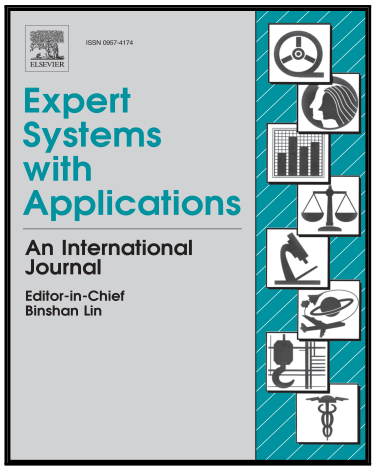
Journal Pre-proof



Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis

Suhyeon Kim, Haecheong Park, Junghye Lee

PII: S0957-4174(20)30225-6
DOI: https://doi.org/10.1016/j.eswa.2020.113401
Reference: ESWA 113401

To appear in: *Expert Systems With Applications*

Received date: 24 November 2019
Revised date: 3 February 2020
Accepted date: 20 March 2020

Please cite this article as: Suhyeon Kim, Haecheong Park, Junghye Lee, Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis, *Expert Systems With Applications* (2020), doi: https://doi.org/10.1016/j.eswa.2020.113401

Highlights

- Blockchain has a considerable value as one of the promising technologies in industrial 4.0.
- We propose a new topic modeling method based on word embedding and clustering.
- The proposed method outperforms an existing method in both qualitative and quantitative views.
- The proposed method contributes to analyzing the research trend of blockchain technology.

Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis

Suhyeon Kim[a], Haecheong Park[a], Junghye Lee[b,*]

[a]*Department of Business Analytics, Graduate School of Interdisciplinary Management, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea*
[b]*School of Management Engineering, UNIST, Ulsan, Korea*

Abstract

Blockchain has become one of the core technologies in Industry 4.0. To help decision-makers establish action plans based on blockchain, it is an urgent task to analyze trends in blockchain technology. However, most of existing studies on blockchain trend analysis are based on effort demanding full-text investigation or traditional bibliometric methods whose study scope is limited to a frequency-based statistical analysis. Therefore, in this paper, we propose a new topic modeling method called Word2vec-based Latent Semantic Analysis (W2V-LSA), which is based on Word2vec and Spherical k-means clustering to better capture and represent the context of a corpus. We then used W2V-LSA to perform an annual trend analysis of blockchain research by country and time for 231 abstracts of blockchain-related papers published over the past five years. The performance of the proposed algorithm was compared to Probabilistic LSA, one of the common topic modeling techniques. The experimental results confirmed the usefulness of W2V-LSA in terms of the accuracy and diversity of topics by quantitative and qualitative evaluation. The proposed method can be a competitive alternative for better topic modeling to provide direction for future research in technology trend analysis and it is applicable to various expert systems related to text mining.

Keywords: Trend Analysis, Topic Modeling, Word2vec, Probabilistic Latent Semantic Analysis, Blockchain

*Corresponding author
Email addresses: suhyeonkim@unist.ac.kr (Suhyeon Kim), haecheongpark@unist.ac.kr (Haecheong Park), junghyelee@unist.ac.kr (Junghye Lee)

1. Introduction

Blockchain refers to a distributed ledger technology in which transactional information on the network is encrypted by hashing and is shared among network members (Zheng et al., 2017). For each transaction, data transformation persists so that data is not able to be arbitrarily manipulated. In addition, it is a highly reliable technology because network members continuously authenticate the data. Since 2008, when the first paper on Bitcoin was published (Nakamoto et al., 2008), the increased attention and understanding of this powerful technology has generated great repercussions around the world. The development of blockchain technology has generated three major innovations commonly referred to as Blockchain 1.0, 2.0, and 3.0. Blockchain 1.0 refers to the evolution of currency and digital payment systems such cryptocurrencies like Bitcoin. Blockchain 2.0 is the application of blockchain technology to the financial sector more broadly. Blockchain 3.0 goes further still by applying the technology to sectors beyond currency or finance (Swan, 2015). In the Blockchain 1.0 and 2.0, especially, cryptocurrency transaction and blockchain architecture are becoming major issues (Peters et al., 2015; Zheng et al., 2017). However, the advent of Blockchain 3.0 has seen new value continually added to fields of interest to Industry 4.0, such as the Internet of Things (IoT), smart contract, eco-systems, and storage systems, as well as to the fields of healthcare, finance, privacy and security (Alharby & van Moorsel, 2017; Dagher et al., 2018; Fan et al., 2018; Miraz & Ali, 2018). While previous iterations of blockchain technology related specifically to virtual currency or financial transactions, recent developments track the broader applications of blockchain technology. These trends indicate that the blockchain-related research will therefore be of interest to any sector in Industry 4.0, and thus the importance of predicting future applications of blockchain technology cannot be overemphasized.

Accordingly, current studies on blockchain trend analysis have been conducted, and our study shares this purpose. The most common approaches to analyze blockchain trends can be summarized as follows: (1) screening review and (2) bibliometrics analysis of the relevant papers. In the first approach, Lu (2019) and Zheng et al. (2017) show overall blockchain research lines. In specific fields, Alonso et al. (2019) and McGhin et al. (2019) conducted a frequency analysis of publications related to blockchain technology, and a number of potential research opportunities are also discussed in eHealth and overall healthcare fields. Considering the social and economic aspects of blockchain technology and associated environmental issues, Giungato et al. (2017) presents

current trends concerned with the sustainability of Bitcoin. On the other hand, in the second approach, the bibliometrics method of the blockchain domain is a statistical analysis of trends using papers or book publications related to blockchain (Dabbagh et al., 2019; Miau & Yang, 2018), simply capturing bibliographic information or using a statistical frequency analysis. Yli-Huumo et al. (2016) proposed a systematic mapping study, which is able to find relevant papers through keywording based on the abstract. Identifying keywords and categories manually for the mapping of the papers, they summarize the challenges and positions and provide recommendations for future research direction. Zeng & Ni (2018) used term-frequency based textual analysis and social network to present blockchain research topics and the researcher-level co-authorship on the basis of Ei Compendex and China National Knowledge Infrastructure database between 2011 and 2017. However, these studies utilized short-term papers or limited databases and the concrete evaluation for their methods remains a challenging task. In brief, the previous studies are based on traditional and naive approaches which just review relevant literature or do simple frequency analysis without providing insights beyond revealed information about blockchain trends, and thus it is urgent to do comprehensive and in-depth trend analysis on blockchain technology.

Therefore, of particular interest to our study is a trend analysis through text mining approach focusing on topic modeling; we can identify the author's opinion or intention by extraction of potential topics from the text. In general, the initial trend analysis was conducted as a simple pattern analysis for 1-dimensional time series data (Kivikunnas, 1998). However, recent developments in text analysis techniques have enabled trend analysis using text data, including user reviews, newspaper articles, papers, patents, keyword analysis that analyzes main words in specific documents, and social network analysis that can examine the association and impact among users (Hung, 2012; Hung & Zhang, 2012; Kim et al., 2015; Kim & Delen, 2018; Terachi et al., 2006; Tseng et al., 2007). In particular, topic modeling has gained a lot of attention recently by researchers in trend analysis since the main purpose of trend analysis based on text data is to detect the up and down trends about frequency of each topic in the target documents (Kang et al., 2019).

