Using Non-Negative Matrix Factorization for Text Segmentation

Aliya Nugumanova[1], Madina Mansurova[2], Yerzhan Baiburin[1], and Yermek Alimzhanov[2]

[1] D. Serikbayev East Kazakhstan State Technical University, Oskemen, Kazakhstan,
[2] Al-Farabi Kazakh National University, Almaty, Kazakhstan
yalisha@yandex.kz,mansurova01@mail.ru,{ebaiburin,aermek81}@gmail.com

Abstract. The aim of this paper is to investigate whether non-negative matrix factorization (NMF) can be useful for semantic segmentation of large full-text documents. NMF is a universal technique that decomposes the monolithic structure of a massive dataset into different trends. In case of textual data these trends can be interpreted as topics. Thereby NMF can associate each document with topics covered in it, however, without linking topics to the certain parts of that document. In this paper, we complement this traditional NMF technique with a new goal: for a given full-text document we build a semantic map which links document's parts with topics covered in it.

Keywords: non-negative matrix factorization, text segmentation, topic modeling.

1 Introduction

Text segmentation is a very interesting challenge in the field of natural language processing. It arises in many information retrieval applications providing users with quick access to document repositories. Since full-text documents stored in such repositories are usually large to read and analyze, information retrieval applications should be able to divide them into chunks and deliver the most relevant chunks to users in accordance to their requests [1].

In this paper, we focus on the task of segmentation of full-text documents from a topic modeling perspective. In recent years, topic modeling is gaining momentum in data mining in general [2], and in particular in the text segmentation field [3]. Recent work in this field has shown that using topic distribution over documents instead of term distribution can significantly increase segmentation performance [4,5,6,7,8,9].

The segmentation mechanism drawn from topic modeling is very simple. At first, each document is divided on small segments (e.g., sentences or paragraphs). At second, topics covered in this document are revealed, and each word in each segment is associated with one topic; thereby, for each segment topic occurrences are defined. At last, adjacent document's segments sharing a certain number of common topics are merged into topical chunks.

One of the most popular approaches to topic modeling is non-negative matrix factorization (NMF). In general, NMF is a well-recognized technique due its ability to extract relevant structures of data and may thus contribute to a deeper understanding of data behavior [10]. This technique, being applied to a collection of full-text documents, maps it into a space of topics. For example, in Fig. 1, NMF is applied to a co-occurrence matrix of a collection, which consists of 5 documents. As we can see from the figure, after matrix factorization, document 1 is represented as a combination of topic 1 and topic 3. Simultaneously, topic 3 is represented as a combination of term 2, term 4 and term 5, and term 5 is the most significant for this topic.

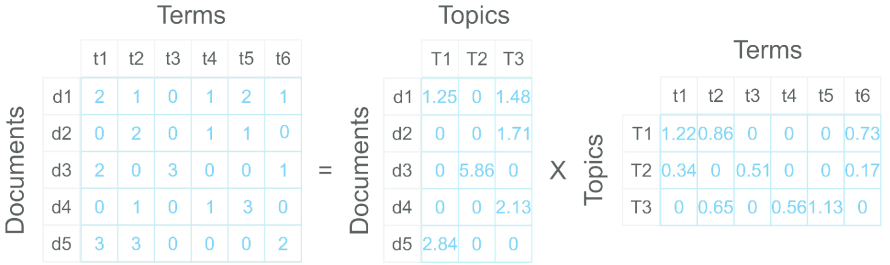

Fig. 1. A sample of non-negative matrix factorization.

As we can see, NMF has two useful applications. Firstly, for each document it defines the most weighted topics, which we call relevant topics. Secondly, for each topic it finds the most weighted terms, which we call support terms. In this paper, we use support terms for the semantic segmentation of full-text documents. Our contribution is to complement the traditional NMF representation with a new goal: the creation of a semantic map of the given document through using support terms as map's nodes. We suppose that, by linking these nodes to the corresponding parts of the document, we achieve its smart and comfortable segmentation.

The rest of this paper is organized as follows. In Section 2, we discuss previous work on text segmentation and explain our reasons to use NMF. In Section 3, we present proposed approach. In particular, we demonstrate how we link support terms with the document parts, and present some experimental results. In Section 4, we formulate conclusions and plans for our future work.

