Topic Modeling for Segment-based Documents

(Extended Abstract)[⋆]

Giovanni Ponti[1], Andrea Tagarelli[2], and George Karypis[3]

[1] ENEA - Portici Research Center, Italy
[2] Department of Electronics, Computer and Systems Sciences,
University of Calabria, Italy
[3] Department of Computer Science & Engineering, Digital Technology Center,
University of Minnesota, Minneapolis, USA

Abstract. Statistical topic models have traditionally assumed that a document is an indivisible unit for the generative process, which may not be appropriate to handle documents that are relatively long and show an explicit multi-topic structure. In this paper we describe a generative model that exploits a given decomposition of documents in smaller, topically cohesive text units, or segments. The key-idea is to introduce a new variable in the generative process to model the document segments in order to relate the word generation not only to the topics but also to the segments. Moreover, the topic latent variable is directly associated to the segments, rather than to the document as a whole. Experimental results have shown the significance of the proposed model and its better support for the document clustering task compared to other existing generative models.

1 Introduction

In recent years, there has been a growing interest towards *statistical topic models* [8, 2, 18, 12, 10, 16], which assume that a document can be represented as a mixture of probability distributions over its constituent terms, where each component of the mixture refers to a main topic. The document representation is obtained by a generative process, i.e., a probabilistic process that expresses document features as being generated by a number of (latent) variables. Unlike conventional vector-space text models, topic models are able to involve latent semantic aspects underlying correlations between words to leverage the structure of topics within a document. This ability becomes particularly relevant when documents explicitly belong to multiple topical classes, and the different topics are discussed at different parts of the text, which is frequent in real-world datasets.

However, classic generative models for documents like PLSA [8] and LDA [2] are not really able to capture topic correlations. A major reason behind this limitation is that they rely on the bag-of-words assumption, which allows for keeping the model's computational complexity acceptable, but also incorrectly assumes independence among the word-topics in the document.

In this work, we present a *Segment-based Generative Model* (SGM) which allows for alleviating the limitations due to the bag-of-words assumption in the context of

[⋆] A full version of this paper appeared at the 14th Int. Conf. on Discovery Science (DS), Espoo, Finland (2011) (multi-topic) documents by exploiting the underlying composition of documents into topically coherent text blocks, or *segments*. Unlike other existing generative models, term generation in SGM is related not only to topics but also to segments. As a consequence, the latent variable that models topics is being associated to the within-document segments, rather than to the document as a whole. In addition, although this model will continue to treat each segment as a bag-of-words, the word-to-topic assignments will be contextualized w.r.t. the various segments, thus generating proper topic distributions for each term according to the segment in which the term occurs.

We evaluated the effectiveness of our generative model in a document clustering task. Experiments conducted on multi-topic document collections have shown that our segment-based approach to document generative modeling improves document clustering performance w.r.t. the other competing models. Moreover, clustering of topically-segmented documents based on our generative model has shown to outperform a traditional document clustering approach in which segments are represented based on the conventional vector-space model.

2 Related Work

The problem of identifying a topic feature space in a given document collection has been originally addressed by mapping the term-document representation to a lower-dimensional latent "semantic" space. Following this line, one of the earliest methods was *Probabilistic Latent Semantic Analysis* (PLSA) [8], in which the conditional probability between documents and terms is modeled as a latent variable. An extension of PLSA, called *Ext-PLSA* [10], has also been proposed to specifically support document clustering. *Latent Dirichlet Allocation* (LDA) [2] is a corpus-oriented model, since the generative process consists of a three-level scheme that involves the whole collection, the documents, and the words in each document. Since exact inference in LDA is not tractable, a number of approximate inference approaches have been developed. Moreover, although possessing a consistent generative semantics, LDA is not able to capture correlations among topics, since the topic proportions as derived from a Dirichlet distribution are substantially independent.

