Using NMF for Analyzing War Logs

Dirk Thorleuchter [1] and Dirk Van den Poel [2]

[1] Fraunhofer INT, Appelsgarten 2, D-53879 Euskirchen, Germany
dirk.thorleuchter@int.fraunhofer.de
[2] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium
dirk.vandenpoel@ugent.be, http://www.crm.UGent.be

Abstract. We investigate a semi-automated identification of technical problems occurred by armed forces weapon systems during mission of war. The proposed methodology is based on a semantic analysis of textual information in reports from soldiers (war logs). Latent semantic indexing (LSI) with non-negative matrix factorization (NMF) as technique from multivariate analysis and linear algebra is used to extract hidden semantic textual patterns from the reports. NMF factorizes the term-by-war log matrix - that consists of weighted term frequencies – into two non-negative matrices. This enables natural parts-based representation of the report information and it leads to an easy evaluation by human experts because human brain also uses parts-based representation. For an improved research and technology planning, the identified technical problems are a valuable source of information. A case study extracts technical problems from military logs of the Afghanistan war. Results are compared to a manual analysis written by journalists of 'Der Spiegel'.

Keywords. Non-negative matrix factorization, NMF, Text Mining

1 Introduction

War logs written by soldiers during mission of war are a valuable source of information. They indicate e.g. technical problems occurred by armed forces weapon systems in use. Considering some of these problems in current research and technology (R&T) projects may be necessary for an increase reliability of future weapon systems. Thus, extracting this feedback from war logs is an important task in R&T planning.

We provide a methodology for a semi-automated identification of technical problems in soldiers' war logs. A manual identification of these problems e.g. by human experts is not possible because of the large amount of the logs. Although war logs describe the events of the war, technical problems are just a part of the content e.g. an event is described in detail and besides the malfunction of a weapon system during that event is also mentioned. A frequently occurred malfunction of a specific system in different events can be discovered by identifying the underlying (hidden) semantic textual patterns from the collection of war logs because different soldiers formulize malfunctions by using different words. This excludes the use of text classification

algorithms based on knowledge structure approaches (e.g. Support Vector Machine, Decision trees) for this identification because they do not consider the aspects of meaning and thus, the identification of hidden semantic textual patterns.

Matrix factorization techniques consider the aspects of meaning [1]. NMF is a matrix factorization technique that can be used for text mining [2]. This algorithm is proposed to identify parts of textual documents [3]. The used parts-based representation is similar to the representation of information in human brain as shown by psychological and physiological studies [4-6]. This makes the results of NMF more comprehensible for a human expert than results of other matrix factorization techniques for text mining e.g. Singular Value Decomposition (SVD) [7]. Based on a term-by-war log matrix of weighted term frequencies, NMF factorizes this m x n matrix $A = [a_{ij}]$ into two non-negative matrices: $U = [u_{ij}]$ and $V = [v_{ij}]$ with m the number of war logs, n the number of different terms, and $r = \text{rank}(A) \leq \min(m,n)$. U represents the (m x r) matrix that shows the similarity of terms and hidden semantic textual patterns. V is the (n x r) matrix that shows the similarity of hidden semantic textual patterns and war logs. Because of many zero values in the matrixes, the rank r can be reduced to $k < r$ with a compressed approximation $A \approx UV^T$ as calculated by

$$a_i \approx \sum_{j=1}^{k} u_j v_{ij} \tag{1}$$

with u_j be the j-th column vector of U. The vectors a_i of A are approximated by the weighted column vectors of U. Thus, a small number of vectors of U is used to approximate a large number of vectors of A. A good approximation can only be archived if the vectors of U discover structure that is latent in the data [8].

2 Methodology

The provided data are textual information contained in a collection of war logs. This unstructured information is prepared by removing specific characters and tags and by correcting typographical errors [9]. Tokenization is used to separate the different terms and all terms are converted in lower case [10], [11]. The number of terms is reduced by part-of-speech tagging, by stop word filtering, by stemming, and by applying Zipf distribution. For each war log a term vector is created based on vector space model [12]. The vector components consist of weighted term frequencies [13]. All vectors are used to create a term-by-war log matrix A. NMF is used to find two non-negative matrices U and V where the product of U and V provides a good approximation to A and where the rank is reduced from r to k. The selection of k is critical [7]. If k is too large then too many column vectors of U exists that represent many irrelevant or unimportant latent semantic textual information. If k is too small then the product of U and V does not provide a good approximation to A and thus, the column vectors of U are not a parts-based representation of the data. We apply a parameter-selection procedure [14] by constructing several rank-k models [15]. A fivefold cross-

validation [16] is used to measure the approximation to A for each rank-k model and the lowest k is selected where the approximation to A is acceptable as defined by a specific threshold [17]. The k column vectors of U represent the latent semantic textual information in form of a parts-based representation. To present the vectors in a comprehensible way to human experts [18], terms are ordered by their corresponding vector components that represent the impact of the terms on a latent semantic textual pattern. As a result, a list of keywords is created for each pattern to support the identification of technical problems in the war logs.

