Национальный исследовательский
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ
ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ
им. В.А. ТРАПЕЗНИКОВА РАН

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2018)

МАТЕРИАЛЫ
XVII Международной конференции
имени А. Ф. Терпугова
10–15 сентября 2018 г.

ТОМСК
«Издательство НТЛ»
2018

УДК 519
ББК 22.17
И74

И74 Информационные технологии и математическое моделирование (ИТММ-2018): Материалы XVII Международной конференции имени А.Ф. Терпугова (10–15 сентября 2018 г.). – Томск: Изд-во НТЛ, 2018. – 442 с.

ISBN 978-5-89503-621-1

Сборник содержит материалы докладов, в которых отражены основные направления XVII Международной конференции имени А.Ф. Терпугова: прикладной вероятностный анализ, теория массового обслуживания и теория телетрафика, интеллектуальный анализ данных, информационные технологии и математическое моделирование.

Для специалистов в области информационных технологий и математического моделирования.

УДК 519
ББК 22.17

Конференция проводится при поддержке
Российского фонда фундаментальных исследований
(*проект № 18-07-20048-г*)

ISBN 978-5-89503-621-1

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

Framework for Arabic Text Classification Based on Topic Modeling

Ghassan Khazal, Alexander Zamyatin

Computer Science Department, Tomsk State University, Tomsk, Russia

Most Arabic text classification researchers represent the text as a Term Frequency-Inverse Document Frequency (TFIDF). The TFIDF is not able to recognize synonyms and semantic relationships between texts. Topic modeling was used to represent words in the documents as topics to process text accurately and efficiently. In this work, we apply topic modeling to cluster Arabic words into topics. These words assigned to the same topic are related semantically. We make a model for three different approaches that used in dimensionality reduction by using topics modeling and show their effectiveness in Arabic text classification by using different classifiers. First, we apply LSI, LDA, and HDP on Arabic text document then we generate three different matrices of topics for LDA and LSI and HDP. The goal of this work is to compare between the feature represented with TFIDF and the topic model using two text classifiers Support Vector Machines (SVM) and Neural Network (NN). The results showed that the topic modeling using LSA and SVM give better performance based on F1 measure equal to 90%.

1. Introduction

Text classification is the field of natural language texts to one or more predefined classes according to their content [1]. Nowadays, Arabic text documents are increasing at an explosive rate. Arabic Text classification has become a key technique for summarizing, organizing, understanding, and searching these documents and also for the information retrieval for the documents with a specific topic. Text classification usually uses supervised machine learning algorithms for learning the classification model [2, 3]. Topic modeling has wide research direction in data mining and natural language processing. Topic modeling used to identify the semantic structure of the text document by using Bayesian analysis on the collected text document.

When we try to access any big text information that we want it requires the complex structure to organize these text data for searching, summarizing or understanding these big text data. These data consist of a lot of topics or any document may be having a mixture of topics. For this topic modeling is useful for auto summarizing, searching, and understanding text data. Most of the Arabic text researchers use vector space model for text document representation. i.e. represent the text as a vector of words and this feature cannot indicate the semantics of the text, and also it will cause a huge size of feature space. Our proposed model for this work we use topic modeling for feature representation to avoid this problem. In this work we use three techniques of topic modeling; Latent Semantic Analysis LSA, Latent Dirichlet Allocation LDA, and Hierarchical Dirichlet Process HDP. The performance based on topic modeling on Arabic text classification using two classifiers SVM and NN. In the experimental analysis, we searching for an answer to the following question: Which classification algorithms and a number of latent topics yield better results and are there optimal variables used of latent topics based representation of Arabic text classification?

This paper is organized as following: Section 2 represent work related to this paper. Section 3 represents a review of topic modeling techniques. Section 4 describes topic model for Arabic text classification. Experiment and results are showed in section 5, and the conclusions in section 6.

