Marcin PIETROŃ, Maciej WIELGOSZ, Michał KARWATOWSKI, Kazimierz WIATR
AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY, 30 Mickiewicza Ave., 30-059 Krakow, Poland
ACC CYFRONET AGH, 11 Nawojki St., 30-950 Krakow, Poland

A study of parallel techniques for dimensionality reduction and its impact on the quality of text processing algorithms

Abstract

The presented algorithms employ the Vector Space Model (VSM) and its enhancements such as TFIDF (Term Frequency Inverse Document Frequency) with Singular Value Decomposition (SVD). TFIDF were applied to emphasize the important features of documents and SVD was used to reduce the analysis space. Consequently, a series of experiments were conducted. They revealed important properties of the algorithms and their accuracy. The accuracy of the algorithms was estimated in terms of their ability to match the human classification of the subject. For unsupervised algorithms the entropy was used as a quality evaluation measure. The combination of VSM, TFIDF, and SVD came out to be the best performing unsupervised algorithm with entropy of 0.16.

Keywords: Singular Value Decomposition, Vector Space Model, TFIDF.

1. Introduction

With the rapid growth of the Internet and other electronic media, the on-line availability of text information has been significantly increased. As a result, the problem of automatic text classification and clustering turns out to be very important because text categorization has become one of the key techniques for handling and organizing data in many applications for industry, entertainment, and digital libraries, which require access and execution of text-based queries. For this purpose, it is often necessary to automatically classify all given texts into predefined classes. Text classification can be used for clustering (creating clusters of texts without any external information or database), information retrieval (retrieving a set of documents that are related to the query), information filtering (rejecting irrelevant documents) and information extraction (extracting the fragments of information, e.g. email addresses, phone numbers, etc.). Possible applications include such tasks as: email spam filtering, organization of web-pages into hierarchical structures, product review analysis, text sentiment mining, organization of papers according to a subject class, and categorization of newspaper articles into topics [5], [6].

Over the past years, several methods of automatic text categorization have been implemented and examined. For instance, in [5] the authors use unsupervised learning for automatic text categorization. Their method does not rely on creation of the training documents by hand, rather it automatically creates a training set using the keyword lists of each category and uses them for classification. The authors use the χ^2 statistic as a feature selection method and the naive Bayes classifier. In [6] the authors use the text summarization techniques to determine whether a sentence is important or not. Two possible text summarization methods are applied: the first measures the sentence importance by the similarity with the title of the document, and the other estimates the importance of the terms in each sentence. Another useful application is the text classification of online reviews. It covers opinion mining on: products [11], movies [10], and political situations [9]. In [7] text categorization is also applied to detection of public opinion from blog posts. The proposed method uses the lexicon based approach combined with the machine learning. The text sentiment mining is proposed in [8] and [4].

2. System description

The system (Fig. 1) consists of the Internet data retrieval module, the text extraction module, text preprocessing, and the set of methods based on TFIDF and SVD for the text classification and clustering [1, 2]. The key for successful implementation of the proposed methods is the construction of the reliable corpus for the area and topic of interest detection. At the moment, the topic of interest detection corpus is being built automatically. The news articles are retrieved from a news portal [3]. The downloaded articles are already classified by journalists, which makes it possible to use them as a reference in selected topic and area of interest.

Figure 1 shows the generic flow of the proposed algorithms. The adopted methods are applied to all texts stored in the repository. The first step involves preprocessing activities which are performed in the following order: removal of all special and redundant characters (e. g. colons, brackets, numbers, etc.), lemmatization, and filtering all stop-words. We use the Vector Space Model which requires conversion of each text into a numerical representation. Thus, the second step consists of symbols (words) converting into numerical values that can be used in the subsequent processing stages. In the next step, a selection of the appropriate classification method is performed. The choice depends on which hypothesis is to be verified. For verification of the first hypothesis an unsupervised learning model is adopted and for the second one supervised learning is use. In the case of supervised classification, a model (corpus) is built based on all texts in the repository. For the unsupervised procedure, the VSM and TFIDF vectors are mapped to the reduced space, which is built with the SVD algorithm.

