Applying Word Co-Occurrence Graph in Enhancing LDA Model for Topic Discovering in Large-Scaled Text Corpus

Phu Pham, Phuc Do

Abstract—Topic modeling, such as LDA is considered as a useful tool for the statistical analysis of text document collections and other text-based data. Recently, topic modeling becomes an attractive researching field due to its wide applications. However, there are remained disadvantages of traditional topic modeling like as LDA due the shortcoming of bag-of-words (BOW) model as well as low-performance in handle large text corpus. Therefore, in this paper, we present a novel approach of topic model, called LDA-GOW, which is the combination of word co-occurrence, also called: graph-of-words (GOW) model and traditional LDA topic discovering model. The LDA-GOW topic model not only enable to extract more informative topics from text but also be able to leverage the topic discovering process from large-scaled text corpus. We test our proposed model in comparing with the traditional LDA topic model, within several standardized datasets, include: WebKB, Reuters-R8 and annotated scientific documents which are collected from ACM digital library to demonstrate the effectiveness of our proposed model. For overall experiments, our proposed LDA-GOW model gains approximately 70.86% in accuracy.

Index Terms: topic model, LDA, graph-of-words (GOW), frequent subgraph mining, word co-occurrence graph, graph-based concept.

I. INTRODUCTION

There is no double that, discovering topic from text is one of the most important task of text mining. Recently, topic modeling is one of the most interesting topics which achieves a lot of attentions from researchers. Topic model help to address the problem of extracting text documents in multiple latent themes or topics. These extracted latent topics are corresponded with the probabilistic distributions over words in each document of corpus. Topic modeling is wide applied in multiple disciplines, include: information retrieval (IR), supporting for large-scaled text document semantic indexing, raw-text topic mining, text classification, clustering, etc. From the view of text mining and processing, text document retrieval is a high-dimensional undirected searching task. Understanding as well as extracting distinctive features from text is considered as a difficult and complex process, due to the diversity in human-written language and grammatical structure. From years, scientists have been finding solutions for better understanding of existed information which are available in text corpora, there are two main approaches, include: text summarization and topic modeling. *Table 1* shows examples of topics and their related words.

Table 1. Example of topics and set of common related words

Topic	**Related words**
Arts	film, shows, musical, theater, actress, actors, opera, play, etc.
Education	college, student, university, teacher, high school, elementary school, etc.
Finance	Money, tax, budget, plan, loan, billionaire, market place, stock, etc.

LDA topic model is one of the most common approach of topic modeling which is firstly introduced by David Blei. et al., (2003) [1] [2]. The mechanism of LDA topic model is mainly based on the principles of multinomial probabilistic distributions of word occurrence within document to uncover the latent topics over document's collections. However, from the first time of introduction, through multiple developing stages, the traditional LDA topic model as well as extended versions still has remained drawbacks. First of all, the main challenging of LDA topic model is the bag-of-word (BOW) in topic representation. The disadvantage of this method is the failure in document's term co-occurrence recognition. There is no doubt that word co-occurrence plays an important role in carrying important information of text document. Moreover, the low-performance in time-consuming for large-scaled text corpus is also considered as significant challenge of traditional LDA topic model. Because LDA topic model is worked on the mechanism of probabilistically evaluating each term independently, therefore, with a large-scaled text corpus with huge amount of separated terms, it needs much more time to complete the overall processes. In order to overcome aforementioned challenges, in this paper, we present the LDA-GOW model which is a combination of using co-occurrence term in document representation and traditional LDA topic model. Our contributions in this paper are three-folds. First, we propose an approach of using word co-occurrence graph in documentrepresentation, also called graph-of-words (GOW). After that, we use the gSpan to extract the graph-based concepts from these transformed GOW-based documents. Then, these graph-based concepts are used to represent back the given

Revised Version Manuscript Received on August 19, 2019.

Phu Pham, University of Information Technology (UIT), VNU-HCM, Vietnam.Asia.

Phuc Do, University of Information Technology (UIT), VNU-HCM, Vietnam. Asia.

documents, instead of distinct words. Second, the documents which are represented by graph-based concepts are used to feed the LDA topic model. The extracted topics are composed by set of concepts instead of keywords in traditional LDA topic model. Finally, we test the performance of our proposed LDA-GOW model by using the model to solve the text classification task via different out-off-the-shelve classification algorithm such as: SVM, Naïve Bayes and Decision Tree, within different standardized dataset. We compared the output results with previous classic LDA model in order to demonstrate the effectiveness of our proposed approach in both model accuracy and time-consuming performance.