2. Related work

There is previous work that uses topic modeling as feature representation for text classification to enhance text the performance, the undertaken tasks on text classification. In [1] investigate generative statistical topic modeling for multi-label documents that associate with individual tokens with different labels. In [2] propose a novel approach to build classification system and to classify documents using multi-latent Dirichlet allocation (LDA). And [3] use conventional representation schemes for text documents with large text as data reduction technique. The author evaluates the performance of LDA model for text classification by using five classifiers. In [4] use the topic hidden analysis of LDA for feature representation then compare it with the classical feature representation in text classification. For the experiments, he uses SVM as the classifier and different type of weighting ranking for terms. The [5] focus on studying the LDA model then compare it with bigram feature. The [6] propose a hierarchical model for classification by using topic modeling and integrating additional term features from neighboring pages.

The [7] focuses on the topic identification for the Arabic language by using topic modeling with different Arabic stemmers on LDA. The author shows LDA is an efficient for Arabic identification, especially with the appropriate hyper-parameters. In [8] apply the topic modeling on clustering the words into topics then compare between the feature technique of BOW and the topic modeling using LDA. Finally [9] focus on the filtering approach for dimensionality reduction. This filtering approach consists of two stages; feature ranking and thresholding. The experiments done by using four English datasets and two Arabic datasets and its indicate the possibility of using this approach in Arabic TC.

3. Topic modeling techniques

In text classification, topic modeling has an important effect. For the collected text documents probability of the topic modeling will compute set of polynomial distribution by using parameter estimation. This distribution called topic, by capturing related information of the words. To extract stable latent structure and finding a short relative description of the text document in the dataset. In topic modeling the basic idea is assume that each document is a mix of topics. Each topic represents the probability distribution of the words. Also, topic models represent as text generation model. I.e., when generates a new text, a topic distribution will be gained first for each word in the text, then a topic is obtained randomly by the topic distribution. In the end, a word will be got randomly by the word distribution of the topic [5].

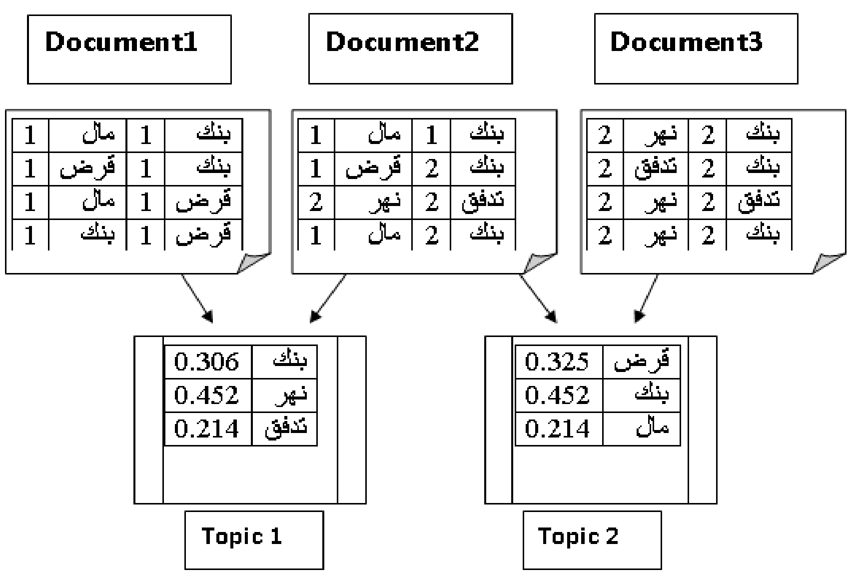

Fig. 1. Generative process of topic modeling

A. Latent Semantic Analysis (*LSA*)

This type of modeling used to represent the terms as a vector and it useful for finding the similarities among texts by computing the distance among vectors. First LSA represents the text as a matrix the row used for terms and columns for the documents.

