Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 369–376

Procedia Computer Science

www.elsevier.com/locate/procedia

没有什么价值的论文

3rd International Conference on Computer Science and Computational Intelligence 2018

Knowledge Base Ontology Building For Fraud Detection Using Topic Modeling

Girija Attigeri, Manohara Pai M M*, Radhika M Pai, Rahul Kulkarni

Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

Abstract

Moving towards the digitization and cashless economy tests the existing IT infrastructure for security and fraud controls substantially. Transition from traditional to cashless economy requires to banks to have more secure system to fight fraud. To understand and transform the needs for more secure banking system it is necessary to understand the domain of fraud and create knowledge base for fraud. It helps bridge the gap between business level and IT levels of banking. So that anti-fraud regulations could be automatically imbibed in the system. Hence the paper focuses on analyzing existing fraud case documentations and understand the significant terms involved in the fraud. For this TF-IDF weighting, topic modeling with LDA is used for identifying the group of words (topic) representing particular type of fraud. Using these knowledge base ontology is extracted which can be used for building fraud detection system. Experiment is performed on extracted fraud documents and ontology is built using the latent topics identified.

Keywords: TF-IDF; Topic modeling; Fraud detection; Ontology;

1. Introduction

With the advent of a digital economy and an increasingly growing emphasis on online transactions, the need for sophisticated techniques to identify discrepancies in these transactions has never been more urgent. Even though online transactions add transparency and efficiency to economies, they can harbor major fraud and lead to devastating effects for both institutions and individuals. This project, therefore, aims at formulating data mining techniques for the successful detection of digital fraudulent activities. Fraud is defined as wrongful or criminal deceit intended to have financial or personal gain. Online transactions are a thing of the present and future. Most of the advanced nations on our planet have put a heavy emphasis on a 'plastic economy', where the cash flowing in the economy is purposely reduced and more and more transactions are done transparently. This increases effectiveness and makes transactions

* Corresponding author. Tel.: +91-866-073-6005
E-mail address: mmm.pai@manipal.edu

10.1016/j.procs.2018.08.186

transparent. Even in India, with the latest 'Digital India' push from the government, more and more transactions are being performed online. To put this into perspective, PayTm, the e-wallet application, has 122 million active users, which indicates a huge shift towards electronic transactions. Though this is the future, there are safeguards that need to be put into place to ensure that the transactions are safe in addition to being transparent. The paper aims at aiding institutions in both the public and private sectors. According to a report by ASSOCHAM, the extensive use of internet has evolved the cyberspace and around 45% of the transaction are done using mobile and other digital devices. It also predicts growth of 65% in the number of mobile frauds by 2017 [1]. Debit and credit card fraud cases are on the top in the charts of cyber-crime and have increased six times during the last three years. Especially after the digital India push by the current government, preparation for digital fraud must me of utmost importance. Hence, these techniques aim to help both the private corporations like preventing a repeat of the Satyam fraud case as well as government organizations.

The word, 'Fraud' refers to the abuse of an organization's profit. With the advent of a digital economy and an increasingly growing emphasis on online transactions, the need for sophisticated techniques to identify discrepancies in these transactions has never been more urgent. Even though online transactions add transparency and efficiency to economies, they can harbor major fraud and lead to devastating effects for both institutions and individuals. Detection of frauds need analysis of structured and behavioral analysis of data. It requires analysis of previous fraud cases and learning from them. This learning can be represented as knowledge base in the form of ontology. The ontology helps in integrating the data from multiple sources and enables semantic queries, aiding fraud analysis.

The paper aims at formulating knowledge base ontology which can be used for the successful detection of digital fraudulent activities. Methods used to achieve this included TF-IDF weighting, Topic Modeling and ontology building. The data that was used to formulate this result was obtained from various online news portals, which include, and are not limited to, NDTV, The Times of India, and the Hindustan Times. Hundred articles were used for the purpose of this paper.