2 Related Work

The most simple and intuitive algorithm of text segmentation is TextTiling [11]. It uses a sliding window to move through a document and capture text blocks (tiles). The similarity between consecutive blocks are calculated on the base of cosine metrics. The calculated values are used to draw a similarities curve that

tracks topics changes between consecutive blocks so that the segment boundaries are chosen at the local minima of the curve. The main disadvantage of TextTiling is low accuracy because of the sparsity of text blocks.

Another simple algorithm of text segmentation is C99 [12]. At first, it divides the input document into minimal blocks (sentences) and for each block calculates its rank based on the blocks similarities. Then it performs divisive clustering starting with the whole document and splitting it to parts in accordance with blocks' ranks. In [13] C99 algorithm is improved by applying Latent Semantic Analysis for calculating the blocks' similarity matrix.

C99 algorithm also is used in [1]. This work addresses the issue of providing topic driven access to full-text documents. Authors of this work apply C99 to subdivide documents into smaller thematically homogeneous parts that can be used as link targets. They try to perform segmentation as accurate as possible: document parts should be of such sizes that "shrinking them would cause relevant information to be left, and expanding them would bring in too much non-relevant information" [1]. However, they concentrate only on the segmentation phase without details of designing a whole navigation system.

As we have mentioned in the introduction, a considerable line of research explores text segmentation methods based on topic modeling. The most popular algorithms for topic modeling are Latent Dirichlet Allocation (LDA) [3], [8,9,10] and Non-negative Matrix Factorization (NMF) [10,11]. Although output of LDA is very similar to the output of NMF, these models are fundamentally different in nature: LDA is based on a Bayesian probabilistic model; whereas NMF is based on algorithms of linear algebra that fit root mean squared error. As it's shown in [14], both LDA and NMF can discover concise and coherent topics and demonstrate similar performance, however NMF learns more incoherent topics than LDA. Authors of [15] also compare LDA and NMF, and conclude that NMF better than LDA "from the perspectives of consistency from multiple runs and early empirical convergence".

We choose NMF here because of its basis sparseness [16]. Basis sparseness means that NMF uses less basis features (terms) than LDA. This makes NMF topics more overlapped, i.e. more semantically related to each other than LDA ones (see an example represented by Table 1). We consider that these relations are essential for the understanding of how the document's semantic map should be organized.

Fig. 2 gives a visual interpretation of Table 1. The four topics are shown in Fig.2, and some of them are linked with other topics through specific key terms.

3 Proposed Approach

Our method of text segmentation consists of 5 steps. Firstly, we should subdivide a given full-text document into units and build units-by-terms co-occurrence matrix. Secondly, we should define a reasonable number of topics (K) and apply NMF to factorize the co-occurrence matrix and obtain 2 matrices: units-by-topics and topics-by-terms. Thirdly, for each extracted topic we should sort topic terms

Table 1. Number of topics intersected by basis terms in "Geology" text collection

#	Basis term		Number of topics to which this term is assigned	
	in Russian	in English	through NMF	through LDA
1	горный	mining	3	1
2	порода	rock	3	1
3	поверхность	surface	2	1
4	склон	slope	2	1
5	процесс	process	2	2
6	динамика	dynamics	2	1
7	система	system	2	1
8	зона	zone	2	1
9	состав	composition	2	1
10	кора	crust	2	2

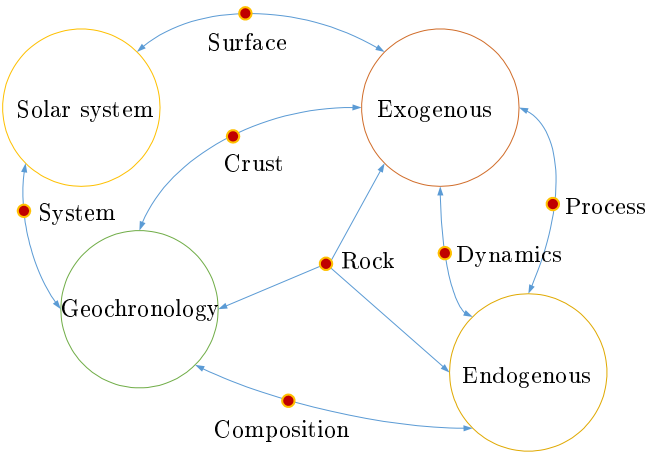

Fig. 2. The visual interpretation of overlapped topics of "Geology" text collection.

by their weights and select only the most weighted terms (no more than 10% of all terms). We call these terms support terms.

