ALGORITHM FOR CLASSIFICATION OF TEXTUAL DOCUMENTS REPRESENTED BY TANDEM ANALISIS

Jasminka Dobša

Faculty of Organization and Informatics
University of Zagreb

Outline

- Introduction
- Aim of research
- Description of algorithm
- Design of experiment and results
- Conclusion and further work

Introduction

- In term-document matrix documents are represented by terms which occurre in the document
- Sometimes such a representation is not optimal due to the problems of synonymy and polysemy
- Aim of dimensionality reduction of original term-document matrix is to represent a collection of documents in a more compact way which could:
  - save memory space
  - speed up tastks (classification, information retrieval)
  - reduce the effect of noise in the data

Introduction

- Method of latent semantic indexing (LSI) presents benchmark in the field of representation of documents in the space of reduced dimensionality
- LSI has some disadvantages in fulfilling the task of classification since its application could remove some significant information concerning structure of the classes
- In this research dimensionality reduction of the original representation of documents in term-documents matrix is obtained sequentially in two steps in procedure called Tandem analysis

Aim of research

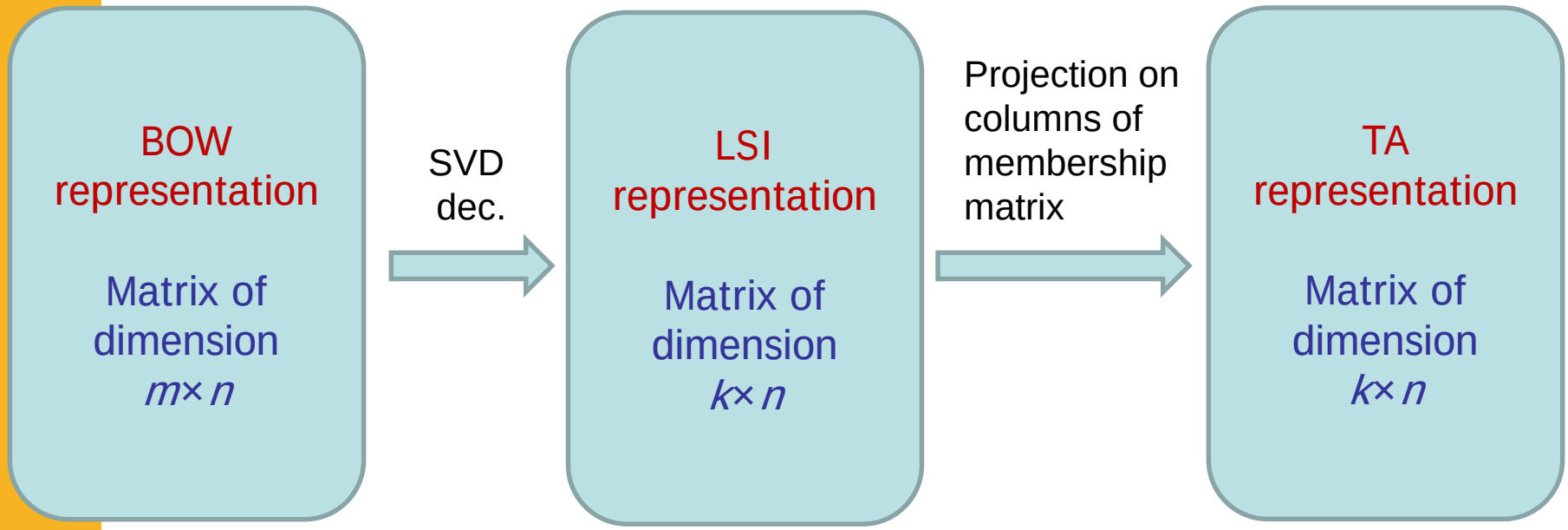
- Introduction of algorithm for classification of textual documents which is carried out in two steps
  - 1. step: classification of textual documents represented in BOW representation
  - 2. step: classification of textual documents represented in space of reduced dimension obtained by Tandem analysis
- Representation of textual documents in the space of reduced dimension is also conducted in two steps:
  - 1. step: representation of documents by method of LSI
  - 2. step: projection of documents represented in LSI space on the space spread by columns of membership matrix which defines membership of documents in classes

Notation

- **A**=[a_{ij}] is term-document matrix
  - **A** is of dimension $m \times n$ where
    - m is number of index terms
    - n is number of documents in collection
- **M**=[m_{ij}] is membership matrix
  - **M** is of dimension $n \times k$ where
    - k is number of classes in collection of documents

$$m_{ij} = \begin{cases} 1 & \text{if document } i \text{ belongs to class } j \\ 0 & \text{otherwise} \end{cases}$$

Dimensionality reduction by Tandem analysis



Why two steps of classification?