2. Background

In the paper [2] the authors explore existing fraud detection methods and categorize these into two parts as supervised and unsupervised. In prior, models are evaluated based on the samples which are labeled as fraudulent and legitimate transactions. In later outliers or unusual transactions are recognized as potential fraud cases. Both these methods give the probability of transaction being fraud in any given dataset. They use two data mining approaches, random forest and support vector machines, random forests and logistic regression, to better detect credit card fraud. The study is done on real-life international credit card transactions. The survey in [3] summarizes, compares and categorizes technical and review articles published within last 10 years in automated fraud detection. It explains the professional fraudsters, categorizes the types and subtypes of fraud. Also presents the nature of fraud evidence collected within affected industries. This survey covers many technical articles and proposes alternative data and solutions from related domains. Authors in [4] explore the significance of data mining techniques in detecting firms that issue fraudulent financial statements. It provides details on identification of factors associated to financial fraud statements. They emphasize that task of management in fraud detection, auditors should be facilitated by Data Mining techniques. The study evaluates use of Decision Trees, Neural Networks and Bayesian Belief Networks in detecting fraudulent financial statements. The input is composed of financial ratios derived from statements. These papers emphasize the significance of data mining and machine learning algorithms for fraud detection which requires data to be readily available to process. However authors in [5] mentions the difficulty of moving from business level to IT level and the need for automatic transition from business level to IT level. Hence business process modeling with the use of semantic analysis is important. Theoretic perspective and use of TF-IDF for identifying word relevance in the document is discussed in [6]. The domain specific ontology creation by analyzing the unstructured or semistructured documents is discussed in [7],[8]. Rani et. al. in [8] used topic model for constructing ontology graph for news group dataset. [9]. It has been used in various applications such as key phrase extraction in news articles to make decision to continue reading or not. In this paper TF-IDF is used to find the important phrases indicating the occurrence of fraud from fraud case documentations. LDA is used widely in various domains for natural language processing and information retrieval such as text classification [10], feature identification [11], text segmentation [12] and LDA model is used in various research applications such as defect reporting in software and software evolution [13 14 15], fraud detection in

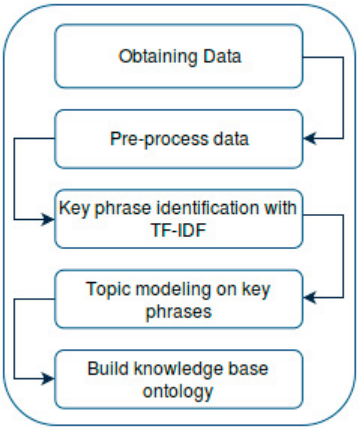

Fig. 1. Methodology for fraud knowledge base ontology creation

telecommunication[16 17], micro-array expression analysis in bioinformatics[18 19] , analyzing genomic data[20], a model which evolves with time as a dynamic model was discussed by Blei et. al. In the current work the topic modeling is used for identification of important group (topic) words indicating fraud. Use of ontology for business process mining and modeling is emphasised in[21]. Authors in[22] propose ontology based anti-money laundering expert system for identifying suspicious transactions. They filter transactions which appear to be normal based on certain rules and process rest of the transactions. They have used OWL for constructing knowledge base and built rules on top of it. Using these and SWRL rules transactions are tested. However the system could not be validated automatically as the data used was not labeled.

3. Methodology

The data collected involves news articles. The aim was to collect as many as documents about fraudulent cases. Articles were collected from news websites. Each article was stored as a document. The articles had many words which are not significant for analysis. Articles are preprocessed by applying removal of stop words and punctuation marks. This collection of preprocessed articles is use for further processing. In the next step important key phrases in the fraud documentation are analyzed using TF-IDF scores. The words with higher scores are the words providing relevant information about the fraud. However many terms may be discussing about a particular concept in fraud domain. In order to identify that topic modeling using LDA is applied and various topics are identified. This gives the top words in a particular topic along with the probability with which they belong to that topic. Finally, the output will be used to highlight suspicious activity or transactions that potentially be of fraudulent nature, through the use of an ontology.

Term frequency-inverse document frequency (TF-IDF) is used in text analysis and information retrieval. It is a statistical measure to evaluate how significant a word is to a document in a collection. As the number of times a word appears in the document increases, importance of the word also increases proportionally and is balanced by the frequency of that word in the collection. This is done by computation of TF and IDF. TF: It measures how frequently a term occurs in a document. As every document is different in length, term might appear much more times in long documents than shorter ones thus, the term frequency is normalized by dividing it by the document length as:

TF^i_w = Number of times word w appears in a document i / Total number of words in the document i

IDF: Measures importance of a word in a document. With TF all words are considered equally important. However certain words such as ”is”, ”the”, ”of”, ”that” etc. appear lot of times but are not important.