Specifically, topic modeling identifies and classifies latent topics of each document. This method has coevolved with advances in machine learning and text mining techniques. Probabilistic latent semantic analysis (PLSA), one of the most widely used techniques of topic modeling, is a probabilistic topic model also known as aspect modeling, which is a latent variable model based on the term-document matrix of co-occurrence data (Hofmann, 1999). The superiority of PLSA

was demonstrated by comparison with k-means and Latent Semantic Analysis (LSA) (Newman & Block, 2006). As a variant or extension of PLSA, Latent Dirichlet Allocation (LDA) uses the Bayesian approach for parameter estimation to complement the incompleteness of PLSA on topic probability distribution (Blei et al., 2003). However, it is difficult to interpret LDA without prior knowledge of latent topics and hyper-parameters. Alghamdi & Alfalqi (2015) presented a paper comparing the techniques of topic modeling such as LSA, PLSA, and LDA.

The aforementioned probability-based statistical topic modeling techniques fail to capture the entire context of the document because it usually uses a uni-gram representation that considers a word independently (Lu & Zhai, 2008). Alternatively, it is possible to use a n-gram representation, which considers multiple words simultaneously, but the efficiency of the model decreases rapidly due to the curse of dimensionality (Bengio et al., 2003). Accordingly, Word2vec quantifies the word into a vector considering the context to solve the limitation of this representation (Mikolov et al., 2013a). In other words, it creates representations of words so that similar words are located in a similar space. Although this new representation is widely used and its performance has been demonstrated in recent text analyses (Asghari et al., 2018; Van Hooland et al., 2017; Zhang et al., 2015), very few attempts have been made to develop a new topic model based on Word2vec.

In short, two types of existing studies on blockchain performed trend analysis in a limited scope and have their own limitation. Screening review-based ones require a great deal of time and effort for screening and summarizing all literature. Bibliometrics analysis-based ones are not suitable to discover underlying patterns that lie in blockchain-related fields. Furthermore, topic models commonly-used for trend analysis in other fields are generally based on uni-gram based word vector representations, which are non-contextual and sparse. In this paper, to overcome these problems, we propose a new topic modeling approach called Word2vec based latent semantic analysis (W2V-LSA) which makes use of Word2vec, contextual word embedding algorithm along with spherical k-means clustering. This technique allows one to quantify a word's contextual meaning in a vector format and to group the words with cosine similarity. We use the proposed method to perform blockchain-specific trend analysis, which can play a role as an advanced and useful alternative to extract meaningful topics involved in the current trends of blockchain.

Our Contributions. The major contributions of this study are summarized as follows:

- As blockchain technology becomes more popular and the number of related technologies and studies increases, the topics of blockchain research become more diverse and precise. Our contri-

bution lies in the fact that we can provide different aspects against the bibliometrics method for the blockchain trend, therefore capturing the topics from the available literature based on a new topic model. In specific, this study also shows characteristics about blockchain technology trends of several leading countries in the blockchain-related research.

- We propose a novel approach to extract more related topics to blockchain research than existing studies, which combines contextual embedding and clustering in a harmonized way. Firstly, we adopt a neural network-based word embedding algorithm which can generate representations of words to capture the context of the documents. Next, we use the cosine similarity-based clustering method to construct topic clusters. Finally, we propose a new topic allocation method via document vector construction and similarity calculation procedure between the topic cluster and the document.
- We demonstrate the performance of the proposed method to show its usefulness on real text data related to blockchain technology. The results show that our method can produce the highly coherent topics and to what extent the topics contain the core meaning of the documents found by topic coherence measures and keyword matching score, respectively. We also present qualitative evaluation on the actual documents to confirm its accuracy of topic detection.
- This study provides comprehensive and intensive understanding of emerging technology specifically for blockchain. It helps the professional leading the blockchain-related research to readily determine future directions of their studies. In addition, our proposed method is an informative tool for anyone responsible for strategic decision-making in the blockchain-related industries. By capturing the trends of various technology fields using blockchain, our method can be utilized to identify the prospects and marketability of each field.

This paper is organized as follows. In Section 2, we represent the material and data preprocessing works. A new topic-modeling method that complements the limitations of existing techniques is proposed, which will be discussed in detail in Section 3. Section 4 presents the results of the trend analysis and compares results with the previous method. Section 5 and 6 contain a discussion of the results and the conclusion of this study.

2. Material

Fig. 1 is the process of data collection and preprocessing used to conduct topic modeling about blockchain. For blockchain technology trend analysis, we collected abstracts of blockchain-related papers from six paper database such as Scopus, ScienceDirect, Web of Science, IEEE Xplore,

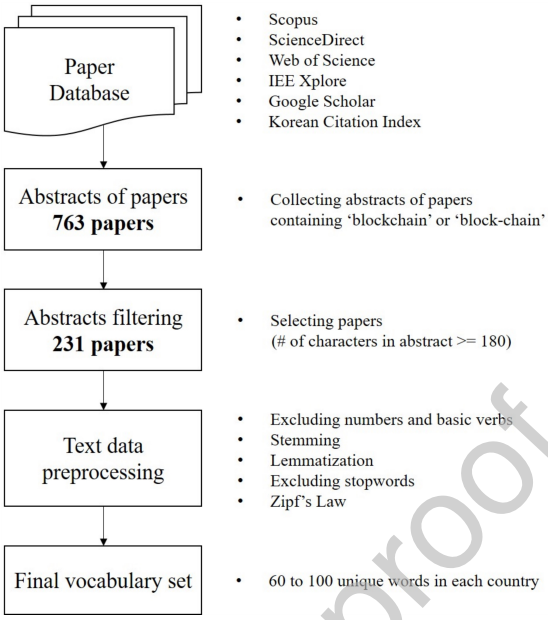


Fig. 1: Process of data collection and preprocessing

Google Scholar, and Korean Citation Index. A total of 763 abstracts of papers were collected, whose keywords and abstracts contain the words such as'Blockchain,' 'Block chain,' and 'Block-chain' from 2014 to August 2018. In this case, conference papers were excluded. To ensure a minimum amount of information in the text for topic modeling, we selected the abstracts whose character count is greater than 180, and a total of 231 abstracts were utilized for experiments. For the collected data, we performed preprocessing by excluding numbers and basic verbs, stemming and lemmatization. Specific words such as 'Blockchain' and 'Technology,' which are not meaningful as a topic index in the blockchain trend analysis, were designated as stopwords and excluded from analysis. Based on the frequency of words in a corpus, we employed Zipf's Law, a method to remove either too common or too rare words. In each country, a final vocabulary set, to be used in analysis, was constructed by extracting about 60 to 100 unique words.

Fig. 2 represents the number of blockchain-related papers published per year and by country. Since 2016, the number of published papers has risen sharply, and the growth rate of papers in 2017 is about 52%. The number of papers in the top three countries -Korea, the US and China- accounts for about 69% of the total number of papers. In particular, the growth rate of papers in the second quarter of 2018 is 34% in the US and 62% in China.