The data dimensionality reduction process is crucial in many data and text mining algorithms and systems. It can tremendously reduce the efficiency of data analysis. There are few popular methods for data dimensionality reduction (e.g. SVD, Random Projection). In our paper we concentrate on singular value decomposition. The singular value decomposition (SVD) is a factorization of a real or complex matrix to $\boldsymbol{U} \times \boldsymbol{\Sigma} \times \boldsymbol{V}$, where $\boldsymbol{U}$ is the matrix of the left-singular vectors, $\boldsymbol{V}$ is the matrix of the right singular vectors and $\boldsymbol{\Sigma}$ is the diagonal matrix with singular values. The SVD factorization is strictly related to the Principal Component Analysis (PCA) algorithm with finding iteratively components with maximal variance.

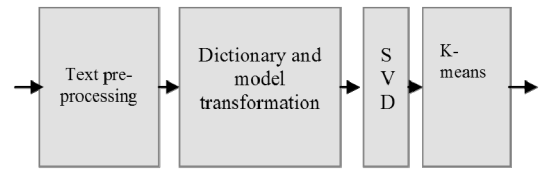


Fig. 1. Block diagram of the system

3. Experiments

In order to verify the adopted hypothesis, several experiments were conducted. All subsequent experiments use the same repository which contains 4200 texts divided into five categories: business, culture, automotive, science and sport. The texts were downloaded from Polish language news website.

For quality evaluation the following metrics are used: precision (1), recall (2), and F-measure (3). They are defined accordingly:

$$\boldsymbol{precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp+fp}} \quad (1)$$

$$\boldsymbol{recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp+fn}} \quad (2)$$

$$\boldsymbol{F} = \boldsymbol{2} \times \frac{\boldsymbol{precision \times recall}}{\boldsymbol{precision+recall}} \quad (3)$$

where *tp* stands for the number of true positives, *fp* is "false positive" and *fn* denotes the number of false negatives. If the assigned category is the same as the one which was predefined by a human, it is a true positive case. A false positive is a case when the texts from other categories are assigned to the category under consideration. For instance, for "business" category, a false positive is a text from "sport" assigned to "business" category. A false negative occurs when a text from the category under consideration is assigned to other class (e. g. "business" text is assigned to "culture" category). One of the most popular clustering quality measure is entropy. Thus the authors decided to employ it in the system. The entropy of a cluster $\boldsymbol{C_i}$ with a size $\boldsymbol{n_i}$ is defined by equation (4):

$$\boldsymbol{E(C_i)} = -\sum_{\boldsymbol{h=1}}^{\boldsymbol{k}} \frac{\boldsymbol{n_i^h}}{\boldsymbol{n_i}} \boldsymbol{\log}\left(\frac{\boldsymbol{n_i^h}}{\boldsymbol{n_i}}\right) \quad (4)$$

where *k* is the total number of categories in the data set and $\boldsymbol{n_i^h}$ is the number of documents from the h^{th} class that were assigned to the cluster $\boldsymbol{C_i}$. The entropy of the cluster will be zero for the ideal cluster with documents only from a single category. In general, the smaller the entropy value, the better the quality of the cluster. Moreover, the average entropy of the overall solution is defined to be the weighted sum of the individual entropy values of each cluster, which is given by equation (5)

$$\boldsymbol{Entropy} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \frac{\boldsymbol{n_i}}{\boldsymbol{n}} \boldsymbol{E(C_i)} \quad (5)$$

where $\boldsymbol{n_i}$ denotes the number of files in each cluster, *n* is the total number of files, and $\boldsymbol{C_i}$ is defined by equation (4).

The experiment involved combination of the VSM, TFIDF and K-means algorithms. In case of the SVD algorithm, the choice of a number of singular values is critical for the performance of the algorithm and should be carefully considered. In this implementation, the authors decided to use only the first two SVD components.

