ИНФОРМАТИКА

UDC 025.4.03:[004.4:351.852]
MSC 68T50

Modification biterm topic model input feature for detecting topic in thematic virtual museums

S. Anggai, I. S. Blekanov, S. L. Sergeev

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg,
199034, Russian Federation

For citation: Anggai S., Blekanov I. S., Sergeev S. L. Modification biterm topic model input feature for detecting topic in thematic virtual museums. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2018, vol. 14, iss. 3, pp. 243–251. https://doi.org/10.21638/11702/spbu10.2018.305

This paper describes the method for detecting topic in short text documents developed by the authors. The method called Feature BTM, based on the modification of the third step of the generative process of the well-known BTM model. The authors conducted experiments of quality evaluation that have shown the advantage of efficiency by the modified Feature BTM model before the Standard BTM model. The thematic clustering technology of documents necessary for the creation of thematic virtual museums has described. The authors performed a performance evaluation that shows a slight loss of speed (less than 30 seconds), more effective using the Feature-BTM for clustering the virtual museum collection than the Standard BTM model.

Keywords: topic model, biterm, short text, BTM, clustering, thematic virtual museums.

Introduction. Recent year topic model is becoming a popular method to identify and organise hidden topic in document collections. The topic model can discover and determine latent topic from a large number of unstructured texts in a corpus automatically using bag of words techniques. In the virtual museum, a curator or museum administrator are analysing and organising numerous online exhibitions of museum object collections to communicate their existence, contextual, value, and many reasons behind the objects. However, they are relying on label information and metadata from the structured database for providing online or thematic exhibitions, and some of the museum institutions do not have thematic exhibitions [1–4].

In development latent information and discovering a topic from a document corpus, there are several techniques have been proposed such as latent semantic indexing (LSI) [5], which offering dimensionality reduction using singular value decomposition and extended calculation from traditional vector space model (VSM). LSI has solved problems about a word or phrase that means exactly or nearly the same as another word or phrase in the same language (synonym) and the coexistence of many possible meanings for a word or phrase (polysemy). LSI also produces a representation of the underlying "latent" semantic structure of the information. Retrieving information in LSI overcomes some of the problems of keyword matching by retrieval based on the higher level semantic structure rather than just the surface level word choice [6].

In 1998, Hofmann introduced unsupervised learning technique called probabilistic latent semantic indexing (PLSI) that had a solid statistical foundation. Since it based on the likelihood principle, defines a proper generative model of the data, identifying and distinguishing between different contexts of word usage without recourse to a dictionary or thesaurus [7, 8] it assumes that in the document contain topics mixtures. In 2003, David Blei et al., proposed latent Dirichlet allocation (LDA) [9], which used the generative probabilistic model of a corpus and represented a mixture of topics in normal document texts. However, due to poor conventional topic models such as PLSI and LDA, Yan et al., proposed generative biterm topic model (BTM) [10, 11] to overcome short texts in a document, and this method outperform LDA even on normal texts.

In this research work, we conduct experiments to exploit BTM feature parameter by modifying input feature of third step BTM generative process in order to improve topic quality and discover themes from virtual museum document collections automatically.

Biterm Topic Model. The BTM basic concept was generating biterm or word-pair from the whole corpus, where the word-pair co-occurrence pattern was an unordering from the fixed sliding window. The generated co-occurrence word from document sliding window and built a set of word-pair of the whole corpus made BTM enhanced topic learning and solved the problem of the sparse word at the document level. The data generation process under BTM had the result the corpus consists of a mixture of topics, and each biterm drew from a specific topic [10, 11]. The graphical BTM plate representation as shown in Fig. 1 [10].

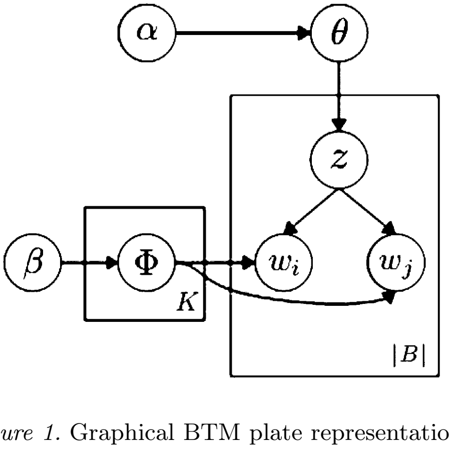

Figure 1. Graphical BTM plate representation

In generating biterm from a document corpus, BTM directly removes stop word and then generate biterm based on the initialised fixed-size sliding window. This probability method drew of couple words or biterm to a specific topic. The steps of BTM generative process introduced in [10, 11] can be written as the following:

1) for each topic z:
   a) draw a topic-specific word distribution $\Phi_z \sim \text{Dir}\ (\beta)$;
2) draw a topic proportion vector $\theta \sim \text{Dir}\ (\alpha)$ for the whole collection;
3) for each biterm b in set B:
   a) draw a topic assignment $z \sim \text{Multi}\ (\theta)$,
   b) draw two words: $w_i, w_j \sim \text{Multi}\ (\Phi_z)$.

In BTM the initial number topic z contained the sum of topic-specific word distribution from the whole collection. It is indicating that for each biterm in set B is assign random topic as the initial state. The detail extraction word-pairs from corpus as the following equation [12]:

$$\text{GenBiterm(words)} = \sum_{i=1}^{N-1} \sum_{j=1}^{N} \text{biterm}(w_i, w_j). \tag{1}$$

In the process of extraction as in equation (1) is necessary to determine the size of sliding window, and the word-pairs is given a unique identifier to prevent duplicate with assumption generated word-pairs biterm(w_i, w_j) is equal to biterm(w_j, w_i). The output from this process is set biterm B, which directly model the word co-occurrences in the whole corpus to make full use of the global information [13].

In the development of BTM, Yen et al., were using collapsed Gibbs Sampling [14] to conjugate out priors, where have contained three latent variables z, Φ, and θ. The latent variables can be integrated out using α and β. They were calculating $P\left(z \mid z_{\neg b}, B, \alpha, \beta\right)$ for each $z_{\neg b}$, where $z_{\neg b}$ denotes the topic assignments for all biterms except b, B is the global biterm set. The joint probability of all the data was using conditional probability as the such equation [10, 15]:

$$P\left(z \mid z_{\neg b}, B, \alpha, \beta\right) \propto \left(n_z + \alpha\right) \frac{\left(n_{w_i|z} + \beta\right)\left(n_{w_j|z} + \beta\right)}{\left(\sum_w n_{w_i|z} + M\beta\right)^2}, \tag{2}$$

where n_z is the number of times of the biterm b assigned to the topic z, and $n_{w|z}$ is the number of times of the word w assigned to the topic z. In [10, 11] have determined that when the biterm has assigned to a topic, both of w_i and w_j actually assigned to the same topic.

For each biterm in set B iteration always calculate and update the biterm information by assigning to a specific topic using equation (2). After period times of iteration has been performed, it will be easily estimated topic-word distribution Φ and global-topic distribution θ using the equations [10]

$$\Phi_{w|z} = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta}, \tag{3}$$

$$\theta_z = \frac{n_z + \alpha}{\mid B \mid + K\alpha}. \tag{4}$$

The output of topic-word distribution in equation (3) and global-topic distribution in equation (4) can be stored in a file or database in the table form, where the rows are all unique words in the entire documents collection or single row for global-topic; the columns are all topics of the collection.

BTM can infer a topic in a document by using the equation

$$P\left(z \mid d\right) = \sum_{b} P\left(z \mid d\right) P\left(b \mid d\right), \tag{5}$$

here the $P\left(z \mid b\right)$ can be calculated by using Bayes formula as follows:

$$P\left(z \mid b\right) = \sum P\left(z \mid d\right) P\left(b \mid d\right) \frac{P\left(z\right) P\left(w_i \mid z\right) P\left(w_j \mid z\right)}{\sum_z P\left(z\right) P\left(w_i \mid z\right) P\left(w_j \mid z\right)}. \tag{6}$$

In formula (6) $P\left(z\right)$ is global topic proportion θ_z; $P\left(w \mid z\right)$ is topic-word distribution $\Phi_{i|z}$. To obtain probability $P\left(b \mid d\right)$ as in equation (5), the distribution of biterm in the document can be estimated by the following equation:

$$P\left(b \mid d\right) = \frac{n_d\left(b\right)}{\sum_b n_d\left(b\right)}, \tag{7}$$

where n_d is the frequency of generated biterm b in document d.