Also, it uses TFIDF to calculate the interaction between words and the document. TFIDF represented in equation 1and 2. Where TF in each document represents term frequency and IDF represent term important in all documents.

$$\mathrm{TF}_{\mathrm{IDF}} = \mathrm{TF} \times \mathrm{IDF}, \tag{1}$$

$$\mathrm{IDF} = \log \frac{N}{N_t}. \tag{2}$$

To refine Term document matrix (TDM), LSA use technique called Singular Value Decomposition (SVD). LSA uses Singular Value Decomposition SVD techniques to enhance the Term Document Matrix TDM, SVD decomposes TDM to three matrices to confirm the robust relation among words and the documents also remove the noise. The decomposition displays all characteristics of theses matrices. Equation (3) represents SVD, and equation (4) represents decomposition of $m \times n$ matrix of the SVD [10].

$$\mathrm{TDM} = U \times S \times V^T. \tag{3}$$

Where U: m dimensional matrix, V: n dimensional matrix, and S: diagonal matrix [3].

$$A_{n\times d} = U_{n\times r}\, D_{r\times r}\, V^T_{r\times d}. \tag{4}$$

B. Latent Dirichlet Allocation (*LDA*)

LDA is a generative probabilistic model for group of documents. The main idea of this method (LDA) is to represent the documents as mixtures of topics. The probability distribution is use to represent the topic of the term and each article is distribution probability of topics [11]. LDA used for identification of topics in different fields such as classification, filtering, and information retrieval. There are three levels to represent LDA. The parameters α and β represents corpus level, these parameters are used to generating a corpus process. α is uniform Dirichlet prior distribution of per document topic distributions. β is the uniform Dirichlet prior distribution of per topic word. θ is a document level parameter used to sample one time per document. z and w are word-level parameters and sampled only one time for each word in each document. Finally, N represents the number of the terms tokens in a document and M represents the number of documents [12, 13].

The LDA introduces set of (K) latent parameter, i.e topics. Each word in the document is generated by one of the topics. The generative process for each document w can be described as follows [11]:

$1.\ Choose\ N \sim Poisson(\varepsilon)$

$2.\ Choose\ \theta \sim Dir(\alpha)$

$3.\ for\ each\ of\ the\ N\ words\ W_n :$

$a)\ choose\ a\ topic\ z_n \sim Multionmial\ (\theta)$

$b)\ choose\ a\ word\ w_n\ from\ P(w_n|z_n,\beta)\ a\ multinomial$
$probability\ condition\ on\ the\ topic\ z_n$

$$[\beta]_{k\times v}\ \ \beta_{ij} = p(w^j = 1 \mid z^j = 1) \tag{5}$$

A k-dimensional Dirichlet random variable θ can take values in the (k–1)-simplex, and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}, \tag{6}$$

$$p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\beta), \tag{7}$$

$$p(\boldsymbol{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d^k\theta, \tag{8}$$

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d^k\theta_d. \tag{9}$$

To explain how LDA generates a document. Suppose that a text corpus contains four topics, such as political, economics, techniques and movies. The document describe the production of movie process may contain both the topic of technology and movie. Technology topic may have a many words that related to technology. These words have a probability, probability of these words appearing in the document with the topic technology. Similarly, there are many words that related to movies in the movie’s topic, and also have probability. When generating a document about movie production, the topics will be chosen randomly first. The probability of choosing two topics of technology and movies will be higher. And then, a word will be selected. The probability of choosing the words that are related to the two topics will be higher. So, the selection of a word is completed. After selecting N words continually, a document is created [12, 14].

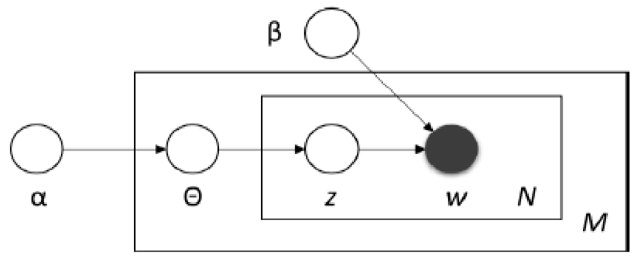

Fig. 2. LDA representation

C. Hierarchical Dirichlet Process HDP

HDP is generative model for the document in the text corpus and can be described as document d, d is a sequence of words drowns fro vocabulary V with $|V| = k$. The probability of the document d given the class i in the hierarchy given by [13]:

$$P(d|i) = P(|d|)\frac{(\sum_j d_j)!}{\prod_j d_j!}\prod_{j=1}^{k}(\theta_{ij})^{d_j}, \tag{10}$$

where d_j is the number of occurrences of the jth word of the vocabulary in document d;

$$\theta_i = (\theta_{i1},\ldots\ldots,\theta_{ik})^T. \tag{11}$$

The parameter vector describes the parameters for a multinomial distribution. We note that, contrary to most previous work where documents are classed only on the leaves of the concept hierarchy, we consider every node in the concept hierarchy as a valid class into which some of the documents belong [8, 11].

Also, the vector θi have dirichlet prior distribution:

$$\theta i \sim Dir(1,\ldots..,1)\, if\ i\ is\ the\ root\ node\ ,$$

$$\theta i\,|\,\theta pa(i) \sim Dir\,\sigma\,\theta pa(i)\, other\ wise\ ,$$

where $pa(i)$ identifies the parent of node I and σ is constant smoothing parameter chosen in advance.

4. Experiments of the proposed work

The goal of this work is to design topic model to show the effect of this feature representation on the classification accuracy of Arabic text using two classifiers. Figure 3 represents the proposed text classification models.

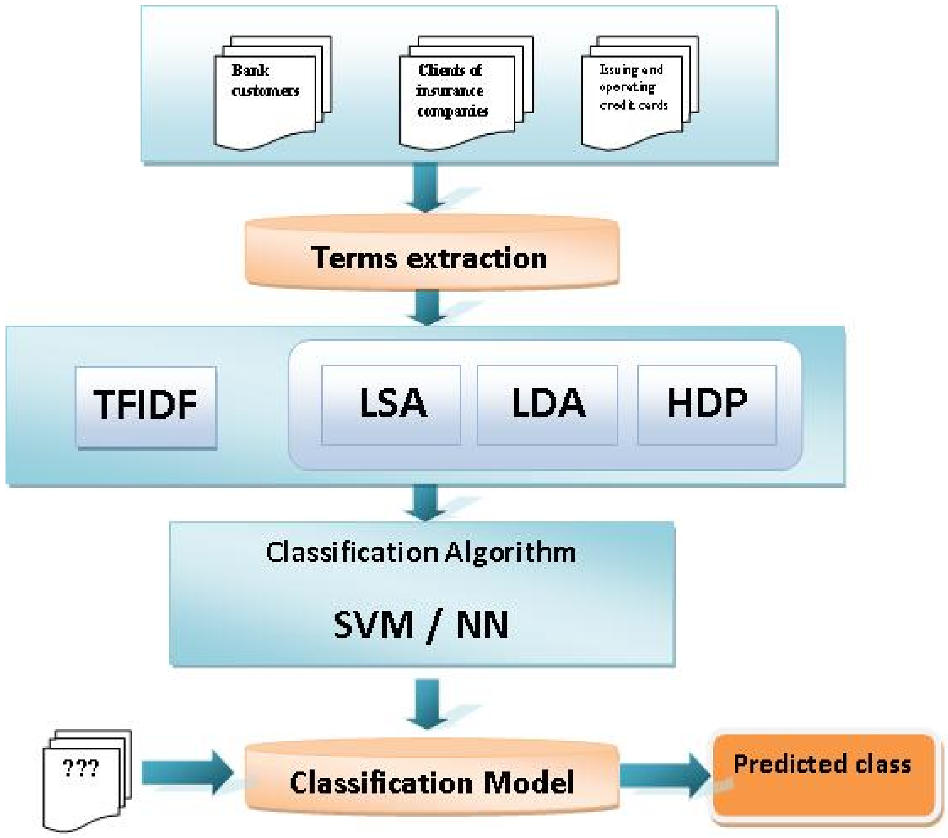

Fig. 3. represents the proposed text classification models

The keyword-based for our model for classification is constructed of five steps:

Step 1: Gathering Arabic text data to apply our model.