$IDF_w = log_e$ (Total number of documents / Number of documents with word w in it)

TF-IDF = TF * IDF

Topic modeling is a method used to analyze large volumes of text. A topic consists of a cluster of words with probability of occurrence in the topic. These topics can be connected using contextual clues, which connect words

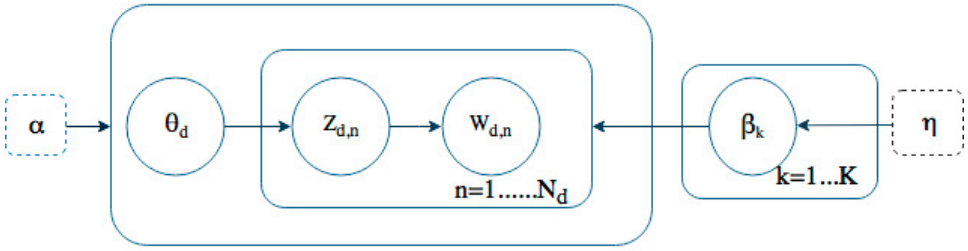

Fig. 2. LDA Model

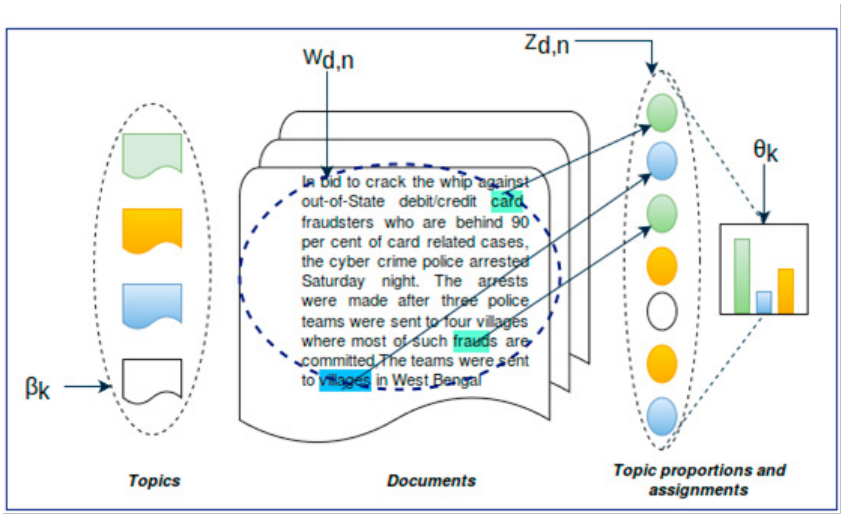

Fig. 3. Parameters of topic modeling in LDA

with similar meanings and uses of words with multiple meanings[11]. One way of performing topic modeling is Latent Dirichlet Allocation (LDA) .It is a graph-based algorithm to extract relevant key phrases. In the LDA model, each document is viewed as a mixture of topics that are present in the collection. LDA model is pictorially represented in Fig. 2. It has the following parameters to be defined for topic identification. K = number of topics, Nd= number of words in document d, D=number of documents, Z{d,n}= cluster chosen for the word n from document d, θ=document composition, β= key words for topic, a and h are the constants chosen for controlling the sparsity for the multinomial distribution.

After obtaining latent topics from LDA, these are used for ontology creation. There are many semantic relations that can be presented between topics of ontology graph such as subclass_of between topics of different levels , equivalent_to if topics are synonyms, disjoint_of between topics which represent antonyms, part_of between topics that complete others can be defined. To define these relations, the latent features are analyzed using Wordnet, verifying possible associations them. In order to simplify the process pairwise relation analysis is done and ontology graph is created.

After construction the knowledge base ontology for fraudulent transactions can be used as a reference for detecting new fraudulent transactions. For this whenever a new transaction takes place it has to be stored in an ontology graph. All the relevant data about the transaction also needs to be considered for this purpose. Then the transaction ontology has to be compared with the knowledge base ontology to identify it as legal or fraudulent transaction.