In our case we have used "Geology" textbook written in Russian [17], and divided it into 89 units (by number of chapters). Then we have chosen K=5 and extracted 5 topics and 500 support terms. By the way we have proposed information about term distribution over these 5 topics to a geology expert for analysis. Based on the information the expert has concluded that it is best to name extracted topics as "Endogenous", "Solar system", "Oceans", "Exogenous" and "Geochronology". Tables 2-6 represent top 10 support terms for each of 5 extracted topics.

Table 2. Support terms for topic 1 "Endogenous" (top 10)

#	Through NMF		Through LDA	
	Russian	English	Russian	English
1	порода	rock	движение	motion
2	процесс	process	геологический	geological
3	движение	movement	скорость	speed
4	минерал	mineral	землетрясение	earthquake
5	землетрясение	earthquake	метод	method
6	горный	mining	время	time
7	внутренний	internal	процесс	process
8	образование	formation	год	year
9	метаморфизм	metamorphism	возраст	age
10	динамика	dynamics	Земля	Earth

Table 3. Support terms for topic 2 "Solar system" (top 10)

#	Through NMF		Through LDA	
	Russian	English	Russian	English
1	Земля	Earth	Земля	Earth
2	магнитный	magnetic	система	system
3	планета	planet	магнитный	magnetic
4	солнечный	solar	планета	planet
5	поле	field	верхний	upper
6	система	system	солнечный	solar
7	ядро	core	поле	field
8	поверхность	surface	поверхность	surface
9	солнце	Sun	средний	average
10	атмосфера	atmosphere	температура	temperature

Table 4. Support terms for topic 3 "Oceans" (top 10)

#	Through NMF		Through LDA	
	Russian	English	Russian	English
1	океан	ocean	океан	ocean
2	зона	zone	зона	zone
3	континентальный	continental	глубина	depth
4	континент	continent	континентальный	continental
5	пояс	belt	континент	continent
6	глубина	depth	волна	wave
7	платформа	platform	пояс	belt
8	разлом	rift	мощность	power
9	кора	crust	дно	bottom
10	мощность	power	подводный	underwater

Table 5. Support terms for topic 4 "Exogenous" (top 10)

#	Through NMF		Through LDA	
	Russian	English	Russian	English
1	вода	water	вода	water
2	процесс	process	процесс	process
3	порода	rock	динамика	dynamics
4	экзогенный	exogenous	внешний	external
5	внешний	external	экзогенный	exogenous
6	динамика	dynamics	материал	material
7	материал	material	склон	slope
8	склон	slope	поверхность	surface
9	выветривание	erosion	озеро	lake
10	поверхность	surface	подземный	underground

Table 6. Support terms for topic 5 "Geochronology" (top 10)

#	Through NMF		Through LDA	
	Russian	English	Russian	English
1	кора	crust	порода	rock
2	химический	chemical	минерал	mineral
3	минерал	mineral	горный	mining
4	земной	terrestrial	процесс	process
5	желтый	yellow	химический	chemical
6	элемент	element	кора	crust
7	порода	rock	состав	composition
8	верхний	upper	земной	terrestrial
9	метод	method	образование	formation
10	таблица	table	являться	to be

The fourth step is the most important in our method. We should associate our units with topics taking into account topics' support terms. If we use only the traditional NMF representation we miss opportunities to exploit the distributional power of support terms.

For example, let's consider the Unit #30 in the given Geology textbook. The unit describes the history of glaciations as well as the impact of glaciers on the Earth's crust (see Table 7). So the geology expert has associated this unit with the topic "Exogenous" as well as with the topic "Geochronology". In contrast, the traditional NMF algorithm evaluates highly the relation of this unit with the topic "Exogenous" and very lowly the relation with the topic "Geochronology". But if one analyses the support terms used in the Unit #30, one finds that the topic "Geochronology" is well represented in the unit with the help of support terms such as "period", "year", "history", "epoch", "time", "Holocene", "cycle" etc.