Text segmentation is concerned with the fragmentation of an input text into smaller units (e.g., paragraphs) each possibly discussing a single main topic. Regardless of the presence of logical structure clues in the document, linguistic criteria and statistical similarity measures have been mainly used to identify thematically-coherent, contiguous text blocks in unstructured documents (e.g., [7, 1, 5]). The *TextTiling* algorithm [7] is the exemplary similarity-block-based method, which has been successfully used in several application domains for retrieval purposes. TextTiling is able to subdivide a text into multi-paragraph, contiguous and disjoint blocks that represent passages, or subtopics.

To the best of our knowledge, there are only a few studies that address topic modeling and text segmentation in a combined way [4, 15, 13]. The key idea is generally to improve the performance of text segmentation algorithms under the assumption that topic segments tend to be lexically cohesive and a switch to a topic corresponds to a shift in the term distribution. Our proposal differs from these methods significantly, since it does not define a new topic-based segmentation approach. Rather, we design a document generative model specifically for topically-segmented documents. Thus,

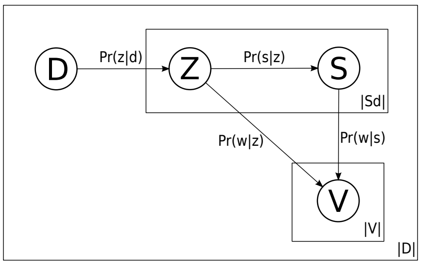

Fig. 1. Plate-based graphical model representation of SGM. The outer plate represents documents (d), whereas the inner plates represent the repeated choice of topics (z) and segments (s) and of words (w) within a document

being able to involve terms as well as text segments in a document in the generative process, our approach aims to lead to a finer-grained identification of topic distributions in the document generative process.

In the same direction as ours, the STM model [6] also exploits the availability of document segments in the generative process. It substantially extends LDA by introducing a further level to represent the document segments. Although our SGM and STM are both generative models that handle document segments, they are quite different in terms of latent variable dependences (STM is a generative model for a corpus, whereas SGM is able to generate model topics for a single document independently of the others in the collection), and of generative process complexity (a four-level model for STM w.r.t. a two-level one for SGM).

3 Segment-based Generative Model

We are given a collection of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$ and a set of words $\mathcal{V} = \{w_1, \ldots, w_M\}$, which represents the vocabulary of $\mathcal{D}$. Each document $d \in \mathcal{D}$ is a sequence of n_d words. We denote with $\mathcal{Z} = \{z_1, \ldots, z_T\}$ the set of hidden topics, where $\mathcal{Z}$ represents a latent variable model that associates topics (unobserved class variables) with word occurrences (observed data). We suppose that each document $d \in \mathcal{D}$ is provided as a set S_d of contiguous, non-overlapping text blocks, or *segments*, and that such segments are obtained by some text segmentation method (cf. Section 2). However, we do not make any particular assumption about the document segmentation strategy (provided that it is in principle coherent to the topical structure of the documents) and the algorithmic choices of the specific text segmentation method used.

Figure 1 illustrates the graphical model representation of SGM, by which nodes correspond to variables and boxes are *plates* representing replicates of the enclosed variables. SGM utilizes one latent variable $\mathcal{Z}$ to model topic distributions, whereas the model variable $\mathcal{S} = \{S_1, \ldots, S_N\}$ is used to represent document segments. The generative process performed by SGM on a corpus $\mathcal{D}$ of segmented documents can be summarized as follows:

1. Select a document d from $\mathcal{D} \Rightarrow \Pr(d)$
2. For each segment $s \in S_d$:
   a) Choose a topic z for the document $d \Rightarrow \Pr(z|d)$
   b) Associate topic-to-segment probability to the segment s for the selected topic z $\Rightarrow \Pr(s|z)$

Table 1. Datasets used in the experiments

dataset	*size (#docs)*	*#words*	*#topic-labels*	*avg #topic-labels per doc*	*#topic-sets*	*avg #docs per topic-set*
IEEE	4,691	129,076	12	4.56	76	61.72
PubMed	3,687	85,771	15	3.20	33	111.73
RCV1	6,588	37,688	23	3.50	49	134.45

c) For each word w in the segment s:
- Choose a word w from the current topic and segment $\Rightarrow \Pr(w|z,s)$

