selection. The length of the context window is the number of adjacent topics that were used to build a context network.

The experiments confirmed that the proposed document representation is preferable for a small number of topics. Better performance can be achieved in a classification task using our representation than using the bag-of-topic representation even with a larger number of dimensions. A small number of dimensions also implies a smaller tendency to overfit effect and allows faster processing of new documents by classification models.

In our future work we would like to enhance the document processing pipeline used for deriving a structured representation of documents. We plan to evaluate how language dependent preprocessing of a text such as stemming influences the quality of derived attributes.

REFERENCES

[1] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, 2007.

[2] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing & Management,* vol. 24, no. 5, pp. 513-523, 1988.

[3] D. Jurafsky and J. H. Martin, Speech and Language Processing, Prentice Hall, 2009.

[4] T. Landauer, P. Foltz and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes,* vol. 25, pp. 259-284, 1998.

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, 1999.

[6] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[7] C. Manning a H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[8] O. Háva, M. Skrbek a P. Kordík, „Contextual latent semantic networks used for document classification," v *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, 2012.

[9] S. Wasserman a K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.

[10] A. P. Dempster, N. M. Laird a D. B. Rubin, „Maximum Likelihood from Incomplete Data via the EM Algorithm," v *Journal of the Royal Statistical Society*, 1977.

[11] S. M. Weiss, N. Indurkhya, T. Zhang and F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer, 2005.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)