Top words indocument 1	*Top words indocument 2*	*Top words indocument 3*
Word: gupta,TF-IDF: 0.0888	Word: gurgaon,TF-IDF: 0.07428	Word: women,TF-IDF: 0.06009
Word: indico,TF-IDF: 0.05396	Word: cyber,TF-IDF: 0.06806	Word:matrimonial, TF-IDF: 0.04484
Word: medicines,TF-IDF: 0.05396	Word: crime,TF-IDF: 0.04471	Word: pretexts,TF-IDF: 0.04019
Word: shetty,TF-IDF: 0.05396	Word: police,TF-IDF: 0.04094	Word: pune,TF-IDF: 0.03689
Word: logistics,TF-IDF: 0.05396	Word: cases,TF-IDF: 0.03944	Word: networking,TF-IDF: 0.03363
Word: company,TF-IDF: 0.0439	Word: phase,TF-IDF: 0.03714	Word: tricksters,TF-IDF: 0.03363
Word: owner,TF-IDF: 0.03668	Word: digital,TF-IDF: 0.02914	Word: profiles,TF-IDF: 0.03363
Word: charges,TF-IDF: 0.03484	Word: duped,TF-IDF: 0.02819	Word: pawar,TF-IDF: 0.03363
Word: ceos,TF-IDF: 0.02698	Word: online,TF-IDF: 0.02819	Word: background,TF-IDF: 0.03363
Word:enterprises, TF-IDF: 0.02698	Word: paytm,TF-IDF: 0.02476	Word: cases,TF-IDF: 0.03265
...		
Top words in document 90	*Top words in document 92*	*Top words indocument 100*
Word: understand,TF-IDF: 0.11394	Word: identities,TF-IDF: 0.0706	Word: identities,TF-IDF: 0.06658
Word: regulation,TF-IDF: 0.10604	Word: false,TF-IDF: 0.06366	Word: dollars,TF-IDF: 0.04572
Word: inflation,TF-IDF: 0.10386	Word: credit,TF-IDF: 0.05502	Word: verma,TF-IDF: 0.04439
Word: subject,TF-IDF: 0.06206	Word: identity,TF-IDF: 0.05354	Word:institutions, TF-IDF: 0.04372
Word: price,TF-IDF: 0.05697	Word: bureaus,TF-IDF: 0.05251	Word: charged,TF-IDF: 0.04191
Word: replying,TF-IDF: 0.05697	Word: enterprise,TF-IDF: 0.04707	Word: thousands,TF-IDF: 0.04075
Word: comes,TF-IDF: 0.04262	Word:conspirators, TF-IDF: 0.04707	Word: businesses,TF-IDF: 0.03561
Word: people,TF-IDF: 0.04033	Word: addresses,TF-IDF: 0.04707	Word: millions,TF-IDF: 0.03366
Word: would,TF-IDF: 0.03576	Word: jersey,TF-IDF: 0.04707	Word: credit,TF-IDF: 0.02965
Word: specific,TF-IDF: 0.03462	Word: defendants,TF-IDF: 0.04321	Word: financial,TF-IDF: 0.02807

Fig. 4. Top TF-IDF words in each document

3.1. Experimentation

Fraud cases are manually collected from Times of India, Yahoo, Hindustan times and RBI websites. These documents are preprocessed by removing stop words. Then top key phrases in each of the documents are identified by computing TF-IDF scores. These top phrases of each of the documents are further used for topic modeling using LDA. Each topic gives set of words with probabilities. Using these topics knowledge base ontology is created. Implementation is carried out using natural language tool kit packages in python for preprocessing and TF-IDF computations. Topic modeling using LDA is implemented using MALLET package in python. Wordnet and Protege tools are used for creation of ontology.

4. Results and Discussion

After collecting and preprocessing hundred fraud cases from different news websites ,the important key phrases form these document samples is shown in Fig 4. These top key phrases from the documents are further used for identifying topics. Based on the requirements number of topics can be selected.

For experimentation number of topics were chosen from two to five. For one of the topics formed with its key phrases is shown in Fig.5. It consists of words that occur in this topic with its probability of occurrence. It was observed that fraud with respect to loan, credit card were among the topics.