The above generative process can be translated into a joint probability model for triadic data, in which each observation is expressed by a triad defined on documents, segments, and words:

$$\Pr(d,s,w) = \Pr(d) \sum_{z \in \mathcal{Z}} \Pr(z|d) \Pr(s|z) \Pr(w|z,s)$$

Model parameter estimation is accomplished by the Expectation Maximization (EM) algorithm. Since SGM has one latent variable ($\mathcal{Z}$) that models the document topics, the E-step consists in estimating the posterior probabilities of $\mathcal{Z}$ given the known model variables:

$$\Pr(z|d,s,w) = \frac{\Pr(z,d,s,w)}{\Pr(d,s,w)} = \frac{\Pr(z|d) \Pr(s|z) \Pr(w|z,s)}{\sum_{z \in \mathcal{Z}} \Pr(z|d) \Pr(s|z) \Pr(w|z,s)}$$

The M-step aims to maximize the expected value of the log-likelihood, $\mathbf{E}[\mathcal{L}]$, which is computed as

$$\mathbf{E}[\mathcal{L}] = \sum_{d \in \mathcal{D}} \sum_{s \in S_d} \sum_{w \in \mathcal{V}} n(d,s,w) \times \sum_{z \in \mathcal{Z}} \Pr(z|d,s,w) \log(\Pr(d,s,w))$$

where $n(d,s,w)$ is the number of occurrences of word w in the segment s of a given document d. Note that the above formula takes into account only the relevant part of the log-likelihood function, since it is trivial to estimate $\Pr(d)$ as proportional to $\sum_{s \in S_d} \sum_{w \in \mathcal{V}} n(d,s,w)$.

4 Evaluation and Results

4.1 Methodology

We conducted an experimental evaluation aimed at assessing the impact of using the SGM representation of documents on the performance of a *document clustering* task. Clustering documents with an inherent multi-topic structure is traditionally accomplished by a *soft* (e.g., fuzzy) clustering method to produce overlapping clusters of documents. However, we pursue the idea that the particular document representation offered by generative models can enable simpler (i.e., *hard*) clustering schemes. Since the generative process produces a topic distribution for each document in the corpus (i.e., $\Pr(z|d)$), documents are represented as *probability mass functions* (pmfs) that are defined over a feature space underlying topics. This topic-feature space is usually lower-dimensional than conventional term-feature space, and is identified by a mixture model of the topic distributions for any given document.

To perform document clustering, we used a centroid-based-linkage agglomerative hierarchical algorithm for clustering document pmfs, which was developed in our earlier work [11]. In the algorithm, the notion of prototype (centroid) of a cluster is defined as a mixture that summarizes the pmfs of the documents within that cluster. Moreover, the cluster merging criterion, which decides the pair of clusters to be merged at each step, utilizes the Hellinger distance to compare the cluster prototypes. Note that the Hellinger distance can be viewed as the information-theoretic counterpart of the popular cosine similarity, since it is derived from the Bhattacharyya coefficient [9] which represents the cosine between any two vectors that are composed by the square root of the probabilities of their mixtures.

For the experimental evaluation, we used three collections of multi-topic documents belonging to different application domains, whose main characteristics are summarized in Table 1. To perform document segmentation, we used TextTiling (cf. Section 2) and set its parameters around the values suggested in [7], by varying the token-sequence size around ± 10 of the default 20 and the text unit size from 3 to 15. We finally selected three configurations, corresponding to the minimum, the average, and the maximum segmentation level (i.e., number of segments produced); we will use symbols SGM^{min}, SGM^{avg}, and SGM^{max} to refer to instances of SGM applied to these three segmentation schemes, for a given document collection.