The rest of our paper is organized into four main parts. For the second part, we mention about previous studies as well motivations. Next, we introduce our approaches of LDA-GOW in the third part. In this part, we also discuss about principal concepts, methodologies as well as system implementation. In the next section, we conduct the empirical studies, carefully describe about our experimental dataset usage, setup, result and discussions. Finally, the fifth section contains our conclusion and future improvements.

II. RELATED WORKS AND MOTIVATION

In text mining, topic model is mostly known in the first introduction of Latent Semantic Indexing (LSI) model which are proposed by Deerwester, Scott et al. (1990) [3]. In fact, LSI model encountered the shortage of evaluating the probabilistic distributions of terms over text document in the manner of topic discovering. From the first idea of LSI, HOFMANN, Thomas et al. (1999) [4] has combined the probabilistic distribution model to the previous latent semantic model to form the probabilistic Latent Semantic Indexing (pLSI), which is considered as the fundamental baseline of LDA topic model of David Blei. et al., (2003). Throughout multiple developing stages, there are several notable improvements in LDA topic modeling for improving the quality of extracted topics be much more informative. Most of the improvements are focused on integrating the supervised learning technique with LDA model to leverage the topic discovering process [5], such as: correspondence LDA (cLDA) [6], supervised LDA (sLDA) [7]. On the other hand, many researchers concentrated on investigating and solving problems related to syntactic sense and semantic relationship among terms inside the given text document collections. This type of approach depends on using NLP-based techniques or analyzing the common-sense concepts from input documents via appropriate available knowledge-based repositories [8], such as: SemLDA model [9] applying WordNet, SemCor, WSD (word-sense disambiguation), etc. repositories in semantic and word's sense identification problem. However, most of previous approaches still have remained drawbacks related to the dependence on available knowledge-based repository such as: WordNet, lexical database, domain ontologies, etc.

III. METHODOLOGY AND SYSTEM DESIGN

In this section we present the methodology of LDA-GOW approach which is the combination of graph-of-words in text representation and the classic LDA model. First of all, the text document is transformed into the graph-based co-occurrence structure, then, we apply the gSpan algorithm to extract the common subgraphs from these graph-based text documents, called graph-based concepts. Then, these graph-based concepts are used to represent back these text documents. The ultimate goal of using graph-based concepts for text document representation is to reduce the dimension of document into fix more informative lower-dimensional space.

Document to Graph-of-Words (GOW) Transformation

Graph-of-words (GOW).

In GOW-based text document representation, giving a specific text document (d) which include $|\mathrm{W}|$ number of distinct words $(\mathrm{w}, \mathrm{w} \in \mathrm{W})$ [10] [11] [12], the document's graph is formed as a directed/undirected graph, denoted as: $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, where $(\mathrm{V}, \mathrm{v} \in \mathrm{V})$ is the vertices, which represent for each word (w) in (W), or $|\mathrm{W}| = |\mathrm{V}|$. The (E) represents for set of edges, represent for the relationship between pairwise vertices/words, often representing for co-occurrence relationship between two nearby words.

For traditional approach, a text document is often represented by set of frequent distinctive keywords, which is considered as bag-of-words (BOW) model. There are several challenges related to BOW model, include failures in word's co-occurrence as well as dependency representations. In text document, most of the words tend to appear surrounding by set of other common words, such as "data" often goes with "mining" or "artificial" goes with "intelligence", etc. Hence, representing the text document by BOW model might lead to the drawbacks of less-informative knowledge capturing from text. Therefore, the GOW model is proposed to tackle these problems. The GOW model [10] underlies the assumption is that all the words which appear in a text document have at least on relationship with the others, two distinct co-occurrence words are represented as two vertices which are linked via an edge represented for co-occurrence relationship of these two words. In the other complicated approaches, the relationships between two words also be vary, might be represented as: semantic or grammatical dependency relationships depending on the grammatical context of the text sentence which these words appear in. However, in the studies of this paper, we only focus on using co-occurrence relationships to form the text document graph. The method for transforming text document into graph-based structure is quite simple. At first, the input text document is tokenized, stemming and converting to lowercase, after that, starting from a specific source term at the begin of document, we slide to a next term, or target term, in order to form a pairwise terms. With specific pairwise terms, if there is no existed edge between these two terms, a new edge will be created to link these source and target term together. The process continues until meet the ending term of a given document. For example, giving a simple text document with following content: "*Java is a programming language.Csharp is also a programming language*", the constructed document graph is illustrated in *Fig.1*.