In order to evaluate topic quality, Mimno et al. [16], proposed topic coherence measure that corresponds well with human coherence judgments and makes it possible to identify specific semantic problems in topic models without human evaluations or external reference corpora as follows:

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=2}^{m-1} \log \frac{D\left(v_m^{(t)} v_l^{(t)}\right) + 1}{D\left(v_l^{(z)}\right)}. \tag{8}$$

In formula (8) $D\left(v\right)$ is word document frequency type v, $D\left(v, v'\right)$ is co-word document frequency type v and v'. $V^{(t)} = \left(v_1^{(t)} \ldots v_M^{(t)}\right)$ is the list top-M most probable word in each topic t. Smoothing count value is one, in order to avoid zero number in logarithm calculation. This coherence measure is sometimes called UMass metric which is more intrinsic in nature, it attempts to confirm that the models learned data known to be in the corpus [17].

Proposed BTM input feature. The fundamental idea of BTM is that if two words co-occur more frequently, they are more likely to belong to the same topic [11] with the assumption that generated word-pair of documents will be drawn independent from the topic. Starting from that assumption, we incorporate TVM indices function to calculate TF-IDF weighting score to adjust a feature for each biterm in set B.

As our concern on providing a list of document collections which thematically similar with given the word or document query, we pay attention on the third step of BTM generative process. This issue also has been revised in d-BTM [18] proposed by Xia et al., where they have focused on biterm discrimination $w_i - w_j$. However, in our method, we prepare a biterm set B, then assign biterm feature based on word-pairs w_i and w_j. The weighting pair of w_i and w_j are using numerical statistic TF-IDF method [19, 20] for measuring how important the word in the document, where term frequency is logarithm (L), document frequency is inverse document frequency of term (T) and normalisation is pivot unique (U). The Logarithm-Term-Pivoted Unique (LTU) combination of TF-IDF as follows:

$$\mathrm{TF} = 1 + \log(tf_{t,d}), \tag{9}$$

$$\mathrm{IDF} = \log(1 + N/df_b), \tag{10}$$

$$\mathrm{Norm} = 1/N_t, \tag{11}$$

$$\text{TF-IDF=TF . IDF . Norm}, \tag{12}$$

here $tf_{t,d}$ is a number of times a term appeared in global generated document sliding window, N is the total number biterm in set B, df_b is a number of times biterm co-occurrence in global generated document sliding window, N_t is the number of unique terms in set B. By taking advantage feature of the term and biterm co-occurrence from document sliding window, we adopt TF-IDF model calculation to formulate weighting score that will be used as a BTM input feature.

Experiments and results. The experiment carried out to evaluate input parameter feature in BTM that have been proposed using TF-IDF variance as in equations (9)–(12) and comparing the output with Standard BTM feature; we also perform topic labelling document cluster using our proposed feature parameter of BTM. In this experiment, we have used Intel Xeon Processor E5-2620 v4 (Broadwell) 2.1 GHz, memory DDR4 32 Gb, and hard disk Skyhawk Surveillance 2 Terabyte. The documents have used in this experiment based on thematic virtual museums (TVM) corpus that contained 29.362 collections [21], that reduced to 23.485 in minimum two terms contained in a document.

In order to compare Standard BTM with our proposed input feature, we were performing difference K-topics number, inferring all words which were containing in K-topics, and applying standard intrinsic UMass method for measuring topic coherence. We count average coherence score of Top-N words for each K-topics [10], the higher score is indicating better performance. In all cases, we defined $a = 50/\mathrm{K}$, $b = 0.01$, and Top-$N = 10$. The calculation result based on UMass coherence measure as shown in Table.

Table. **Calculation result based on UMass coherence measure**

Iteration	Method	K=10	K=20	K=50	K=70	K=100
100	Feature BTM	−33.77	−38.74	−53.42	−52.96	−53.57
	Standard BTM	−61.37	−58.46	−59.85	−59.26	−58.08
200	Feature BTM	−35.56	−39.06	−52.18	−52.62	−53.84
	Standard BTM	−61.93	−59.56	−52.18	−58.56	−58.53
300	Feature BTM	−36.64	−39.72	−52.10	−52.40	−53.92
	Standard BTM	−60.66	−57.70	−59.87	−58.34	−59.35
400	Feature BTM	−35.82	−39.90	−51.87	−52.34	−54.04
	Standard BTM	−60.95	−56.67	−59.72	−58.37	−58.94
500	Feature BTM	−35.97	−40.09	−51.71	−52.06	−54.17
	Standard BTM	−60.02	−57.07	−60.25	−52.06	−58.83
1000	Feature BTM	−35.68	−40.15	−51.95	−52.06	−54.74
	Standard BTM	−60.97	−58.25	−60.22	−60.22	−60.22

In Table above shows the approximately average scores for each K-dimensional topic and number iterations of Standard BTM and Feature BTM, where the calculation of coherence score have performed each ten times iteration in order to get a more precise average score. One can be noticed that our proposed method gives significant improvement of topic quality with a t-test on p-value less than 0.001. The detail graphic visualisation of Table calculation result based on UMass coherence measure as shown in Fig. 2.