- Projection of LSI representations of documents onto column space of membeship matrix **M** is accomplished by solving the least square problem

$$\left\| \mathbf{REP}_{LSI} - \mathbf{MZ} \right\| \rightarrow \min.$$

- A solution to the set problem is given by

$$\mathbf{Z} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{REP}_{LSI}$$

- Representation of documents by Tandem analysis is given by transpose of $n \times k$ matrix

$$\mathbf{B} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{REP}_{LSI}$$

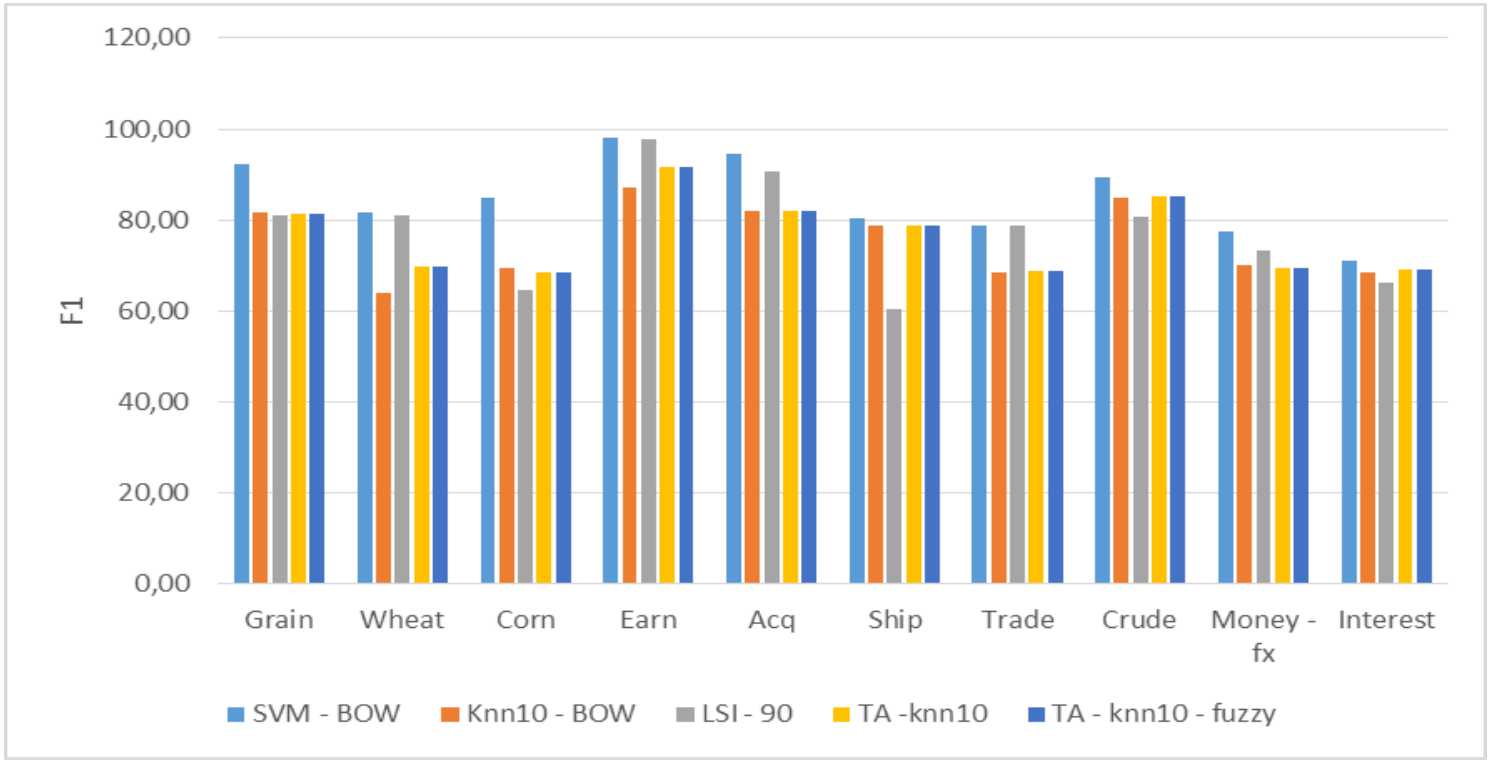
Why two steps of classification?