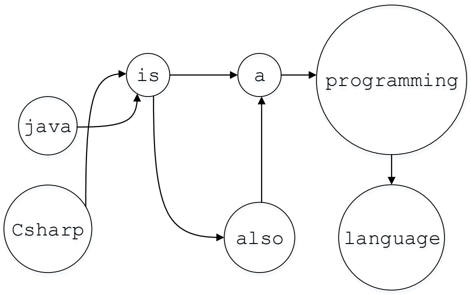

Fig.1. Example of constructed GOW for the given document content

In fact, using co-occurrence document graph transformation is the simple way for representing the text document as graph. There are also multiple complex ways for representing text such as applying text dependency parsing to extract grammatical relationship between terms. However, the use of syntactic dependency parsing in document's graph transformation is much more difficult as well as mostly depending on the third-party NLP tools, such as: Stanford-NLP or Apache Open-NLP, etc. Moreover, currently the text dependency parsing is able to extract the syntactic dependency relationships within only a single sentence, but not for multiple sentences.

Graph-based concept Extraction via gSpan

From the given transformed GOW-based documents, we use gSpan algorithm to extract the set of common subgraphs, called graph-based concepts. These graph-based concepts are considered as common features of given text documents. Then, we used these common graph-based concepts to represent back the given documents. Different from the previous approach of BOW model, the text documents are represented by set of frequent keywords, in the GOW, the distinct features of documents are common subgraphs. In order to extract frequent subgraphs from given document's graphs, there are several, such as: gSpan in subgraph pattern mining [13], FFSM (HUAN, Jun, et al., 2003) [14] applying isomorphic graph presence in frequent subgraph mining and SPIN (HUAN, Jun, et al., 2004) [15]. In this paper, we select the gSpan to extract the common subgraphs from given document's graphs, the overall processes are illustrated in *Fig.2*.

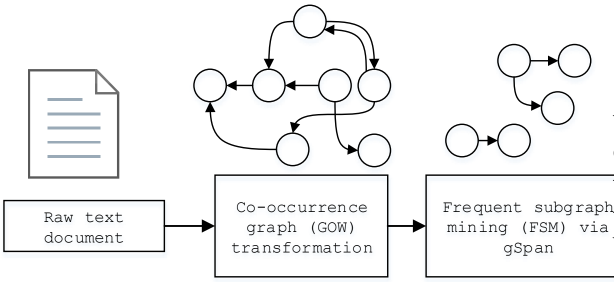
Raw text document
Co-occurrence graph (GOW) transformation
Frequent subgraph mining (FSM) via gSpan

Fig.2. Overall process of extracting feature graph-based concepts from raw text document

LDA-GOW: The combination of Latent Dirictlet Allocation (LDA) and GOW model in text transformation

Back the problem of topic discovering from text, in traditional LDA topic model which is proposed by Blie et al. [2][6] a specific text document (d) is described as a mixture of latent topic distribution, denoted as: θ^{d}_{y}, and each latent topic is a probabilistic distribution over separated words (w), denoted as: ϕ^{t}. In short, we can say that the classic LDA model supports to generate two distributions, which are: $\langle \text{document}, \text{topic} \rangle$ and $\langle \text{topic}, \text{word} \rangle$. As aforementioned problems, the distributions of $\langle \text{topic}, \text{word} \rangle$ is considered as a drawback of BOW model, which means all terms which are distributed over set of extracted latent topics are evaluated independently, just like we have set of frequent separated keywords as shown in *Table 1* without any information about the relationships between them. Therefore, in the LDA-GOW model, instead of extracting the set of documents' distinct terms as the input for the LDA topic, we first transform and extracting common feature graph-based concepts for the given text documents. After that, we use the set of documents which are represented by these set of graph-based concepts to feed the LDA model. Therefore, the outputs of LDA-GOW model will be the two distributions of $\langle \text{document}, \text{topic} \rangle$and $\langle \text{topic}, \text{concept} \rangle$. For specific text document (d_i) which is transformed into graph-based structure, denoted as: ($d_i \rightarrow G_i$) will be associated with the set of $\delta_i{:}\, G_{freq}$ - if $\delta_i{:}\, G_{freq}$is isomorphic with any extracted subgraph in G_i related to "subgraph isomorphism" or "subgraph matching" problems: $D \rightarrow \delta{:}\, G_{freq}$ – there are some notable proposed algorithms such as: VF2 in matching large graphs (CORDELLA, Luigi P., et al., 2004) [16], QuickSI (SHANG, H. et al., 2008) [17] [18]. Finally, we apply GibbsLDA to compute the mixture probabilistic distributions of (ϕ^{t}) (the distributions of [topic (t_j)]-[concept ($\delta_i{:}\, G_{freq}$)]) and [topic]-[document] (θ^{d}). The generative model of LDA-GOW is described in *Fig.3*