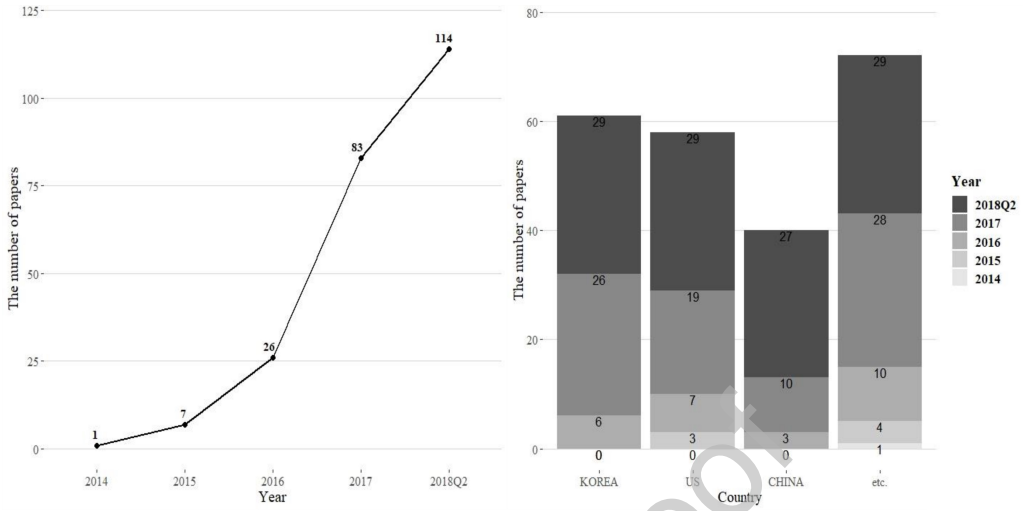


Fig. 2: Growth of the number of blockchain-related papers; Q2 means the second quarter of a calendar year. The detailed information of the etc. group is represented in Appendix.

3. Methodology

3.1. Word2vec

Word2vec, a neural network-based model, represents words in corpus as a vector with contextual comprehension (Mikolov et al., 2013a). In vector space, the closer the distance between two vectors, the higher the similarity of the two words. The result of Word2vec depends on two user-defined parameters: the dimensionality (i.e., size) of the vector representation m, and the maximum distance (i.e., window) between a word and words around the word in a sentence δ. Word2vec is configured in two-ways: skip-gram and continuous bag of words (CBOW). The major difference is that skip-gram is intended to predict the surrounding words by inputting the reference word, whereas CBOW predicts the current word using the surrounding words.

3.2. Spherical k-means clustering

Because it quantifies the degree to which two vectors point in the same direction by measuring their cosines, cosine similarity has been widely used in text data analysis (Dhillon & Modha, 2001). Each word vector $\boldsymbol{x}_i \in R^m, i = 1, \ldots, N$ derived from Word2vec and the inner product with two word vectors represents the semantic similarity with cosine. The centre cluster is calculated by allowing the cluster vector $c(i) \in 1, \ldots, C$ be assigned to $\boldsymbol{x}_i$ and the cosine distance between $\boldsymbol{x}_i$ and $p_q, q = 1, \ldots, C$. The objective is to find the best adjustable cluster to minimize the cosine distance

between $\boldsymbol{x}_i$ and $\boldsymbol{x}_q$. σ_{iq} is a constraint that defines whether clusters q and $\boldsymbol{x}_i$ are equal (i.e. $\sigma_{iq} = 1$) (Buchta et al., 2012).

$$min \sum_{i,q} \sigma_{iq}(1 - cos(\boldsymbol{x}_i, \boldsymbol{p}_q))$$
$$\text{s.t. } \sigma_{iq} = \begin{cases} 1, & \text{if } c(i) = q. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

3.3. Proposed Method

In this paper, we propose W2V-LSA, a new topic-modeling method combining Word2vec and spherical k-means clustering in a harmonized manner. It can significantly increase the quality of topic modeling by overcoming the drawbacks of existing representation-based probabilistic-statistical models, are ill-suited to satisfactorily consider the context of documents. Fig. 3 shows the overall process of W2V-LSA consisting of four steps. Each step will be further explained in detail.

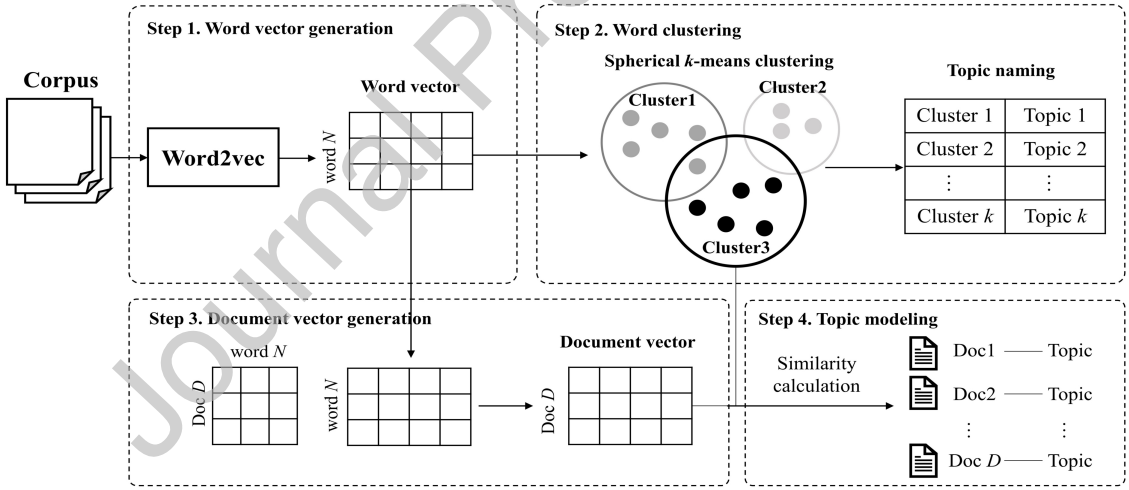


Fig. 3: Overall process of W2V-LSA

Step 1: Each word in a corpus is vectorized as an m-dimensional word vector $\boldsymbol{x}_i \in R^m$ by Word2vec.

Step 2: Word clustering is performed by applying a spherical k-means clustering method to the extracted $\boldsymbol{x}_i$. Each $\boldsymbol{x}_i$ is assigned the closest cluster number by comparing $\boldsymbol{p}_q$ and cosine

similarity of the cluster. In this case, the name of the cluster is defined by considering the characteristics of the words assigned to the cluster, and it is considered as a topic.

Step 3: Each document-specific vector $\boldsymbol{l}_j, j = 1, \ldots, D$, representing the characteristics of the document, is generated by using matrix multiplication between the $\boldsymbol{x}_i$ and $N \times D$ term-document matrix. Fig. 4 is a graphical representation of how to generate $\boldsymbol{l}_j$.

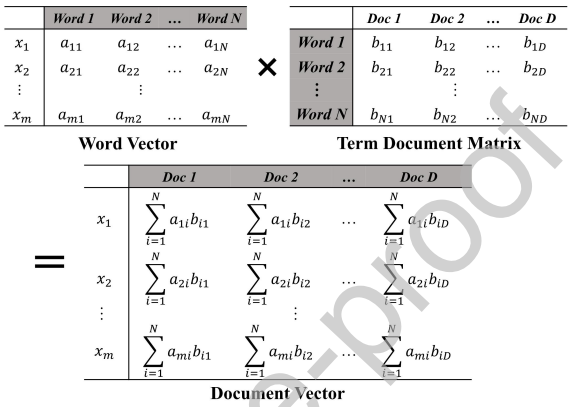


Fig. 4: Example of document vector construction

Step 4: Cosine similarity between $\boldsymbol{x}_i$ in each cluster and $\boldsymbol{l}_j$ is calculated. The final similarity between the cluster and the document is determined by the average value of the cosine similarity with the top t words of each cluster. The topic of the cluster with the highest similarity is assigned to the topic of the document by comparing their final similarity. This process is illustrated in Fig. 5.