Step 2: Text data preprocessing to clean the text from the unwanted text.

Step 3: Apply different feature representation TFIDF, LDA, LSA, and HDP.

Step 4: Construct trained classification Model.

Step 5: Evaluate the results.

In our experiments, we use a collection of text documents for bank questionnaires. We selected 6 classes for each class 50 document. The total number of documents is 1800, as shown in Table 1.

Table 1

Dataset information

Category name in Arabic	Category name in English	Doc. No.
إصدرا وتشغيل بطاقات الإئتمان	Issuing and operating credit cards	300
المنازعات اولمخالفات التأمينية	Disputes and insurance violations	300
حماية عملاء البنوك	Protecting bank customers	300
حماية عملاء شرآات التمويل	Protecting clients of finance companies	300
عملاء البنوك	Bank customers	300
عملاء شرآات التأمين	Clients of insurance companies	300

For preprocessing splitting written string into tokens called tokenization, which it represent the first step in many natural language processing tasks. For most languages, tokenization involves splitting punctuation, and some affixes off of the words [6]. Also to minimize the dimensionality of term space removing non-Arabic characters, removing stop words (articles, prepositions, and pronouns), removing numbers, removing Diacritics, and word length less than three characters.

We apply 8 experiments to show the effect of topic modeling and compare with TFIDF model. The experiment parameters as described in Table 2.

Table 2

Experiment Parameters

Parameters	Description
Dataset	In-house collected Dataset “banking client’s questions and queries” [7, 8]
Training size	70%
Testing Size	30%
Pre-processing	Tokenizing, filtering (stop word removal, word length <3 characters, non-Arabic word removal, numbers removal, districts removal…)
Feature representation	TFIDF, LSA, LDA, or HDP
Classifiers	SVM, NN

To report the evaluation results, we use three metrics: **R**ecall (equation 12), **P**recision (equation 13) and **F**-measure (equation 14).

$$Rcall = \frac{No.of\ Doc.correctly\, labeled}{No.of\ topics\, Doc.}, \tag{12}$$

$$Precision = \frac{No.of\ Doc.correctly\, labeled}{No.of\ labelled\ Doc.}, \tag{13}$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}. \quad (14)$$

Table 3 shows the results in of precision, recall, and F1 for our model. The average values computed using 10-fold cross-validation. The results showed that the SVM accuracy in average better than the NN classifiers for all experiments. The results of all applied experiments on the two classification algorithms are illustrated in Table 3 and Fig. 4.

Table 3

Precision, recall, and F1 results of the proposed model

Method	F1	P	R
LSA – SVM	0.90	0.89	0.91
LDA – SVM	0.80	0.81	0.79
HDP – SVM	0.81	0.82	0.81
TFIDF – SVM	0.85	0.85	0.86
LSA – NN	0.82	0.82	0.83
LDA – NN	0.74	0.71	0.78
HDP – NN	0.75	0.71	0.79
TFIDF – NN	0.79	0.78	0.81