3 Evaluation

In a case study, we use the released US military logs of the Afghanistan war (Kabul War Diary) from January 2008 to December 2009 as published by WikiLeaks.org. The data consists of 42,374 events and for each event a detailed description is given. These descriptions are used in the case study and the methodology is applied as described in Sect. 2. Based on the parameter-selection procedure, k is set to 100 and lists of the 100 latent semantic textual patterns are presented to human experts. To compare the performance of the methodology, we apply a second methodology on the data. The second methodology uses SVD instead of NMF because SVD as matrix factorization technique also can be used for text mining. However, SVD does not use a parts-based representation. Thus, we can measure the effect of parts-based representation on human experts' success for identifying technical problems. To enable comparison, the SVD parameter k is also set to 100 and lists of these 100 latent semantic textual patterns are also presented to human experts.

Human experts identify three patterns from NMF algorithm and two patterns from SVD algorithm where a strong reference to technical problems occurs. Technical problems of ground vehicles, trucks, clinger, convoys etc. as well as hydraulic problems of aircrafts can be identified from patterns of both, NMF and SVD. Further, an NMF pattern shows that in several events, small hand-launched remote-controlled unmanned aerial vehicles have crashed. Many reasons for this can be found in the corresponding list of keywords. Technical reasons are from general system faults via computing errors up to sensor failures. Further reasons are human failures by controlling or the shot down by insurgents. These results are confirmed by 'Der Spiegel' [19] where journalists have analyzed the data manually. They also find out e.g. that very often, unmanned aerial vehicle crash in events by the reasons mentioned above. Further, the human experts say that the NMF results are more comprehensible than SVD results. Thus, the proposed methodology can be used to identify technical problems from war logs semi-automatically, and in this case, the use of NMF outperforms the use of SVD. Further work should focus on a more detailed evaluation.

References

1. Cai, D., He, X., Wu, X., Han, J.: Non-negative Matrix Factorization on Mani-fold. In: 8th IEEE International Conference on Data Mining, pp. 63–72, IEEE Press, New York (2008)

2. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2), 111–126 (1994)
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
4. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual Review of Neuroscience 19, 577–621 (1996)
5. Palmer, S.E.: Hierarchical structure in perceptual representation. Cognitive Psychology 9, 441–474 (1977)
6. Wachsmuth, E., Oram, M.W., Perrett, D.I.: Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. Cerebral Cortex 4, 509–522 (1994)
7. Thorleuchter, D., Van den Poel, D., Prinzie, A.: Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Syst. Appl. 39 (3), 2597–2605 (2012)
8. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562. MIT Press, Cambridge, MA (2001)
9. Thorleuchter, D., Van den Poel, D.: Companies Website Optimising concerning Consumer's searching for new Products. In: International Conference on Uncertainty Reasoning and Knowledge Engineering, pp. 40–43. IEEE Press, New York (2011)
10. Thorleuchter, D., Van den Poel, D.: Improved Emergency Management by a Loosely Coupled Logistic System. In: Future Security 2012. CCIS. Springer, Berlin (2012) in press
11. Thorleuchter, D., Van den Poel, D.: Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website. Expert Syst. Appl. (2012) in press
12. Thorleuchter, D., Herberz, S., Van den Poel, D.: Mining Social Behavior Ideas of Przewalski Horses. In: Wu, Y. (ed.) 3CA 2011. LNEE, vol. 121, pp. 649–656. Springer, Berlin (2012)
13. Thorleuchter, D., Van den Poel, D.: Extraction of Ideas from Microsystems Technology. In: Jin, D., Lin, S. (eds.) CSIE 2012. Advances in Intelligent and Soft Computing, vol. 168, pp. 563-568. Springer, Berlin (2012)
14. Thorleuchter, D., Van den Poel, D.: High Granular Multi-Level-Security Model for Improved Usability. In: 2nd International Conference on System science, Engineering design and Manufacturing informatization, pp. 191–194. IEEE Press, New York (2011)
15. Thorleuchter, D., Weck, G., Van den Poel, D.: Usability based Modeling for Advanced IT-Security - an Electronic Engineering approach. In: International Conference on Mechanical and Electronic Engineering 2012. LNEE, Springer, Berlin (2012) in press
16. Thorleuchter, D., Van den Poel, D.: Granular Deleting in Multi Level Security Models - an Electronic Engineering approach. In: International Conference on Mechanical and Electronic Engineering 2012. LNEE, Springer, Berlin (2012) in press.
17. Thorleuchter, D., Van den Poel, D.: Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships. In: Annual SRII Global Conference 2012. IEEE Press, New York (2012) in press
18. Thorleuchter, D., Van den Poel, D.: Rapid Scenario Generation with Generic Systems. In: International Conference on Management Sciences and Information Technology 2012. Lecture Notes in Information Technology. IERI, Delaware (2012) in press
19. Gebauer, M., Goetz, J., Hoyng, H., Koelbl, S., Rosenbach, M., Schmitz, G.P.: Die Afghanistan-Protokolle. Der Spiegel 30, 70–86 (2010)