Word cloud for the topics without removing the unimportant words like ”said”, ”also” etc. is represented in Fig.6(a). Word cloud with important words for a topic is shown in Fig 6(b). Word cloud considering all the topics is shown in Fig 7. It can be observed that word cloud consists of major defaulters of the loan, major areas in which fraud

[(0,'0.025*"bank" + 0.018*"said" + 0.013*"case"+ 0.011*"fraud" + 0.009*"crore" + 0.009*"police"+ 0.007*"account" + 0.007*"card" + 0.006*"agency"+ 0.005*"money"), (1, '0.018*"said" +0.016*"bank" + 0.009*"case" + 0.007*"fraud"+ 0.006*"police" + 0.005*"company" + 0.005*"loan"+ 0.005*"million" + 0.005*"money" +0.004*"account"')]

Fig. 5. Topic model key phrases and weights for one topic

Fig. 6. (a)Word cloud with all the words from the topic (b)Word cloud with important words

Fig. 7. Word cloud with all topics

occurs etc. It serves as important background study required for fraud analysts. These are the phrases considered as important for understanding and combating fraud. Hence knowledge base using these key phrases was built using ontology representation.

Generic fraud ontology is shown in Fig. 8. It connects the fraudster, case registered and motivation to commit fraud. Ontology representation for the topic shown in Fig. 5. is represented in Fig. 9. It shows mapping of key phrases of the topic and relating them, which can be used for fraud analysis

5. Conclusion

TF-IDF weighting is an extremely effective measure for gauzing the importance of a term in a collection of documents. Similarly, topic modeling is a useful tool for unsupervised clusters. Similarly, based on the results obtained, it can be concluded that the bank and branch of the bank is important in a transaction where fraud is prevalent, in case

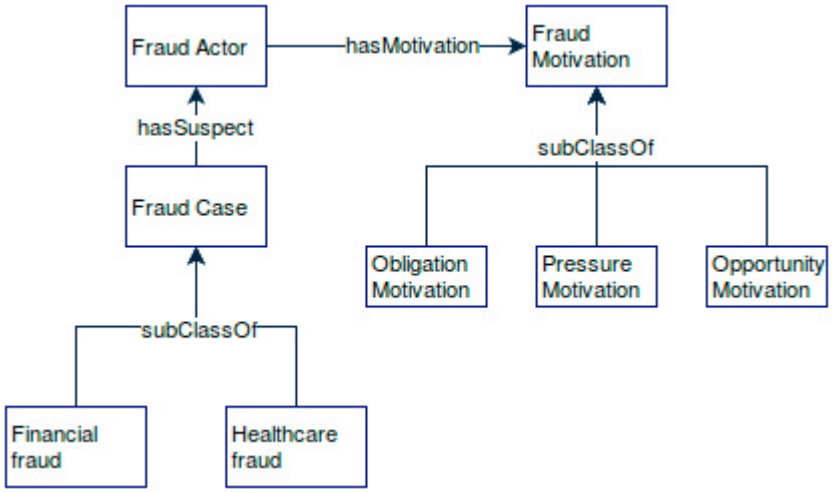

Fig. 8. Generic ontology for fraud

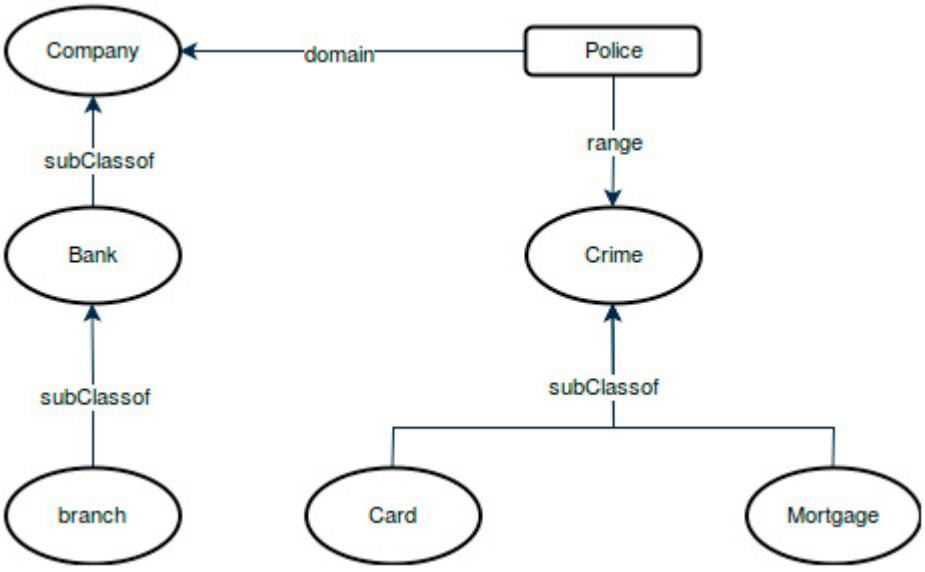

Fig. 9. Ontology for fraud from topic selected

of corporate fraud. In case of police involvement, the range is the crime itself and the domain is the company. Finally, from the ontology and topic modeling results, Word cloud and further ontology knowledge domain can be used to determine which transactions are susceptible to fraud. For future purposes, it is proposed that a bigger data set is used for the purpose of a broader and more accurate result. Also, a more detailed ontology will be useful to establish a better relationship between related terms.