We adopted an external cluster validity approach, in order to assess how well a document clustering solution fits the *topic-set-based reference classification* for a given dataset. We derived such reference classifications by exploiting the availability of topic-labels in each dataset: since topic distributions identify the set of covered topics in each document, documents that are clustered together tend to have similar profiles of their mixtures of topics. We call a *topic-set* θ a subset of topics in $\mathcal{Z}$ that is entirely covered by at least one document. Topic-sets are regarded as sets of topic-labels that may overlap, whereas documents are kept organized in disjoint groups. Therefore, the assignment of topic-sets to documents allows for inducing a multi-topic, hard classification for the documents in a given dataset, which can be exploited as a reference classification for clustering evaluation purposes. The last two columns of Table 1 report on statistics about the topic-sets that were identified on each of the evaluation datasets, with a coverage of at least 20 documents per topic-set. As an example of topic-set construction, consider a set of documents $\mathcal{D} = \{d_1, \ldots, d_7\}$ and a set of topic-labels $\mathcal{Z} = \{z_1, \ldots, z_5\}$ in $\mathcal{D}$. Suppose that an external document labeling information produces an assignment of each document in $\mathcal{D}$ with a subset of topics in $\mathcal{Z}$ as follows: $d_1 \leftarrow \{z_3, z_5\}, d_2 \leftarrow \{z_1, z_4\}, d_3 \leftarrow \{z_1, z_2, z_5\}, d_4 \leftarrow \{z_1, z_4\}, d_5 \leftarrow \{z_3, z_5\}, d_6 \leftarrow \{z_1, z_4\}, d_7 \leftarrow \{z_1, z_2, z_5\}$. Three distinct topic-sets are hence present in $\mathcal{D}$, i.e., $\theta_1 = \{z_3, z_5\}, \theta_2 = \{z_1, z_4\}, \theta_3 = \{z_1, z_2, z_5\}$, which correspond to a 3-class partition of $\mathcal{D}$: $\{\{d_1, d_5\}, \{d_2, d_4, d_6\}, \{d_3, d_7\}\}$.

To compare clustering solutions and reference classification, we resorted to three widely used criteria in document clustering, namely *F-measure* (*F*), *Entropy* (*E*), and *Normalized Mutual Information* (*NMI*). Higher (resp. lower) values of *F* and *NMI* (resp. *E*) correspond to better clustering quality.

4.2 Results

We present here document clustering results obtained by SGM and the selected competing models, namely PLSA, LDA and Ext-PLSA. The generative processes of the various models were set in such a way that the topic variable assumed the same number of values as the number of topic-labels given for each dataset. Ext-PLSA also required a further latent variable related to the size of the desired clustering solutions. For this

Table 2. SGM-based clustering performance on IEEE with different segmentations

segmentation setting	*#segments*	*F*	*E*	*NMI*
SGM^{avg}	155,828	0.64	0.58	0.49
SGM^{min}	89,539	0.59	0.62	0.45
SGM^{max}	179,491	0.58	0.60	0.47

Table 3. Summary of clustering results

	F				*E*				*NMI*			
	PLSA	Ext-PLSA	LDA	**SGM**	PLSA	Ext-PLSA	LDA	**SGM**	PLSA	Ext-PLSA	LDA	**SGM**
IEEE	0.53	0.56	0.46	0.64	0.70	0.73	0.62	0.58	0.37	0.32	0.44	0.49
PubMed	0.48	0.50	0.43	0.58	0.57	0.54	0.49	0.42	0.50	0.52	0.58	0.64
RCV1	0.49	0.54	0.42	0.56	0.57	0.59	0.51	0.48	0.49	0.46	0.54	0.59
avg score	*0.50*	*0.53*	*0.44*	*0.59*	*0.61*	*0.62*	*0.54*	*0.49*	*0.45*	*0.43*	*0.52*	*0.57*
avg gain	*+0.09*	*+0.06*	*+0.16*	—	*+0.12*	*+0.13*	*+0.05*	—	*+0.12*	*+0.14*	*+0.05*	—

study, we carried out the algorithms on CRESCO HPC system,[4] which is integrated into ENEA-GRID infrastructure. CRESCO is a general purpose system composed by 382 nodes with more than 3300 cores. We executed our experiments on a CentOS 5.5 platform, with Linux 2.6.18 kernel, 64GB memory, 4 Intel(R) Xeon(R) CPU E7330, 2.40GHz quadcore [3].