4. Results

In this section, we compare W2V-LSA with PLSA, a representative probabilistic topic model. This section consists of three parts. First, we show blockchain trend analysis results from PLSA and W2V-LSA. We then evaluate the performance of W2V-LSA quantitatively and qualitatively in terms of the accuracy of topic allocation and the relevance of the words in each topic, respectively.

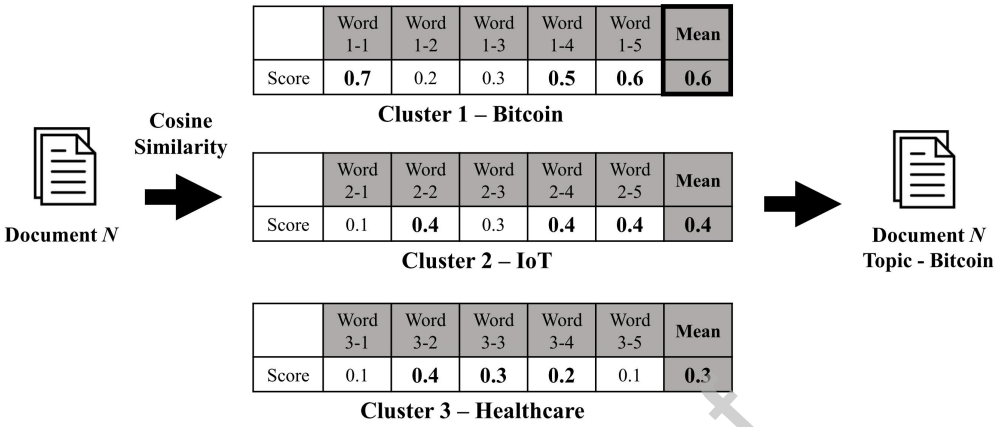


Fig. 5: Topic Modeling of W2V-LSA

4.1. Blockchain trend analysis

4.1.1. PLSA

We implemented PLSA to each term-document matrix based on uni-gram representation for a single country. Bayesian information criterion (BIC) is used to minimize overfitting problem caused by increasing the number of parameters while maximizing log-likelihood function (Schwarz et al., 1978). The number of topics in each country was determined where the BIC value is the smallest; the group comprised of Korea, the US, China, and the etc. group have 7, 9, 6, and 9 topics respectively. The top 3 to 5 words in order of probability are designated as the main topic of the document. The PLSA results are presented in Table 1.

It is noteworthy that in Table 1 the ratios are uniformly distributed. There are some characteristic results in each group. In the case of Korea, we found 'Fintech' and 'Regulation,' which are absent from others. 'Healthcare/Privacy' accounts for a large share in the US and China. Also, the etc. group has a variety of topics compared to others as well as one unique topic, 'Real Estate.'

Table 2 shows the results of using PLSA to identify topics that change over time in each country. In Korea, topics such as 'Security/Network,' 'Finance/Fintech,' and 'Virtual Currency/Bitcoin' were prominent in 2016 and 2017, while topics related to application fields in Blockchain 3.0, including 'Service/Trade', 'IoT', and 'Energy/Transaction,' comprised a big part of topic ratio in 2018. In the US, only 'Distributed Ledger' and 'Bitcoin/Transaction' topics appear in 2015, but other topics such as 'Healthcare/Privacy' and 'IoT/Smart Contract' rose to dominance in 2016. Except for a few specific topics such as 'Energy/Cryptocurrency' and 'Cloud,' various topics are

Table 1: PLSA based topic results for blockchain related papers by country; Ratio (%) indicates the percentage of the topic among the topics of the entire document

KOREA		US	
Topic	Ratio (%)	Topic	Ratio (%)
Finance/Fintech	16.4	Healthcare/Privacy	17.2
Security/Network	16.4	Cloud	15.5
Service/Trade	16.4	Energy/Cryptocurrency	12.1
IoT	14.8	Security	12.1
Electricity/Transaction	13.1	Distributed Ledger	10.3
Virtual Currency/Bitcoin	13.1	IoT/Smart Contract	10.3
Regulation/Cryptocurrency	9.8	Bitcoin/Transaction	8.6
		Finance/Service	8.6
		Network	5.2
CHINA		etc.	
Topic	Ratio (%)	Topic	Ratio (%)
Healthcare/Privacy	25	Bitcoin	13.9
Electricity/Smart Contract	17.5	Market/Cryptocurrency	13.9
Security	15	Smart Contract	13.9
Storage/Cloud	15	Transaction/Network	13.9
Transaction/Bitcoin	15	Distributed Ledger/Service	11.1
Service	12.5	IoT/Security	11.1
		Healthcare/Privacy	9.7
		Finance	6.9
		Real Estate/Energy	5.6

uniformly distributed in 2017 and 2018. In China, 'Storage/Cloud' takes up a big share of 2016, but they are replaced by 'Transaction/Bitcoin' in 2017. From 2016 to 2018, documents related to 'Healthcare/Privacy' have consistently been predominant. In the etc. group, beginning with the document on 'Bitcoin' in 2014, studies on various fields have been carried out by 2018.

4.1.2. W2V-LSA

To create unique word vectors for each country, we applied Word2vec to documents for each country. We used the Skip-gram method and set m and δ to 100 and 12 respectively. When implementing spherical k-means clustering on $\boldsymbol{x}_i \in R^{100}$, k, the optimal number of clusters, was decided by a silhouette measure that can estimate k considering the distance and density of the clusters (Rousseeuw, 1987); the values of k for Korea, the US, China, and the etc. group were determined to be 6, 6, 7, and 7 respectively. In this study, we defined t as 3 in Step 4 to calculate the final similarity between the clusters and the documents. Table 3 shows the results of topic modeling using W2V-LSA.

The top-ranked topics in Table 3 are distributed differently by country, denoting their characteristics. In Korea, there is a preponderance of papers related to 'Virtual Currency,' 'Regulation,' 'Economy' and 'Fintech,' which represents Korea's interest in the financial sector. In other countries, there are various topics including 'Healthcare' and 'Cloud' not seen in Korea. Especially

Table 2: PLSA based topic results for blockchain related papers over time by country

KOREA						US					
Topic	Ratio by Year (%)					Topic	Ratio by Year (%)				
	2014	2015	2016	2017	2018		2014	2015	2016	2017	2018
Finance/Fintech	-	-	17	19	14	Healthcare/Privacy	-	0	29	11	21
Security/Network			33	15	14	Cloud		0	14	11	21
Service/Trade			0	12	24	Energy/Cryptocurrency		0	0	21	10
IoT			17	15	14	Security		0	14	16	10
Energy/Transaction			17	12	14	Distributed Ledger		33	0	11	10
Virtual Currency/Bitcoin			17	19	7	IoT/Smart Contract		0	29	11	7
Regulation/Cryptocurrency			0	8	14	Bitcoin/Transaction		67	0	0	10
						Finance/Service		0	0	11	10
						Network		0	14	11	0
CHINA						etc.					
Topic	Ratio by Year (%)					Topic	Ratio by Year (%)				
	2014	2015	2016	2017	2018		2014	2015	2016	2017	2018
Healthcare/Privacy	-	-	33	20	26	Bitcoin	100	0	20	14	10
Electricity/Smart Contract			0	10	22	Market/Cryptocurrency	0	0	10	11	21
Security			0	20	15	Smart Contract	0	25	10	14	17
Storage/Cloud			67	10	11	Transaction/Network	0	25	10	14	14
Transaction/Bitcoin			0	40	7	Distributed Ledger/Service	0	50	10	7	10
Service			0	0	19	IoT/Security	0	0	20	11	10
						Healthcare/Privacy	0	0	10	14	7
						Finance	0	0	10	11	3
						Real Estate/Energy	0	25	0	4	7

noteworthy is unique topics such as ‘Real Estate’ in the etc. group.