- For representation of documents of train set it is used membership matrix $\mathbf{M}_{train}$
- Membeship matrix $\mathbf{M}_{test}$ is not allowed to be used in the learning phase
- Matrix $\mathbf{M}_{test}$ is approximated by results of the first step of classification

Design of experiment

- Experiments are conducted on the 10 largest classes of standard Reuters21578 collection
- “ModApte” split is used: 9603 training and 3299 test documents
- List of index terms
  - Stop words are removed
  - Words that occurred in less than 4 documents are removed
  - The list contains 9867 terms
- Classification of documents is conducted by
  - k-nn algorithm for k=10 or SVM algorithm in the first step
  - by SVM algorithm in the second step
- LSI method is conducted for k=90
- Representations obtained by Tandem analysis also use LSI with k=90

Results: 1st step knn10, 2nd step SVM



- Last two columns show F_1 measure for classifiction of documents represented by Tandem analysis
- Best results are obtained for BOW representation
- Representations by Tandem analysis outpreform representation by LSI method for 5 out of 10 classes

Modification: Method TA knn10-fuzzy

- The fifth column shows results of F_1 measure for classification of documents represented by Tandem analysis with modification in comparison to method TA knn10 (Method TA knn10-fuzzy)
- In the first step of classification the decision about membership of a document to a class is not made categoricaly
- Element m_{ik} of a membership matrix contains proportion of 10 nearest documents to i^{th} document contained in k^{th} class
- Modification did not result in improvement of F_1 measure in comparison to method TA knn10

Results: SVM in both steps

	Bag of words			Tandem analysis			Tandem analysis modified		
Class	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Grain	**97.04**	87.92	**92.26**	**97.04**	87.92	**92.26**	85.98	**94.63**	90.10
Wheat	**89.83**	74.65	81.54	**89.83**	74.65	81.54	78.75	**88.73**	**83.44**
Corn	**97.67**	75.00	**84.85**	**97.67**	75.00	**84.85**	81.82	**80.36**	81.08
Earn	**98.43**	97.79	**98.11**	**98.43**	97.79	**98.11**	93.36	**99.54**	96.35
Acq	**97.49**	91.66	**94.49**	**97.49**	91.66	**94.49**	89.14	**98.19**	93.45
Ship	**92.65**	70.79	80.26	**92.65**	70.79	80.26	75.00	**94.38**	**83.58**
Trade	**85.15**	73.50	**78.90**	**85.15**	73.50	**78.90**	62.21	**91.45**	74.05
Crude	**91.21**	87.83	**89.49**	**91.21**	87.83	**89.49**	76.39	**94.18**	84.36
Money - fx	**81.99**	73.74	77.65	**81.99**	73.74	77.65	72.32	**90.50**	**80.40**
Interest	**89.53**	58.78	70.97	**89.53**	58.78	70.97	73.33	**75.57**	**74.43**
Macroaverage	92.10	79.17	84.85	92.10	79.17	84.85	78.83	90.75	84.12

Results of precision, recall and F_1 measure are exactly the same for a BOW representation and representation obtained by Tandem analysis!

Results: SVM in both steps

- Tandem analysis modified:

$$m_{ik} = \begin{cases} 1 & \text{if Prediction} > 0.6 \\ 0 & \text{if Prediction} < -0.6 \\ 0.5 & \text{Otherwise} \end{cases}$$

  - By modification recall is improved for all classes, but at the same time precision is deteriorated for all classes

Conclusion and further work

- Representation of documents by Tandem analysis can improve performance of classification in comaparison to LSI method
- Classification performance is limited by approximation of membership matrix obtained in the first step of classification
- In the further work method will be tested for a task of information retrieval and cross-lingual information retrieval
- Algorithm of Tandem analysis will be compared with algorithm of simultaneous performance of both dimensionality reduction steps (reduction of variables/terms and reduction of objects/documents)