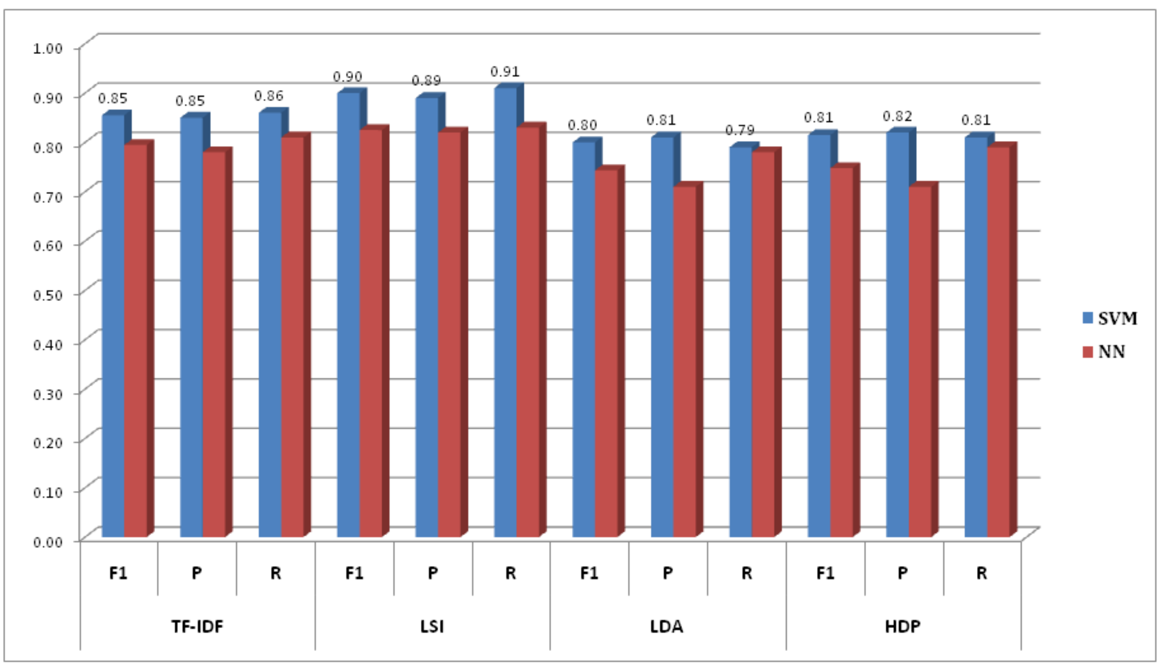

Fig. 4. Shows the distribution of average accuracy achieved with different classifiers

Also, the approach of LSA gives higher performance compared to applying the TFIDF with both classification methods. And the other method gives a good performance that reaches the performance of TFIDF model.

For LDA approach gives better performance with SVM, we use α to 0.1 and β to 0.01 according to the work of [9, 16].

5. Conclusions

In this research, three types of topic model are used for clustering term features into a group of topics. First, we find the same topic for terms that are semantically related. Then we compared classification performance among four feature representation techniques, TFIDF model and three techniques of topic model. We use two text categorization algorithms: SVM and NN for evaluation. Our results reported that the SVM accuracy in average is better than NN classifiers for all experiments. Also, the approach of feature representation with the LSA model achieves the better performance with F-measure (90 %) with SVM and (82 %) with NN; an improvement of about 5 % and 3% over the TFIDF model respectively. LDA and HPD not give any enhancement in performance, we think that was because of the relation among words are not enough and also because of noisy data.

REFERENCES

1. *Dumais S.*, *Platt J.*, *Heckerman D.*, *and Sahami M.* Inductive learning algorithms and representations for text categorization // Proc. ACMCIKM' 98. 1998.
2. *Sriurai W.* Improving Text Categorization by using a Topic Model // Adv. Comput. 2011. V. 2. No. 6. P. 21–27.
3. *Ramya Raj K.P.* Malayalam News Classification Based On Topic Modeling // Int. J. Computer & Mathematical Sciences 2017. V. 6. No. 10. P. 18–24.
4. *Rubin T.N.*, *Chambers A.*, *Smyth P.*, *and Steyvers M.* Statistical topic models for multi-label document classification // Mach. Learn. 2012. V. 88. No. 1–2. P. 157–208.
5. *Fu R.*, *Qin B.*, *and Liu T.* Open-categorical text classification based on multi-LDA models // Soft Comput. 2014. V. 19. No. 1. P. 29–38.
6. *Onan A.*, *Korukoglu S.*, *and Bulut H.* LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis // Int. J. Comput. Linguist. Appl. 2016. V. 7. No. 1. P. 101–119.
7. *Taşci Ş. and Güngör T.* LDA-based keyword selection in text categorization // 2009 24th Int. Symp. Comput. Inf. Sci. Isc. 2009. P. 230–235.
8. *Li L. and Zhang Y.* An empirical study of text classification using Latent Dirichlet Allocation // Cs.Cmu. Edu. No. 1.
9. *Sriurai W.*, *Meesad P.*, *and Haruechaiyasak C.* Hierarchical Web Page Classification Based on a Topic Model and Neighboring Pages Integration // International Journal of Computer Science and Information Security. 2010. V. 7. No. 2. P. 166–173.