Table 3: W2V-LSA based topic results for blockchain related papers by country

KOREA		US	
Topic	Ratio (%)	Topic	Ratio (%)
IoT/Network/Smart Contract	29.5	Energy/Healthcare	27.6
Virtual Currency/Tax/Regulation/Real Estate	23	IoT/Economy/Privacy	27.6
Industry 4.0/Economy	19.7	Distributed Ledger/Network	19
Bitcoin/Cryptocurrency/Healthcare/Law	13.1	Bitcoin/Cryptocurrency/Transaction	17.2
Finance/Fintech/Bank	9.8	Smart Contract	5.2
Energy/Transaction	4.9	Finance	3.4
CHINA		etc.	
Topic	Ratio (%)	Topic	Ratio (%)
Smart Contract/Energy/Trade	30	Healthcare/Privacy/Network	30.6
Healthcare	25	Finance/Market	13.9
Cloud/Service	22.5	Bitcoin/Cryptocurrency/Security	12.5
Security/Signature	12.5	Real Estate/Service/Trade	12.5
Bitcoin/Transaction	5	Distributed Ledger/IoT	11.1
Network	2.5	Smart Contract/Energy	9.7

Table 4 shows the results of W2V-LSA. In Korea, from 2016 to 2018, ‘IoT/Network/Smart Contract’ proved to be of continual interest, as were topics regarding the background of blockchain and finance fields such as ‘Industry 4.0/Economy,’ ‘Virtual Currency/Regulation’ and ‘Finance.’ In the US, ‘Bitcoin/Cryptocurrency/Transaction’ was prevalent for much of 2015 but interest in the topic began to wane after 2016. ‘IoT/Economy/Privacy’ was especially popular, accounting

for about 43% of topics in 2016, while 'Energy/Healthcare' has consistently occupied a large portion of 2016. In China, unlike Korea, topics such as 'Smart Contract/Energy/Trade,' 'Healthcare,' and 'Security/Signature' began to trend after 2016. In the etc. group, topics related to 'Bitcoin/Cryptocurrency,' 'Distributed Ledger' and 'Transaction' are dominant during the first two years of the entire period, but topics such as 'Healthcare/Privacy/Network' and 'Real Estate/Service/Trade' appear only after 2016.

Table 4: W2V-LSA based topic results for blockchain related papers over time by country

KOREA						US					
Topic	Ratio by Year (%)					Topic	Ratio by Year (%)				
	2014	2015	2016	2017	2018		2014	2015	2016	2017	2018
IoT/Network/Smart Contract	-	-	33	23	34	Energy/Healthcare	-	0	28.6	31.6	27.6
Virtual Currency/Tax/Regulation/Real Estate			17	27	21	IoT/Economy/Privacy		0	42.9	21.1	31
Industry 4.0/Economy			33	19	17	Distributed Ledger/Network		0	0	26.3	20.7
Bitcoin/Cryptocurrency/Healthcare/Law			0	8	21	Bitcoin/Cryptocurrency/Transaction		66.7	14.3	15.8	13.8
Finance/Fintech/Bank			17	15	3	Smart Contract		33.3	14.3	5.3	0
Energy/Transaction			0	8	3	Finance		0	0	0	6.9
CHINA						etc.					
Topic	Ratio by Year (%)					Topic	Ratio by Year (%)				
	2014	2015	2016	2017	2018		2014	2015	2016	2017	2018
Smart Contract/Energy/Trade	-	-	33	30	29.6	Healthcare/Privacy/Network	0	0	30	28.6	37.9
Healthcare			33	20	25.9	Finance/Market	0	0	10	14.3	17.2
Cloud/Service			0	10	29.6	Bitcoin/Cryptocurrency/Security	100	25	30	14.3	0
Security/Signature			33	20	7.4	Real Estate/Service/Trade	0	0	0	10.7	20.7
Bitcoin/Transaction			0	10	3.7	Distributed Ledger/IoT	0	25	30	7.1	6.9
Network			0	10	0	Smart Contract/Energy	0	0	0	10.7	13.8
Privacy			0	0	3.7	Transaction	0	50	0	14.3	3.4

4.2. Quantitative evaluation

4.2.1. Topic coherence evaluation

Perplexity is referred to as a key evaluation measure in probabilistic topic modeling. However, perplexity is unable to explain the semantic coherence of words for each topic on non-probabilistic models (Chang et al., 2009). Alternatively, topic coherence can measure the quality of a topic with reference to how many the words of a topic coincide within the same documents or the semantic similarity among the words in the topic (Aletras & Stevenson, 2013; Li et al., 2016; Mimno et al., 2011). The higher topic coherence score, the more the words for each topic cohere. In order to evaluate our topic model, we calculate two coherence measures: (1) UMass (Mimno et al., 2011) and (2) normalized pointwise mutual information (NPMI) (Lau et al., 2014). We compute the coherence as we increase T, the number of words for each topic. Top-T words have a large weight, which means the highest probability of the words in PLSA and the highest cosine similarity of the words in W2V-LSA respectively.

Fig. 6 shows UMass and NPMI based average topic coherences for each model, PLSA and W2V-LSA, and T was varied from 3 to 14. As T increases, coherence scores decrease in both W2V-LSA and PLSA. For all conditions based on T values, W2V-LSA model outperforms PLSA. To be specific, the NPMI score gap between W2V-LSA and PLSA is the largest at $T = 3$ and the smallest at $T = 14$.