We initially investigated how clustering performance based on our SGM depends on the selected segmentation strategy. For this purpose, we tested SGM on the evaluation datasets by providing it with different input segmentations, namely SGM^{max}, SGM^{min}, and SGM^{avg}. We report results only for a selected dataset, as conclusions drawn from the remaining datasets were very similar to those here presented. Table 2 shows clustering results obtained on IEEE. In the table, we can observe that neither minimizing nor maximizing the number of segments improved the clustering accuracy obtained based on SGM^{avg}. Nevertheless, a higher number of segments would seem to be preferable to a smaller one. In fact, SGM^{max} achieved a little gain over SGM^{min} in terms of *E* and *NMI* (both around 0.02), while being comparable on *F* based evaluation. This can be explained since more segments would lead to discover (sub)topics that are hierarchically related to the main ones but also would tend to overfit the data, as the occurrences of any specific word will be diluted over the many segments and, consequently, such a topic-word over-specificity will correspond to more topic distributions.

Table 3 summarizes clustering results, where we used SGM^{avg}. A first evident remark is that our SGM led to the best clustering quality results. In fact, improvements in F-measure varied from 0.06 (vs. Ext-PLSA) to 0.16 (vs. LDA). Major improvements in terms of F-measure obtained by SGM were observed on IEEE and PubMed (above 0.08 on average better than the best among the competing models), whereas on RCV1 the performance gain was lower (about 0.02). This would indicate that benefits from text segmentation in document generative modeling are more evident for relatively long documents than short ones. Moreover, looking at the performance based on the other quality measures, average quality gains achieved by our SGM were quite similar to those previously discussed in terms of F-measure. In particular, SGM outperformed the other methods in terms of entropy, from 0.05 (vs. LDA) to 0.13 (vs. Ext-PLSA). In terms of *NMI*, quality improvements were from 0.05 (vs. LDA) to 0.14 (vs. Ext-PLSA). Comparing the performance of the competing methods, LDA outperformed both PLSA and

[4] http://www.cresco.enea.it/

Table 4. Comparison with traditional VSM-based document clustering

	SGM-based clustering			VSM-based clustering		
dataset	*F*	*E*	*NMI*	*F*	*E*	*NMI*
IEEE	0.64	0.58	0.49	0.21	0.84	0.21
PubMed	0.58	0.42	0.64	0.31	0.79	0.28
RCV1	0.56	0.48	0.59	0.39	0.63	0.45
avg score	*0.61*	*0.49*	*0.57*	*0.30*	*0.75*	*0.31*

Ext-PLSA according to entropy and *NMI* (up to 0.11 *E* and 0.12 *NMI* both on IEEE), whereas Ext-PLSA behaved better than the other two methods in the case of F-measure evaluation (up to 0.12 *F* on RCV1). This would suggest that by using LDA clustering solutions tend to be less coarse than those obtained by PLSA and Ext-PLSA—because F-measure is typically biased towards coarser clustering.

Comparison with traditional document clustering. We also compared the performance achieved by clustering the segmented documents based on our SGM with an approach that first performs the clustering of the segments from the document collection (by treating each segment as a single mini-document), and finally derives a document clustering solution. For this purpose, in the baseline method segments were represented by the conventional vector-space model (VSM) equipped with the popular *tf.idf* term relevance weighting scheme. Clustering of the segments was performed by using the *Bisecting K-Means* [14] algorithm, which is widely known to produce high-quality, hard clustering solutions in high-dimensional, large datasets [17]. The segments that belong to the documents in the various collections were represented as conventional *tf.idf* vectors prior to inputting them to the clustering algorithm. Since the partitioning of the segment collection produced by Bisecting K-Means corresponds to a potentially soft clustering of the documents, we devised a simple method to derive a hard assignment of documents to clusters by adopting a majority voting strategy (i.e., each document is assigned to the cluster that contains the majority of its segments). Finally, the document clustering solution derived by this approach was evaluated w.r.t. the reference classification based on topic-sets for any specific dataset.