References

1. ASSOCHAM, . India will see 65% rise in mobile frauds by 2017: Assocham-ey study. December 12, 2016. URL http://www.assocham.org/newsdetail.php?id=6087.
2. Wang, S.. A comprehensive survey of data mining-based accounting-fraud detection research. In: *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*; vol. 1. IEEE; 2010, p. 50–53.

3. Abdallah, A., Maarof, M.A., Zainal, A.. Fraud detection system: A survey. *Journal of Network and Computer Applications* 2016;**68**:90–113.
4. Yao, J., Zhang, J., Wang, L.. A financial statement detection model based on hybrid data mining methods. In: *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE; 2018, .
5. Robertson, S.. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation* 2004;**60**(5):503–520.
6. Paik, J.H.. A novel tf-idf weighting scheme for effective ranking. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM; 2013, p. 343–352.
7. Bharadwaj, S., Chiticariu, L., Danilevsky, M., Dhingra, S., Divekar, S., Carreno-Fuentes, A.. Creation and interaction with large-scale domain-specific knowledge bases. *Proc VLDB Endow* 2017;**10**(12):1965–1968.
8. Rani, M., Dhar, A.K., Vyas, O.. Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence* 2017;**63**:108–125.
9. Li, J., Zhang, K., et al. Keyword extraction based on tf/idf for chinese news document. *Wuhan University Journal of Natural Sciences* 2007; **12**(5):917–921.
10. Olszewski, D.. A probabilistic approach to fraud detection in telecommunications. *Know-Based Syst* 2012;**26**:246–258.
11. Phan, X.H., Nguyen, L.M., Horiguchi, S.. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: *Proceedings of the 17th International Conference on World Wide Web*; WWW '08. ACM. ISBN 978-1-60558-085-2; 2008, p. 91–100.
12. Hemant, M., Yvon, F., Capp, O., Jose, J.. Text segmentation: A topic modeling perspective. *Information Processing & Management* 2011; **47**(4):528–544.
13. Somasundaram, K., Murphy, G.C.. Automatic categorization of bug reports using latent dirichlet allocation. In: *Proceedings of the 5th India Software Engineering Conference*; ISEC '12. New York, NY, USA: ACM; 2012, p. 125–130.
14. Thomas, S.W., Adams, B., Hassan, A.E., Blostein, D.. Studying software evolution using topic models. *Science of Computer Programming* 2014;**80**:457 – 479.
15. Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., et al. Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM Trans Comput Biol Bioinformatics* 2012;**9**(6):1831–1836.
16. Xing, D., Girolami, M.. Employing latent dirichlet allocation for fraud detection in telecommunications. *Pattern Recognition Letters* 2007; **28**:1727–1734.
17. Baldi, P.F., Lopes, C.V., Linstead, E.J., Bajracharya, S.K.. A theory of aspects as latent topics. *SIGPLAN Not* 2008;**43**(10):543–562.
18. Liu B Liu L, T.A.. Identifying functional mirnamrna regulatory modules with correspondence latent dirichlet allocation. *IEEE/ACM Trans Comput Biol Bioinformatics* 2010;**16**(24).
19. Nikolenko, S.I., Koltcov, S., Koltsova, O.. Topic modelling for qualitative studies. *Journal of Information Science* 2017;**43**(1):88–102.
20. La Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.. Genomic sequence classification using probabilistic topic modeling 2014;:49–61.
21. Bistarelli, S., Di Noia, T., Mongiello, M., Nocera, F.. Pronto: an ontology driven business process mining tool. *Procedia Computer Science* 2017;**112**:306–315.
22. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C.. Data mining for credit card fraud: A comparative study. *Decis Support Syst* 2011;**50**(3):602–613.