In Fig. 2, the average coherence score presented for each K-dimensional topics in the single graph can be more clearly investigated, where the average gap coherence score

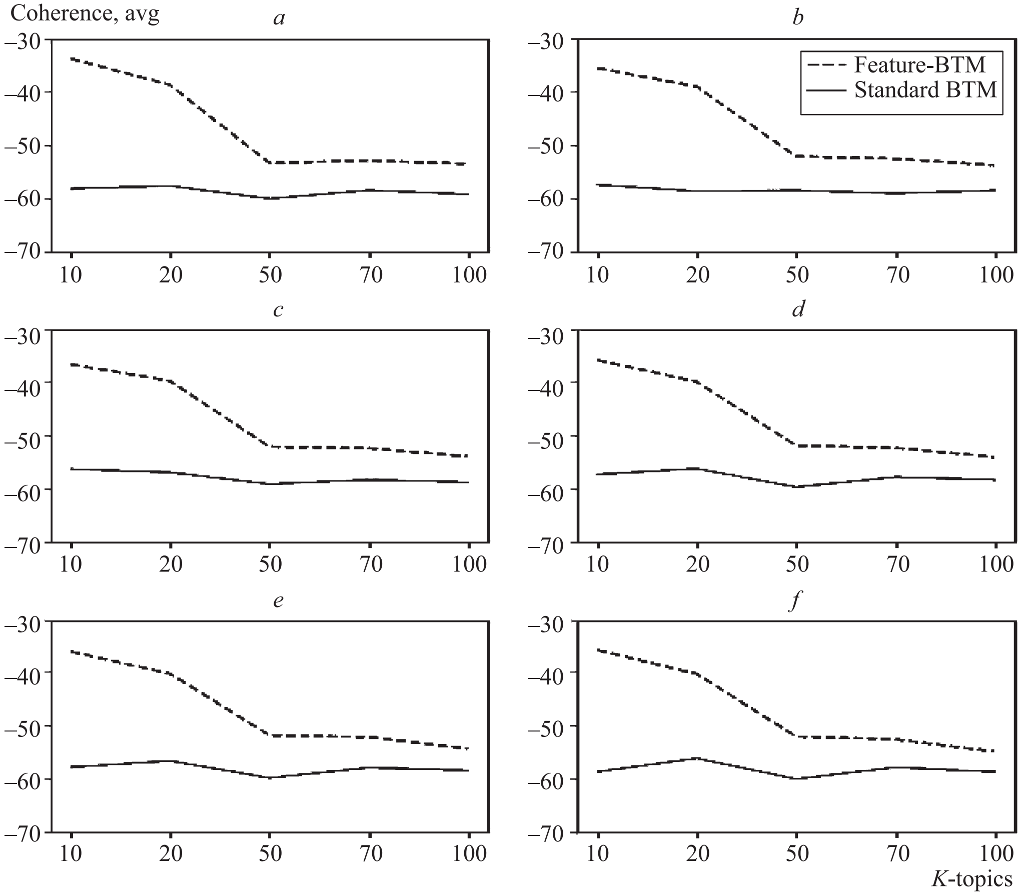

Figure 2. Calculation result based on UMass coherence measure

Iterations: a — 100, b — 200, c — 300, d — 400, e — 500, f — 1000.

between Standard BTM and our proposed method were small when the K-dimensional increased. This explained by the fact that at K number from 10 to 50, the method well identifies and underestimates the weight of biterms frequently used, which in the case of Standard BTM with such K fall into almost all topics.