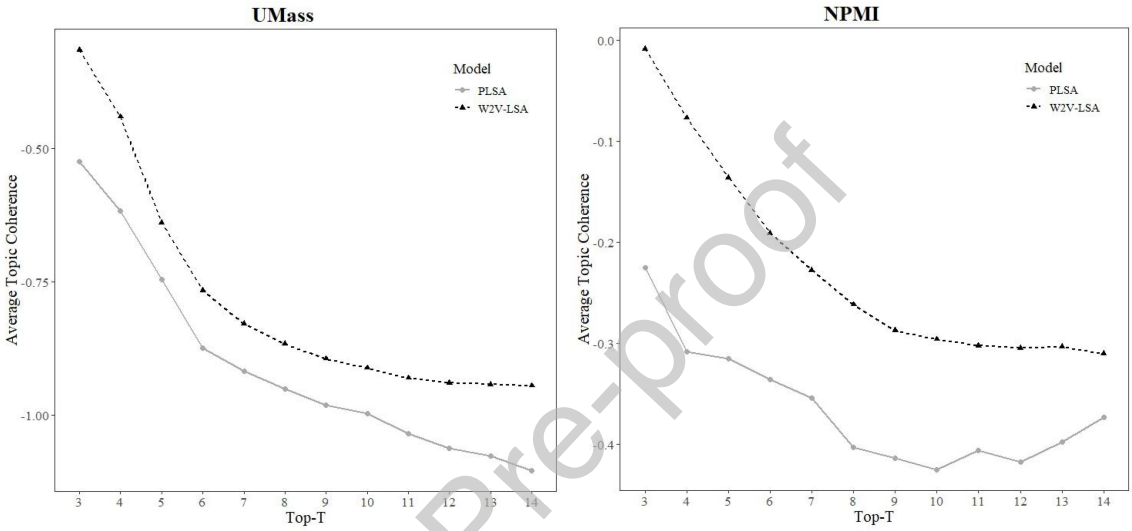


Fig. 6: UMass and NPMI scores for each model

4.2.2. Keyword matching evaluation

For measuring the accuracy of allocated topics to the documents, existing studies have used the data for text classification, which was already categorized or assigned to the topic. Since there are no exact labels for our data, we propose a quantitative evaluation method: keyword matching score (KMS). Unlike ambiguous results of existing topic modeling, this approach has the advantage of numerically measuring the accuracy. For computing KMS, we gathered keywords from each document and we counted how many top-T words of the topic exactly match the keywords.

KMS is:

$$KMS = \sum_{t=1}^{T} u_t \tag{2}$$

where u_t is the sum of the number of words that exactly match the keywords.

The weighted KMS is:

$$weighted\ KMS\ = \frac{\sum_{t=1}^{T} w_t u_t}{\sum_{t=1}^{T} t} \tag{3}$$

where $w_1, w_2, \ldots, w_T$ are the weights assigned to the top-T words for each topic in case of the top-T weighted KMS.

KMS was computed for each model, PLSA and W2V-LSA, for several T (Fig. 7). The KMS of W2V-LSA before top-4 and after top-12 is larger in W2V-LSA than PLSA. In the case of the top-5 weighted KMS, the score in the W2V-LSA model is significantly larger than that of the PLSA in all top-T words if only weighted scores were given to the top-5 words.

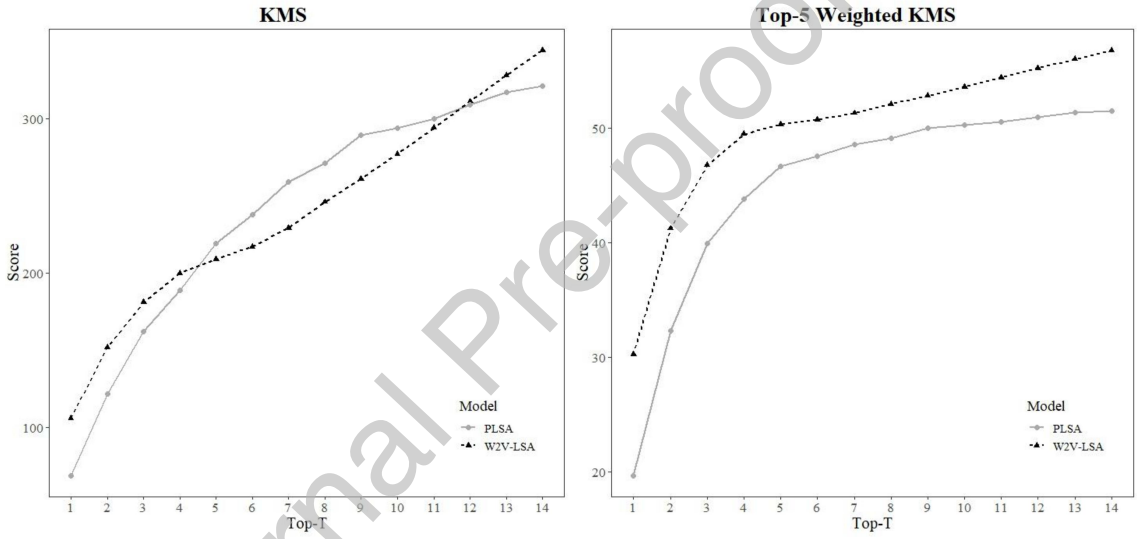


Fig. 7: KMS and top-5 weighted KMS for each model

4.3. Qualitative evaluation

Results show that W2V-LSA is able to extract more detailed topics than PLSA for documents. This also means that W2V-LSA has the advantage of assigning more suitable topics to each document than PLSA. For example, the paper in Fig. 8 is related to the blockchain in the healthcare industry. PLSA and W2V-LSA assigned this paper to the topic of 'Service/Trade' and 'Healthcare' respectively. It is because words such as "healthcare" or "medical" barely appear in the entire corpus of Korea compared with the word "service", and PLSA as a word frequency-based topic-modeling technique suffers from capturing precise information. Fig. 9 is one of the documents in

Document Number
60th Document in Korea

Topic
PLSA – Service/Trade,
W2V-LSA – Bitcoin/Cryptocurrency/Healthcare/Law

Title
A Study to Accept Block Chain System on Medical Industry

Abstract
The purpose of this study was to investigate characteristics of Blockchain to introduce Blockchain technology to the medical industry. Thus, the 5 factors such as Security, Availability, Reliability, Diversity and Economic feasibility were used as the characteristics of Blockchain, which are independent variable, based on precedent studies. Also, Technology Acceptance Model(TAM) that are widely used in the research on Acceptance Intention in order to introduce new technology accept was utilized for intervening variable and dependent variable. For the purpose of this study, the health professionals and ordinary people were classified into health care providers and medical service consumers respectively to conduct survey and comparative analysis. The researching findings discovered that Hypothesis1-1(Security-Perceived Easiness- Acceptance Intention), Hypothesis2-1(Security-Perceived Usefulness-Acceptance Intention) and Hypothesis 3-1(Security-Perceived Easiness-Perceived Usefulness-Acceptance Intention) were rejected in all participants including health professionals and ordinary people. Additionally, all characteristics of Blockchain such as Availability, Reliability, Diversity and Economic feasibility were rejected in Hypothesis3-2 and Hypothesis 3-5 that pass through Easiness and Usefulness to Acceptance Intention, in case of the medical service consumers. Thus, this study discovered and confirmed that there's difference between the health care providers and medical service in the characteristics of Blockchain, as they pass through Perceived Easiness and Perceived Usefulness to Acceptance Intention. The significance of this study can be found, as it suggested activation plan through empirical analysis, which was hardly ever used in the study on Blockchain activation.

Fig. 8: Example for comparison of PLSA (marked in light gray) and W2V-LSA (marked in dark gray)

the US and its main content is about the application of the decentralized network system based on cryptocurrency for the security of banking monetary technocracy. For the same reason as in the previous example, PLSA pointed out 'Network' as a topic for the document shown in Fig. 9, while W2V-LSA identified 'Bitcoin/Cryptocurrency'.