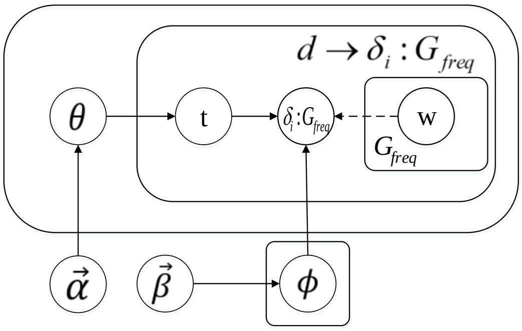

Fig.3. Overall generative model processes of LDA-GOW

In fact, the LDA-GOW enables not only to capture the word's relationship within text documents but also make the extracted topic be more informative in compare with the traditional LDA model. In LDA-GOW model, the extracted topics are the distributions of concepts which are composed by common co-occurrence words which absolutely be able to prevent problems related to BOW model. Moreover, the LDA-GOW also demonstrate the effectiveness in performance of time-consuming while handling the large-scaled text corpus, because in charge of evaluating the probabilistic distributions of separated keywords, which is considered very large, in each text document, the LDA-GOW model only needs to evaluate the small number of concepts which are existed in each document, ($|W| \gg |\delta: G_{freq}|$). In the next section, we demonstrate the empirical studies on the performance of LDA-GOW model in comparing with the traditional one in handling large-scaled text corpus.

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we conduct the experiments on the proposed LDA-GOW model in different standardized dataset in order to show the effectiveness of the combination between graph-of-words in text document transformation and LDA model in topic discovering from large-scaled text corpus.

For overall experiments, we use three main annotated datasets with different number of class, include: WebKB (4 classes) [19], Reuters-R8 (8 classes) [20] and 20K abstract content of categorized documents (belong to 10 classified topics) which are collected from ACM digital library [21]. One of the most common way for evaluating the accuracy of LDA topic model is to let the topic distribution outputs solve the problem of text classification and clustering.

Model accuracy evaluation on text classification task.

In this paper, we applied the both LDA-GOW and classical LDA model to extract the topic distributions from the three given testing text corpora. Then, these topic distributions are used as the feature vectors to feed the classifier. In this study, we use three most well-known out-of-the-shelf classification algorithms, include: SVM, Naïve Bayes and Decision Tree (J48) to test the model accuracy in solving text classification problem. With different classification algorithms, the F-measure metric is used to evaluate the classification accuracy. Then, we get the average accuracy outputs of all classification approaches as the final result.

Table 2. Accuracy outputs between LDA-GOW and traditional LDA in WebKB dataset

	Traditional LDA	LDA-GOW
Precision	0.71726	0.75627
Recall	0.68278	0.71232
F-measure	0.69960	**0.73364**

As shown in experimental results, the LDA-GOW model outperforms the traditional LDA topic model in all three experimental datasets. Table 5 shows the F-measure accuracy scores between two approaches. The experiments are conducted with different size of testing set (the left dataset is used for training the classifier), from 30% to 90%. The outputs demonstrate that the proposed LDA-GOW outperforms the traditional LDA topic model. For all datasets, the average accuracy scores of proposed LDA-GOW model always gains better performance than the traditional LDA model.

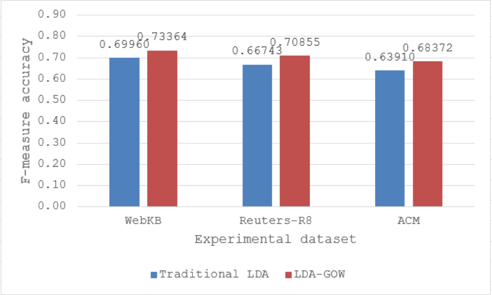

Fig.4. Comparisons between LDA-GOW and traditional LDA model in different datasets