Table 7. Topics distribution over the Unit #30 of the Geology textbook [17, p.318-319]

Unit #30	История оледенений. Ледниковые эры, периоды, эпохи. В истории Земли неоднократно возникали великие оледенения, при которых площади ледниковых покровов возрастали до десятков миллионов квадратных километров. Интервалы времени длительностью в миллионы лет с характерными для них похолоданием климата и разрастанием оледенений получили название ледниковых периодов. Последний из них, продолжающийся до сих пор и называемый плейстоценовым, или четвертичным, начался 2,5-3 млн лет назад			
#	NMF representation		Expert representation	
	Topic	Weight	Topic	Mark
1	Exogenous	125.99	Exogenous	5
2	Solar system	93.70	Geochronology	5
3	Oceans	92.30	Oceans	3
4	Geochronology	10.56	Solar system	0
5	Endogenous	0	Endogenous	0

Therefore, in order to more accurately define topics for each document's unit we should complement the traditional NMF approach by analyzing support terms distribution in this unit. We should analyze next 3 factors:

1. What support terms related to the certain topic are occurred in this unit?
2. How frequently they are occurred?
3. How important are they for the certain topic (how many their weights in the topic)?

As a result we should decide can we associate this unit with some support terms and with some topics represented via these terms. The decision rule can be summarizes as follows:

$$Decision = \begin{cases} 1, \sum_{st \in Topic \cap Unit} tf(st, Unit) * weight(st, Topic) \geq Th \\ 0, \sum_{st \in Topic \cap Unit} tf(st, Unit) * weight(st, Topic) < Th, \end{cases}$$

where st is a support term related to the topic, $tf(st, Unit)$ is its frequency in the unit, $weight(st, Topic)$ is its weight in the topic, Th is a threshold value above which the topic is recognized as related to the unit. In this paper we set $Th = 0.1$.

At the fifth step, we construct a navigation (semantic) map that contains three layers: layers of document units, layers of support terms and layers of topics. In Fig. 3 a part of the Geology textbook's map is illustrated. This map consists of 5 top-level nodes which correspond to textbook topics and 500 middle-level nodes which correspond to support terms. Middle-level nodes can be moved up or down or rolled up and stored away until one activates them again. Active middle-level nodes point out the related units which are bottom-level nodes. If middle level nodes are rolled up, access to bottom-level units is enable directly through top-level nodes.

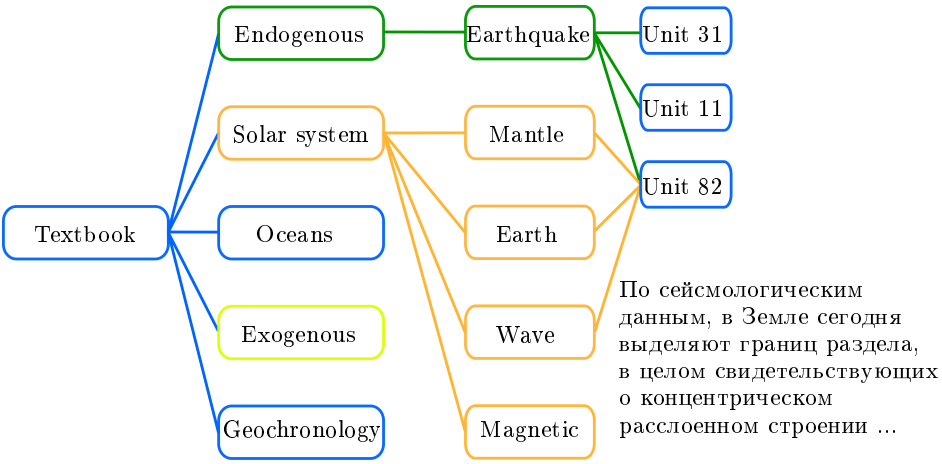

Fig. 3. A fragment of the Geology textbook semantic map.