5. Discussion

There are several methodological implications of this study. First of all, the use of the context-embedding representation is the considerable advantage of W2V-LSA compared with other topic models based on uni-gram based representation. To best of our knowledge, most of the studies on natural language models have demonstrated that the contextual embedding method outperformed the classical n-gram based models via empirical experiments (Mikolov et al., 2013a,b; Schnabel et al., 2015; Sharma et al., 2017). W2V-LSA resolves the issues of sparseness and high dimensionality of the n-gram representation, and unlike probability-based statistical topic models it does not require any distributional assumptions which often degrade algorithm performance. Secondly, we

Document Number
8th Document in US

Topic
PLSA – Network ,
W2V-LSA – Bitcoin/Cryptocurrency/Transaction
Title
Decentralized banking Monetary technocracy in the digital age

Abstract
Bitcoin has ushered in the age of blockchain based digital currency systems. Secured by cryptography and computing power, and distributed across a decentralized network of anonymous nodes, these novel systems could potentially disrupt the way that monetary policy is administered—moving away from today's human-fallible central bankers and towards a technocratic, rules-based algorithmic approach. It can be argued that modern central banks have failed to stem macroeconomic crises, and may have, in fact, exacerbated negative outcomes by incen-tivizing excessive risk-taking and moral hazard via unconventional monetary tools such as quantitative easing and negative interest rates. A central bank typically serves three primary functions to issue and regulate the supply of money to serve as clearinghouse for settlement of payments transactions and to serve as lender of last resort. Could a digital currency system serve as a rational substitute for a central bank. This perspective paper examines that question, and then suggests that indeed it could be plausible. While Bitcoin in its current form will prove to be inadequate to function as monetary authority, I put forward what an operative case could resemble.

Fig. 9: Example for comparison of PLSA (marked in light gray) and W2V-LSA (marked in dark gray)

demonstrated the feasibility and usefulness of W2V-LSA by comparing with PLSA in quantitative and qualitative ways and confirmed that W2V-LSA has a relative advantage over PLSA in finding precise topics for documents. W2V-LSA can extract topics not found in PLSA and topics derived from W2V-LSA are usually more detailed and definite than ones of PLSA; The words assigned to each topic are more relevant and meaningful than those of PLSA. PLSA tends to place topics with high-frequency words on top of other topics, while W2V-LSA captures diverse and distinct topics appropriately and it also forces words to belong to one cluster exclusively. This may be because W2V-LSA can learn the word periphery using the cosine similarity, making it feasible to derive the main words in one document even though their frequency is low in terms of the whole corpus. Finally, W2V-LSA can be used universally for any other studies using topic modeling as well as technology trend analysis, which creates the added value to the field of semantic expert systems.

Furthermore, this study has several managerial implications. First, it provides a data-driven text mining approach called W2V-LSA that allows to effectively and efficiently discover trends in blockchain technology without anyone investigating full texts of every document. As a content-based analysis technique, W2V-LSA can extract new information not found in the existing blockchain trend analysis at both national and global levels. Second, we can provide valuable insights to the blockchain-related academia and industry, and present the future implementing fields

of blockchain technology. In the early blockchain research, a virtual currency like Bitcoin attracted much attention as a promising field of future research. This is particularly so in Korea, where virtual currency and its regulations have been discussed nationwide in 2017. Recently, global research on blockchain technology has been oriented toward other applications beyond virtual currencies, such as healthcare, smart contract, energy, cloud, and IoT. These emerging fields of study promise to have a significant impact in the near future. Besides, security still remains an important research topic because of attack attempts for blockchain technology itself. Therefore, it is also valued that the research on security of a decentralized network, which is one of the advantages of blockchain. Lastly, this study provides direction to enable industrial sectors such as new technology-based firms (NTBFs, i.e., technology-based startups), willing to leverage blockchain, to preoccupy a potential application domain based on blockchain. From the perspective of investment, it can bring real business value to NTBFs and promote crowdfunding and investment of venture capital firms (Fiedler & Sandner, 2017).

6. Conclusions

This paper proposed a novel technique for topic modeling called W2V-LSA based on Word2vec and Spherical k-means clustering. We collected blockchain-related 231 documents and applied our method to analyze blockchain trends by country and time. We then presented current trends in blockchain technology and demonstrated the usefulness of the new method by comparing it with PLSA from quantitative and qualitative perspectives. The significance of this study lies in developing a new topic-modeling method as well as providing an indicator to present the future direction of blockchain study.

Although this study has a lot of contributions to technology trend analysis, but at the same time there are several limitations as well. We conducted the experiments in a limited scale to serve as a proof of concept; we compared our proposed method only with PLSA under a small number of documents. We plan to expand the scope of our analysis to a larger-scale analysis of other advanced technologies along with a comparison to several other comparative methods. In addition, it should be noted that the optimal values of the user-defined parameters are data-dependent, which makes it hard to select those a priori. There is definitely a need to study this problem using a principled approach.

Possibility for several future research directions is worth investigating. This study can be applied to the trend analysis for any other domains not only for blockchain. In addition, it is expected to be widely used in several topics of research in which text data from different sources are collected such as patent analysis (Lee et al., 2009; Xie & Miyazaki, 2013; Noh et al., 2015), customer online review analysis (Jung & Suh, 2019; Korfiatis et al., 2019), and text based-recommendation system (dos Santos et al., 2018), which are recently attracting attention. Further, it would be interesting to investigate the performance of our proposed method when Word2vec in W2V-LSA is replaced by state-of-the-art word embedding methods (Pennington et al., 2014; Devlin et al., 2018).

Appendix A. Detailed information for the etc. group

Table A.1 shows the growth of the number of blockchain-related papers published by 21 countries in the etc. group.

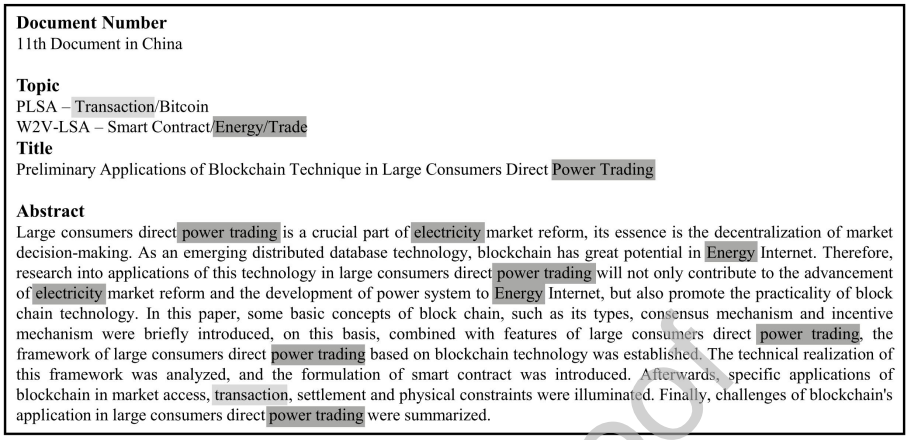
Table A.1: Growth of the number of blockchain-related papers by country in the etc. group

	2014	2015	2016	2017	2018Q2	Total
etc.	1	4	10	28	29	72
UK	1	4	1	5	8	19
Australia			3	3	2	8
Russia				3	4	7
Germany			1	1	4	6
Italy			2	1	3	6
India				1	3	4
Brazil				2	1	3
Slovenia				1	2	3
France			1	1		2
Switzerland			1	1		2
UAE					2	2
Canada				1		1
Denmark				1		1
Greece				1		1
Hongkong				1		1
Japan				1		1
Mexico				1		1
Malaysia				1		1
Taiwan				1		1
Ghana				1		1
Netherlands					1	1

Appendix B. Examples for qualitative evaluation

We showed only two examples for the comparison of PLSA and W2V-LSA (Fig. 8 and Fig. 9). The figures in Fig. B.1 and Fig. B.2 are other examples for qualitative evaluation of W2V-LSA.