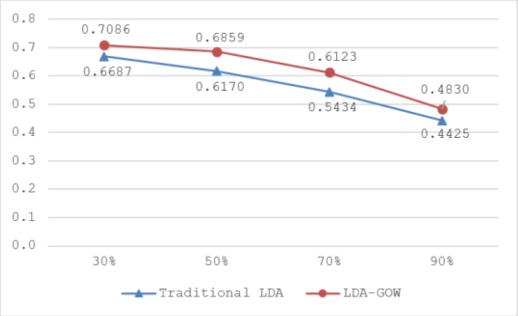

Fig.5. Average accuracy scores of LDA-GOW and traditional LDA in three datasets with different size of test set (%)

Table 3. Accuracy outputs between LDA-GOW and traditional LDA in Reuters-R8 dataset

	Traditional LDA	LDA-GOW
Precision	0.67212	0.71913
Recall	0.66281	0.69827
F-measure	0.66743	**0.70855**

Table 4. Accuracy outputs between LDA-GOW and traditional LDA in ACM dataset

	Traditional LDA	LDA-GOW
Precision	0.63563	0.68927
Recall	0.64262	0.67826
F-measure	0.63910	**0.68372**

Table 5. F-measure accuracy outputs of LDA-GOW and traditional LDA in 3 datasets with different size of test set (%)

TestSet (%)		30%	50%	70%	90%
Traditional LDA	WebKB	0.69961	0.63221	0.56271	0.48765
	Reuters-R8	0.66743	0.62148	0.55278	0.41223
	ACM	0.63915	0.59726	0.51472	0.42776
LDA-GOW	WebKB	**0.73364**	0.71231	0.63271	0.51962
	Reuters-R8	**0.70855**	0.67821	0.61287	0.45726
	ACM	**0.68372**	0.66721	0.59123	0.47225

Model accuracy evaluation on text clustering task.

In this part, we conduct the experiments to compare accuracy of both LDA-GOW and traditional LDA approaches by solving the text clustering task. In this test, we use the most well-known k-means clustering algorithm in the extracted topic distributions to obtain clusters. For evaluating the accuracy of text clustering outputs, we use two metrics: purity and NMI. Table 6 shows the output accuracy of both LDA-GOW and traditional LDA approaches in three datasets.

Table 6. Purity and NMI scores for two approaches in different datasets

		Purity	NMI
Traditional LDA	WebKB	0.56113	0.53242
	Reuters-R8	0.58267	0.51256
	ACM	0.51525	0.48725
LDA-GOW	WebKB	0.55872	**0.57713**
	Reuters-R8	0.58251	**0.56261**
	ACM	0.62123	**0.51872**

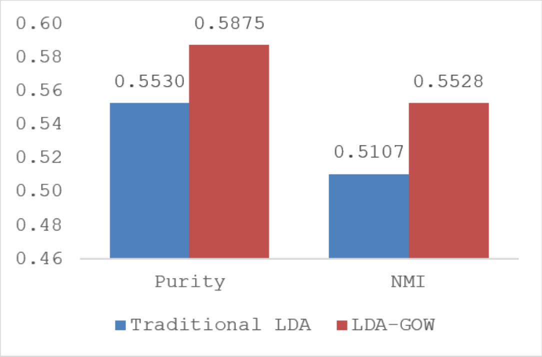

Fig.6. Average purity and NMI scores for both approaches in three datasets

The experimental outputs demonstrate that our proposed LDA-GOW model gains higher performance than the traditional LDA model. Especially, in Reuters-R8 dataset, it improves about 9.7% in NMI accuracy of text clustering task and about 8.23% in average for overall three datasets.

Time-consuming performance.

In this part, we compare the capability of proposed LDA-GOW and traditional LDA model in handle large-scaled text corpus by evaluating the overall execution time (in seconds) of two models on a same dataset. We select the 20K documents of ACM as the main dataset, both models are implemented and run simultaneously with multiple test-cases, each test-case is different in number of topics and keywords/concepts. As shown from the experimental results which are in

Table 7, for overall test-cases, the LDA-GOW model outperforms the previous LDA model in the time-consuming performance, therefore the proposed LDA-GOW model is much suitable for handling large-scaled text corpus.