Tab. 1. Precision, recall and f-measure for SVD and K-means for 2 singular values

	number of clusters	Precision	recall	F-measure
business	3.9(0.3)	0.81(0.022)	0.56(0.077)	0.66(0.034)
culture	3(0)	0.37(0.015)	0.7(0.061)	0.48(0.024)
automotive	4.8(0.4)	0.39(0.007)	0.56(0.021)	0.45(0.01)
science	2.1(0.3)	0.39(0.014)	0.74(0.016)	0.51(0.014)
sport	4.8(0.4)	0.39(0.007)	0.56(0.021)	0.45(0.01)

The computation of the presented combination of algorithms, Tab. 1, is the most demanding of all presented in this work. It absorbed 31 GB of RAM memory and took 12 hours. The authors decided to compare all the algorithms covered in this section in terms of entropy (see Tab. 2 and Fig. 2). The results meet the expectations, i.e. the usage of more advanced and sophisticated methods such as TFIDF and SVD rather than pure Vector Space Model, improves the entropy, which is the overall measure of the clustering quality. However, it is worth emphasizing that the presented solutions may be significantly improved by optimization of the parameters of the algorithms such as, e.g., the number of singular values.

Tab. 2. Entropy of the employed clustering methods

employed algorithms	entropy
vsm+kmeans	0.28(0.012)
vsm+tfidf+kmeans	0.17(0.019)
vsm+tfidf+svd+kmeans	0.16(0.006)

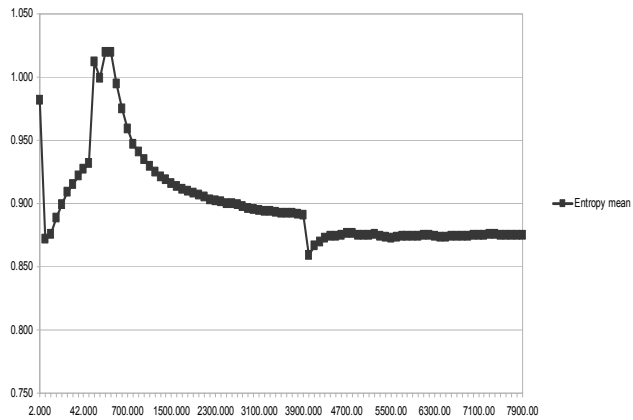


Fig. 2. Entropy for a different computed number of singular values

As it is mentioned above, Gensim SVD computation took about ten hours for our corpus. Therefore GPU parallel NIPALS algorithm was implemented. The NIPALS algorithm is an iterative algorithm. The sequence of iterations must be preserved. The single iteration is only a part of the algorithm which can be parallelized. There are few linear algebra operations (e.g. vector and matrixes multiplications) inside iteration. In our implementation, CUBLAS library was used for parallelization of this section of the code. The results are presented in Tab. 3.

Tab. 3. Times of the NIPALS algorithm implemented in GPU and CPU for a different number of singular values

reduction size	GPGPU [ms]	CPU [ms]
10	33	80
20	77	305
30	107	420
40	161	624

4. Conclusions and future work

Our work shows that dimensionality reduction in text mining can significantly improve computational complexity of algorithms while maintaining the accuracy level. Additionally, the presented method is scalable and can be speed up on parallel hardware platforms. Future work will concentrate on adding to our system other data dimensionality reduction algorithms, especially such as the Random Projection algorithm, and compare the system effectiveness and accuracy with the existing implemented algorithms.

This research is supported by the European Regional Development Program no. POIG.02.03.00-12-137/13 PL-Grid Core.