10. *Said D.A.* Dimensionality Reduction Techniques for Enhancing Automatic Text Categorization, 2007.
11. *Agrawal A.*, *Fu W.*, *and Menzies T.* What is wrong with topic modeling? And how to fix it using search-based software engineering // Inf. Softw. Technol. 2018. V. 98. P. 74–88.
12. *Hadeel Alazzam*, *Abdulsalam Alsmady*. A distributed Arabic text classification approach using latent semantic analysis for big data // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 2017.
13. *Pavlinek M. and Podgorelec V.* Text classification method based on self-training and LDA topic models // Expert Syst. Appl. 2017. V. 80. P. 83–93.
14. *Zhou S.*, *Li K.*, *and Liu Y.* Text categorization based on topic model // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2008. V. 5009 LNAI. No. 4. P. 572–579.
15. *Sriurai W.* Improving Text Categorization by using a Topic Model // Adv. Comput. 2011. V. 2. No. 6. P. 21–27.
16. *Naili M.*, *Chaibi A.*, *and Ghézala H.* Arabic topic identification based on empirical studies of topic models // Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, INRIA, 2017. V. 27. Special issue CARI 2016. <hal-01444574v2>.
17. *Blei D.M.*, *Ng A.Y.*, *Jordan M.I.* Latent dirichlet allocation // Journal of Machine Learning Research. 2003. No. 3. P. 993–1022.
18. *Lee S. and Kim J.* An Extension of Topic Models for Semi-supervised Text Classification: a Term Weighting Approach // BigComp. 2015. V. 8. P. 8.
19. *Veeramachaneni S.*, *Sona D.*, *and Avesani P.* Hierarchical Dirichlet model for document classification // Proc. 22nd Int. Conf. Mach. Learn. ICML ’05. 2005. P. 928–935.
20. *Green S.*, *Denero J.* A class-based agreement model for generating accurately inflected translations // Proc. 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. 2012. V. 1. P. 146–155.
21. *SAMA*. faqs Banks consumers, 2017. Disponivel em: <http://www.sama.gov.sa/ar-sa/ConsumerProtection/Pages/faqsBanksconsumers.aspx>.
22. *DORAR*, 2017. Disponivel em: <https://www.dorar-aliraq.net/threads/445044>.
23. *Zhao W.*, *Chen J. and Zen W.* Best practices in building topic models with LDA for mining regulatory textual documents // Cder 9th November, 2015.

Khazal Ghassan, аспирант; ghassan_khazal@yahoo.com;
Zamyatin Alexander, д.т.н., профессор; avzamyatin@inbox.ru

СОДЕРЖАНИЕ

МОДЕЛИ И МЕТОДЫ МАССОВОГО ОБСЛУЖИВАНИЯ

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ИХ ПРИМЕНЕНИЕ

ПРИКЛАДНОЙ ВЕРОЯТНОСТНЫЙ АНАЛИЗ И ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Научное издание

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2018)

Материалы XVII Международной конференции имени А. Ф. Терпугова 10–15 сентября 2018 г.

Редактор *Т.С. Портнова*

Дизайн, верстка *Д.В. Фортеса*

ООО «Издательство научно-технической литературы»
634050, Томск, пл. Новособорная, 1, тел. (3822) 533-335

Изд. лиц. ИД № 04000 от 12.02.2001. Подписано к печати 22.08.2018.
Формат 60 × 84 $^{1}/_{16}$. Бумага офсетная. Печать офсетная. Гарнитура «Таймс».
Усл. п. л. 25,69. Уч.-изд. л. 28,77. Тираж 100 экз. Заказ № 21.