Table 7. Execution time of LDA-GOW and traditional LDA model in different test-cases

# of topic	**# of Concept/keyword**	**Execution time (seconds)**	
		Traditional LDA	**LLDA-GOW**
5	5	126.726	128.l97
10	20	212.213	198.872
20	50	421.212	387.827
30	100	672.821	567.521

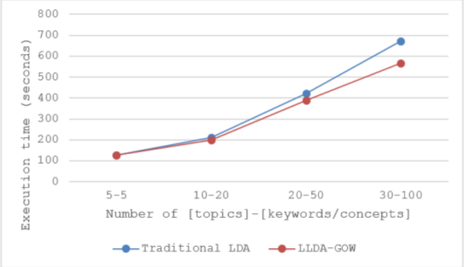

Fig.7. Comparisons in execution time of LDA-GOW and traditional LDA model

V. CONCLUSION

In this paper, we present an approach of LDA-GOW model which is the combination of graph-of-words (GOW) in text transformation and the classical LDA model. Instead of feeding the LDA model with set of separated independent keywords from raw text document, in LDA-GOW model, we first apply the GOW model to transform the document into the GOW-based structure. After that, the gSpan frequent subgraph mining technique is applied to extract common graph-based concepts from the document's graphs, as the document representative features which are used to feed the LDA topic model. Over experimental studies with different standardized dataset, our proposed LDA-GOW model has been shown that not only enables to leverage the quality of discovered topics from text but also supports to effectively handle the large-scaled text corpus.

VI. ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCMC) under the grant number B2017-26-02.

REFERENCES

1. BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I, "Latent dirichlet allocation," Journal of machine Learning research, pp. 993-1022, 2003.
2. D. M. BLEI, "Probabilistic topic models," Communications of the ACM, pp. 77-84, 2012.
3. DEERWESTER, Scott, et al., "Indexing by latent semantic analysis," Journal of the American society for information science, 1990.
4. HOFMANN, Thomas., "Probabilistic latent semantic analysis," Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann
Publishers Inc., pp. 289-296, 1999.

5. MIMNO, David, et al., "Optimizing semantic coherence in topic models," Proceedings of the conference on empirical methods in natural language processing, pp. 262-272, 2011.
6. BLEI, David M.; JORDAN, Michael I., "Modeling annotated data," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, pp. 127-134, 2003.
7. MCAULIFFE, Jon D.; BLEI, David M, "Supervised topic models," Advances in neural information processing systems, pp. 121-128, 2008.
8. RAJAGOPAL, Dheeraj, et al., "Commonsense-based topic modeling," Proceedings of the second international workshop on issues of sentiment discovery and opinion mining, p. 6, 2013.
9. FERRUGENTO, Adriana, et al., "Can Topic Modelling benefit from Word Sense Information?," LREC, 2016.
10. ROUSSEAU, François; VAZIRGIANNIS, Michalis., "Graph-of-word and TW-IDF: new approach to ad hoc IR," Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, pp. 59-68, 2013.
11. MELADIANOS, Polykarpos, et al., "Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream," ICWSM, pp. 248-257, 2015.
12. ROUSSEAU, François; KIAGIAS, Emmanouil; VAZIRGIANNIS, Michalis, "Text Categorization as a Graph Classification Problem," ACL (1), pp. 1702-1712, 2015.
13. YAN, Xifeng; HAN, Jiawei., "gspan: Graph-based substructure pattern mining," Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, pp. 721-724, 2002.
14. HUAN, Jun; WANG, Wei; PRINS, Jan., "Efficient mining of frequent subgraphs in the presence of isomorphism," Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, pp. 549-552, 2003.
15. HUAN, Jun, et al., "Spin: mining maximal frequent subgraphs from graph databases," Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 581-586, 2004.
16. CORDELLA, Luigi P., et al., "A (sub) graph isomorphism algorithm for matching large graphs," IEEE transactions on pattern analysis and machine intelligence, pp. 1367-1372, 2004.
17. SHANG, Haichuan, et al., "Taming verification hardness: an efficient algorithm for testing subgraph isomorphism," Proceedings of the VLDB Endowment, pp. 364-375, 2008.
18. LEE, Jinsoo, et al., "An in-depth comparison of subgraph isomorphism algorithms in graph databases," Proceedings of the VLDB Endowment. VLDB Endowment, pp. 133-144, 2012.
19. WebKB dataset: https://www.cs.umb.edu/~smimarog/textmining/datasets/webkb-train-stemmed.txt, Accessed February 10, 2019.
20. Reuters-R8 dataset: https://www.cs.umb.edu/~smimarog/textmining/datasets/r8-train-stemmed.txt, Accessed February 10, 2019.
21. ACM digital library: https://dl.acm.org/, Accessed February 10, 2019.