5. References

[1] Russek P., Pietroń M, Żurek D, Janiszewski M, Wiatr K., Jamro E., Wielgosz M: Implementation of algorithms for fast text search and

files comparison. Proceedings of the High Performance Computer Users Conference KU KDM 2013, pp. 83-84. Academic Computer Centre Cyfronet AGH, Academic Computer Centre Cyfronet AGH, 2013.
[2] Janiszewski M., Pietroń M., Russek P., Jamro E., Dabrowska-Boruch A., Wiatr K., Wielgosz M., Koryciak S.: Parallel mpi implementation of n-gram algorithm for document comparison. ACACES 2013 : the 9th international summer school on Advanced Computer Architecture and Compilation for High-performance and Embedded Systems, pages 217-220, 2013.
[3] Interia.pl. http://interia.pl
[4] Niaz~Arifin S.M., Dasgupta S.: Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. Proceedings of the 21st international Computational Linguistics Conference, pp. 611-618, 2006.
[5] Seo J., Ko Y.: Automatic text categorization by unsupervised learning. Proceedings of the 18th international conference on computational linguistics}, pages 453-459, 2000.
[6] Seo J., Ko Y., Park J.: Improving text categorization using the importance of sentences. Information Processing and Management, pp. 65-79, 2004.
[7] Boughanem M., Saad Missen M.M.: Using wordnet's semantic relations for opinion detection in blogs. In Advances in Information Retrieval, vol. 5478, Lecture Notes in Computer Science, pp. 729-733. Springer Berlin, Heidelberg, 2009.
[8] Ghose A. K., Polpinij J.: An ontology-based sentiment classification methodology for online consumer reviews. In Proceedings of the IEEE international conference on Web Intelligence and Intelligent Agent, pp. 518-524, 2008.
[9] Smith M. D., Durant K. T.: Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In Advances in Web Mining and Web Usage Analysis, vol. 4811, Lecture Notes in Computer Science, pp. 187-206. Springer Berlin, Heidelberg, 2007.
[10] Chunping L. Zhao L.: Ontology based opinion mining for movie reviews. In Knowledge Science, Engineering and Management, vol. 5914, Lecture Notes in Computer Science, pp. 204-214. Springer Berlin, Heidelberg, 2009.
[11] Montoyo A., Balahur A.: A feature dependent method for opinion mining and classification. In Proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering, pp. 1-7.

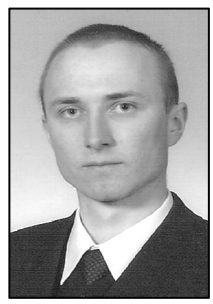
Received: 21.04.2015 *Paper reviewed* *Accepted: 02.06.2015*

Marcin PIETROŃ, PhD

He received MSc degree in electronic engineering and in computer science in 2003 and PhD in 2013 from the AGH University of Science and Technology, Kraków, Poland. He currently works in Academic Computing Centre CYFRONET AGH and University of Science and Technology. His research interests include parallel computing, automatic parallelization and data mining.

e-mail: pietron@agh.edu.pl

Maciej WIELGOSZ, PhD



He received MSc and PhD degrees in electronic engineering in 2005 and 2010 respectively from the AGH University of Science and Technology, Kraków, Poland. He currently works in Academic Computing Centre CYFRONET AGH and University of Science and Technology. His research interests include hardware acceleration, text mining and hardware architectures for artificial intelligence.

e-mail: wielgosz@agh.edu.pl

Michal KARWATOWSKI, MSc, eng.

He received the BSc Eng. and MSc degrees in electronic engineering in 2013 and 2014 respectively from the AGH University of Science and Technology, Kraków, Poland. Currently PhD student. His research interests include usage of hardware accelerators in complex computations, control and energy efficient systems of small and big scale, mainly using Field-Programmable Gate Arrays.

e-mail: mkarwat@agh.edu.pl

Prof. Kazimierz WIATR, DSc, eng.

He received the MSc and PhD. degrees in electrical engineering from the AGH University of Science and Technology, Kraków, Poland, in 1980 and 1987, respectively, and the DSc degree in electronics from the University of Technology of Łódź in 1999. Received the professor title in 2002. His research interests include design and performance of dedicated hardware structures and reconfigurable processors employing FPGAs for acceleration computing. He currently is a director of Academic Computing Centre CYFORNET AGH.

e-mail: wiatr@agh.edu.pl