In viewing topic model as a method for dimensional reduction, we performed experiments to cluster the documents based on the TVM corpus and assign to a specific cluster label. The process of assigned a topic cluster to a document as follows. The first step, for each topic $j\,(j \in K)$ infer N documents and for each document N_i the total probability of the all words occurrence of this document in each topic (or the overall relevance of the document to the topic) was calculated — P_{ij}. Second, we choose the highest probability score for each document $\max f\,(P_{ij})$. Finally, each cluster j assigned only to those documents that have the highest relevance value.

In this experiment, we define number of topic $K = 100$ with 1000 iterations for performing dimensional reduction by clustering TVM corpus based on a modification of BTM input feature, the cluster results based on UMass coherence measure as shown in Fig. 3.

Time for calculating input parameter weighting score was exemplarily 0.002 seconds, the total time for each update iteration was exemplarily 27.26 minutes, and the average

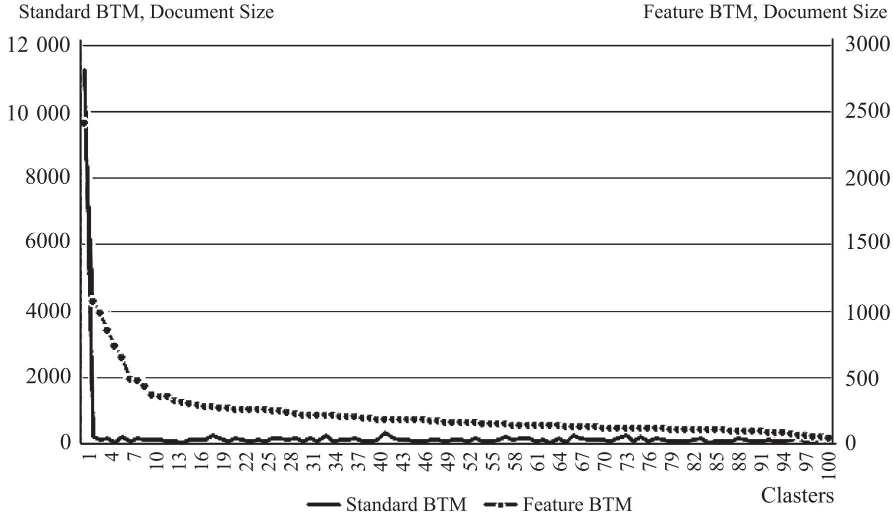

Figure 3. Cluster results based on UMass coherence measure

was exemplarily 0.027 minutes, calculating and normalising Φ and θ was exemplarily 9.25 seconds, while total time for inferring document collections and assigning topic label was exemplarily 17.29 seconds. Based on these calculation, we can estimate the total time needed for calculating proposed input feature of BTM is approximately 27.75 minutes. As shown in Fig. 3, minimum cluster size of the TVM corpus after the Standard BTM applied was 34, maximum class size was 11.275 documents, while our proposed BTM input feature, the minimum cluster size was 39 with maximal class size was 2.421 documents. We found that our proposed method gives better number document proportion of clusters than the Standard BTM. By performing topic cluster, we have reduced query time operation for retrieving relevant information in the whole documents to local documents in a cluster which related to a given query document.

Conclusion. In this paper we have proposed to exploit BTM input feature parameter based on the modification of the third step of the generative process. Experimental results shown the advantage of efficiency by the modified Feature BTM model before the Standard BTM model. The thematic clustering technology of documents necessary for creation of thematic virtual museums has described. The authors performed a performance evaluation, that shown a slight loss of speed (less than 30 seconds), more effective used the Feature BTM for clustering the virtual museum collection than the Standard BTM model.

References

1. Anggai S. The design and implementation of social networking at virtual museum of Indonesia (a case study museum of geology). Bandung, Indonesia, Bandung Institute of Technology, 2012, 60 p. (see pp. 1–60).

2. Foo S. Online virtual exhibitions: concepts and design considerations. *Journal of Library and Information Technology*, 2008, vol. 28, pp. 1–19.

3. Champion E. Entertaining the similarities and distinctions between serious games and virtual heritage projects. *Entertainment Computing*, 2016, vol. 14, pp. 67–74.

4. Palombini A. Storytelling and telling history. Towards a grammar of narratives for cultural heritage dissemination in the digital era. *Journal of Cultural Heritage*, 2017, vol. 24, pp. 134–139.

5. Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, vol. 41, pp. 391–407.

6. Foltz P. W. Using latent semantic indexing for information filtering. *SIGOIS Bull.*, 1990, vol. 11, pp. 40–47.

7. Hofmann T. Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR conference on Research and Development in Information Retrieval.* Berkeley, California, USA, 1999, pp. 50–57.

8. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001, vol. 42, pp. 177–196.

9. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 2003, vol. 3, no. 2, pp. 993–1022.

10. Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts. *Proceedings of the 22nd International conference on World Wide Web (WWW'13).* Rio de Janeiro, Brazil, 2013, pp. 1445–1456.

11. Cheng X., Yan X., Lan Y., Guo J. Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 2014, pp. 2928–2941.

12. Xu J., Liu P., Wu G., Sun Z., Xu B., Hao H. A fast matching method based on semantic similarity for short texts. *Natural Language Processing and Chinese Computing*, 2013, vol. 400, pp. 299–309.

13. Wang P., Zhang H., Liu B. X., Hao H. Short text feature enrichment using link analysis on topic-keyword graph. *Natural Language Processing and Chinese Computing*, 2014, vol. 496, pp. 79–90.

14. Griffiths T. *Gibbs sampling in the generative model of latent dirichlet allocation.* Stanford technical report. Stanford, California, USA, 2002, pp. 1–3.

15. He X., Xu H., Li J., He L., Yu L. FastBTM: reducing the sampling time for biterm topic model. *Knowledge-Based Systems*, 2017, vol. 132, pp. 11–20.

16. Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Edinburgh, United Kingdom, 2011, pp. 262–272.

17. Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D. Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Jeju Island, Korea, 2012, pp. 952–961.

18. Xia Y., Tang N., Hussain A., Cambria E. Discriminative Bi-Term topic model for headline-based social news clustering. *The Twenty-Eighth International Florida Artificial Intelligence Research Society Conference.* Florida, North America, 2015, pp. 311–316.

19. Manning C. D., Raghavan P., Schutze H. Introduction to Information Retrieval. New York, USA, Cambridge University Press, 2008, 544 p. (see pp. 61–123).

20. Zhang W., Yoshida T., Tang X. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 2011, vol. 38, no. 2, pp. 2758–2765.

21. Anggai S., Blekanov I. S., Sergeev S. L. Index data structure, functionality and microservices in thematic virtual museums. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2018, vol. 14, iss. 1, pp. 31–39. DOI: 10.21638/11701/spbu10.2018.104

Received: 10.03.2018; accepted: 14.06.2018.

Author's information:

Sajarwo Anggai — postgraduate student; sajarwo@gmail.com

Ivan S. Blekanov — PhD Sci. in Technics, Associate Professor; i.blekanov@spbu.ru

Sergei L. Sergeev — PhD Sci. in Physics and Mathematics, Associate Professor; slsergeev@yandex.ru

Модификация метода тематического моделирования BTM для обнаружения тем в тематических виртуальных музеях

С. Анггаи, И. С. Блеканов, С. Л. Сергеев

Санкт-Петербургский государственный университет, Российская Федерация,
199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Anggai S., Blekanov I. S., Sergeev S. L.* Modification biterm topic model input feature for detecting topic in thematic virtual museums // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14. Вып. 3. С. 243–251. https://doi.org/10.21638/11702/spbu10.2018.305

В статье описывается разработанный авторами метод обнаружения тем в коротких текстовых документах из виртуальной музейной коллекции. Данный метод получил название Feature BTM, поскольку основывается на модификации третьего шага генеративного процесса известной тематической модели BTM. Был поставлен эксперимент по оценке качества, который показал преимущество в эффективности детектирования тем модифицированной моделью Feature BTM перед классической моделью BTM. Была описана технология тематической кластеризации документов, необходимая для построения тематических виртуальных музеев. Проведена оценка производительности, показывающая при незначительной потери скорости (менее 30 с) большую эффективность применения Feature BTM для выполнения кластеризации виртуальной музейной коллекции, чем использования классической модели BTM. Полученный авторами метод позволяет решить проблемы зашумленности и смещения темы при их выявлении, которые имеются в модели BTM.

Ключевые слова: тематическая модель, битерм, короткие тексты, модель BTM, кластеризация, тематический виртуальный музей.

Контактная информация:

Анггаи Сажарво — аспирант; sajarwo@gmail.com

Блеканов Иван Станиславович — канд. техн. наук, доц.; i.blekanov@spbu.ru

Сергеев Сергей Львович — канд. физ.-мат. наук, доц.; slsergeev@yandex.ru