Table 4 summarizes results of this comparative analysis. SGM^{avg}-based clustering always outperformed the VSM-based clustering on all datasets, achieving average improvement per dataset of 0.31 *F*, 0.26 *E*, and 0.26 *NMI*. By modeling segmented documents, SGM was indeed able to directly produce a hard document clustering that corresponds to a finer mapping of documents to topic-sets, which is well-suited for better reflecting the multi-topic nature of documents. Conversely, by treating segments that belong to same document as independent text units to be clustered, the baseline document clustering approach tends to produce solutions whose document clusters are likely to be biased by those topics that are present in most of the segments within the same document.

5 Conclusions

We presented a generative model for topically-segmented documents, which introduces a segment model variable in the generative process. The topics of a document in a given collection are modeled as a mixture of the individual distributions of the topics present in each of the document segments. In this way, the bag-of-words assumption (which is typically exploited in statistical topic modeling) becomes more realistic since it is transferred to smaller text units, i.e., the document segments. As experimentally proved, the topic modeling obtained on the within-document segments is better suited for documents that have an explicit multi-topic class structure.

References

1. D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. *Journal of Machine Learning Research*, 34(1-3):177–210, 1999.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
3. G. Bracco, S. Podda, S. Migliori, P. D'Angelo, A. Quintiliani, D. Giammattei, M. De Rosa, S. Pierattini, G. Furini, R. Guadagni, F. Simoni, A. Perozziello, A. De Gaetano, S. Pecoraro, A. Santoro, C. Scio', A. Rocchi, A. Funel, S. Raia, G. Aprea, U. Ferrara, D. Novi, and G. Guarnieri. CRESCO HPC System Integrated into ENEA-GRID Environment. In *Proc. of the Final Workshop of the Grid Projects of the Italian National Operational Program 2000-2006 – Call 1575*, pages 151–155, 2009.
4. T. Brants, F. Chen, and I. Tsochantaridis. Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. In *Proc. 11th ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 211–218, 2002.
5. F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent Semantic Analysis for Text Segmentation. In *Proc. Int. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 109–117, 2001.
6. L. Du, W. L. Buntine, and H. Jin. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learning*, 81(1):5–19, 2010.
7. M. A. Hearst. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64, 1997.
8. T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196, 2001.
9. T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, 1967.
10. Y. M. Kim, J. F. Pessiot, M. R. Amini, and P. Gallinari. An Extension of PLSA for Document Clustering. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1345–1346, 2008.
11. G. Ponti and A. Tagarelli. Topic-based Hard Clustering of Documents using Generative Models. In *Proc. 18th Int. Symp. on Methodologies for Intelligent Systems (ISMIS)*, pages 231–240, 2009.
12. I. Sato and H. Nakagawa. Knowledge Discovery of Multiple-Topic Document using Parametric Mixture Model with Dirichlet Prior. In *Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 590–598, 2007.
13. M. M. Shafiei and E. E. Milios. A Statistical Model for Topic Segmentation and Clustering. In *Proc. Canadian Conf. on Artificial Intelligence*, pages 283–295, 2008.
14. M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In *Proc. KDD'00 Workshop on Text Mining*, 2000.
15. Q. Sun, R. Li, D. Luo, and X. Wu. Text Segmentation with LDA-based Fisher Kernel. In *Proc. 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT)*, pages 269–272, 2008.
16. J. Zeng, W. K. Cheung, C. Li, and J. Liu. Multirelational Topic Models. In *Proc. 9th IEEE Int. Conf. on Data Mining (ICDM)*, pages 1070–1075, 2009.
17. Y. Zhao and G. Karypis. Empirical and Theoretical Comparison of Selected Criterion Functions for Document Clustering. *Machine Learning*, 55(3):311–331, 2004.
18. S. Zhong and J. Ghosh. Generative Model-Based Document Clustering: a Comparative Study. *Knowledge and Information Systems*, 8(3):374–384, 2005.