4 Discussion

So, the main advantage of NMF in comparison with LDA is a great "naturalness", i.e. the topics in NMF are more widely intersect with each other by a number of representative terms. LDA tries to generate topics in such a way so

that intersect as less as possible, as a result, each topic has its own set of "exclusive" words. In practice, such a result does not look natural because in each document there is a number of general thematic terms which refer to the subject in a whole and can not belong only to one definite topic within the subject. For example, in geology such terms are the words "rock", "process", "dynamics", etc. NMF associates such terms with several topics, while LDA tries to assign each such term to only one topic, as a result its stability suffers. In the new series of experiments, a set of words referred by LDA to one topic can significantly differ from the set of words referred to the same topic in the previous series of experiments. In our experiments, NMF proved to be a more stable method than LDA. The topical dispersion of representative key words does not hinder segmentation of the document, on the contrary, it enhances segmentation not isolating segment from each other but connecting them. The disadvantage of NMF is its computing complexity. But this is the problem of computing technologies, not of the method itself.

5 Conclusion And Future Work

In this paper we considered semantic map as a tool of smart document's segmentation and organization. We presented an approach to automatic creation of semantic maps and showed how this process can benefit from a new interpretation of NMF based on the concept of support terms. Also we performed a little case study to illustrate proposed approach. However, more work should be done to evaluate advantages and disadvantages of this approach, to substantiate choice of the NMF parameters (e.g. the number of reasonable topics, or start number of units, or threshold for topic validation). Efforts must also be devoted to a complete comparison of proposed approach with LDA and NMF.

Acknowledgments. This work was supported in part under grant of Foundation of Ministry of Education and Science of the Republic of Kazakhstan "Development of intellectual high-performance information-analytical search system of processing of semi-structured data" (2015-2017).

References

1. Caracciolo, C., Van Hage, W., De Rijke, M.: Towards topic driven access to full text documents. In: Research and Advanced Technology for Digital Libraries. LNCS, 8th European Conference, ECDL 2004, Bath, UK, September 12-17, 2004, Proceedings. pp. 495–500. Springer, Heidelberg (2006)
2. Xing K. et al.: Adi: Towards a framework of app developer inspection. In: International Conference on Database Systems for Advanced Applications. Springer International Publishing. pp. 266–280. (2014)
3. Blei D. M. Probabilistic topic models. J. Communications of the ACM. Vol. 55, issue 4, pp. 77-84 (2012)

4. Riedl M., Biemann C. Text segmentation with topic models. Journal for Language Technology and Computational Linguistics. Vol. 27, issue 1, pp. 47–69 (2012)
5. Misra H. et al.: Text segmentation via topic modeling: an analytical study. In: 18th ACM conference on Information and knowledge management, ACM, pp. 1553–1556 (2009)
6. Du L., Buntine W. L., Johnson M.: Topic Segmentation with a Structured Topic Model. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies HLT-NAACL, pp. 190–200 (2013)
7. Kazantseva A., Szpakowicz S.: Measuring Lexical Cohesion: Beyond Word Repetition. In: International Conference on Computational Linguistics COLING, pp. 476–485 (2014)
8. Du L., Pate J. K., Johnson M.: Topic models with topic ordering regularities for topic segmentation. In:2014 IEEE International Conference on Data Mining. IEEE, pp. 803–808 (2014)
9. Sun Q. et al.: Text segmentation with LDA-based Fisher kernel. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pp. 269–272 (2008)
10. Frigyesi A., Hoglund M.: Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. J. Cancer informatics, Vol. 6, pp. 275-Ц292 (2008)
11. Hearst M. A.: TextTiling: Segmenting text into multi-paragraph subtopic passages. J. Computational linguistics. Vol. 23, issue 1. 33–64 (1997)
12. Choi F.: Advances in domain independent linear text segmentation. In: 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics, pp. 26–33 (2000)
13. Choi F., Wiemer-Hastings P., Moore J.: Latent semantic analysis for text segmentation. In: EMNLP (2001)
14. Stevens K. et al.: Exploring topic coherence over many models and many topics. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, pp. 952–961. (2012)
15. Choo J. et al.: Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. J. Visualization and Computer Graphics, IEEE Transactions on. Vol. 19, issue 12, pp. 1992–2001 (2013)
16. Hu F., Hao Q. (ed.): Intelligent sensor networks: the integration of sensor networks, signal processing and machine learning. CRC Press (2012)
17. General Geology: in 2 vols. Edited by Professor L.K. Sokolovsky. Moscow: KDU (in Russian) (2006)