The results can be obtained and explained in the same way as Section 4.3.

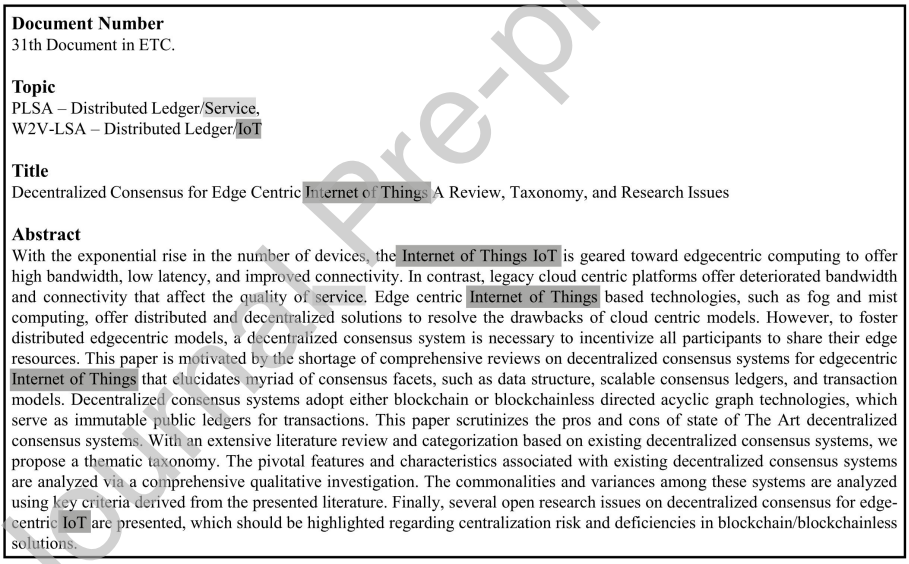


Document Number
11th Document in China

Topic
PLSA – Transaction/Bitcoin
W2V-LSA – Smart Contract/Energy/Trade
Title
Preliminary Applications of Blockchain Technique in Large Consumers Direct Power Trading

Abstract
Large consumers direct power trading is a crucial part of electricity market reform, its essence is the decentralization of market decision-making. As an emerging distributed database technology, blockchain has great potential in Energy Internet. Therefore, research into applications of this technology in large consumers direct power trading will not only contribute to the advancement of electricity market reform and the development of power system to Energy Internet, but also promote the practicality of block chain technology. In this paper, some basic concepts of block chain, such as its types, consensus mechanism and incentive mechanism were briefly introduced, on this basis, combined with features of large consumers direct power trading, the framework of large consumers direct power trading based on blockchain technology was established. The technical realization of this framework was analyzed, and the formulation of smart contract was introduced. Afterwards, specific applications of blockchain in market access, transaction, settlement and physical constraints were illuminated. Finally, challenges of blockchain's application in large consumers direct power trading were summarized.

Fig. B.1: Example for comparison of PLSA (marked in light gray) and W2V-LSA (marked in dark gray)



Document Number
31th Document in ETC.

Topic
PLSA – Distributed Ledger/Service,
W2V-LSA – Distributed Ledger/IoT

Title
Decentralized Consensus for Edge Centric Internet of Things A Review, Taxonomy, and Research Issues

Abstract
With the exponential rise in the number of devices, the Internet of Things IoT is geared toward edgecentric computing to offer high bandwidth, low latency, and improved connectivity. In contrast, legacy cloud centric platforms offer deteriorated bandwidth and connectivity that affect the quality of service. Edge centric Internet of Things based technologies, such as fog and mist computing, offer distributed and decentralized solutions to resolve the drawbacks of cloud centric models. However, to foster distributed edgecentric models, a decentralized consensus system is necessary to incentivize all participants to share their edge resources. This paper is motivated by the shortage of comprehensive reviews on decentralized consensus systems for edgecentric Internet of Things that elucidates myriad of consensus facets, such as data structure, scalable consensus ledgers, and transaction models. Decentralized consensus systems adopt either blockchain or blockchainless directed acyclic graph technologies, which serve as immutable public ledgers for transactions. This paper scrutinizes the pros and cons of state of The Art decentralized consensus systems. With an extensive literature review and categorization based on existing decentralized consensus systems, we propose a thematic taxonomy. The pivotal features and characteristics associated with existing decentralized consensus systems are analyzed via a comprehensive qualitative investigation. The commonalities and variances among these systems are analyzed using key criteria derived from the presented literature. Finally, several open research issues on decentralized consensus for edge-centric IoT are presented, which should be highlighted regarding centralization risk and deficiencies in blockchain/blockchainless solutions.

Fig. B.2: Example for comparison of PLSA (marked in light gray) and W2V-LSA (marked in dark gray)

References

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers* (pp. 13–22).

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *6*.

Alharby, M., & van Moorsel, A. (2017). A systematic mapping study on current research topics in smart contracts. *International Journal of Computer Science & Information Technology*, *9*, 151–164.

Alonso, S. G., Arambarri, J., López-Coronado, M., & de la Torre Díez, I. (2019). Proposing new blockchain challenges in ehealth. *Journal of medical systems*, *43*, 64.

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. (2018). Trends on health in social media: Analysis using twitter topic modeling. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 558–563). IEEE.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*, 1137–1155.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Buchta, C., Kober, M., Feinerer, I., & Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, *50*, 1–22.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Dabbagh, M., Sookhak, M., & Safa, N. S. (2019). The evolution of blockchain: A bibliometric study. *IEEE Access*, *7*, 19212–19221.

Dagher, G. G., Mohler, J., Milojkovic, M., & Marella, P. B. (2018). Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. *Sustainable Cities and Society*, *39*, 283–297.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, .

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, *42*, 143–175.

Fan, K., Wang, S., Ren, Y., Li, H., & Yang, Y. (2018). Medblock: Efficient and secure medical data sharing via blockchain. *Journal of medical systems*, *42*, 136.

Fiedler, M., & Sandner, P. (2017). Identifying leading blockchain startups on a worldwide level. *Frankfurt School Blockchain Center*, .

Giungato, P., Rana, R., Tarabella, A., & Tricase, C. (2017). Current trends in sustainability of bitcoins and related blockchain technology. *Sustainability*, *9*, 2214.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296). Morgan Kaufmann Publishers Inc.

Hung, J.-l. (2012). Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. *British Journal of Educational Technology*, *43*, 5–16.

Hung, J.-L., & Zhang, K. (2012). Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques. *Journal of Computing in Higher education*, *24*, 1–17.

Jung, Y., & Suh, Y. (2019). Mining the voice of employees: A text mining approach to identifying and analyzing job

satisfaction factors from online employee reviews. *Decision Support Systems*, *123*, 113074.

Kang, H. J., Kim, C., & Kang, K. (2019). Analysis of the trends in biochemical research using latent dirichlet allocation (lda). *Processes*, *7*, 379.

Kim, H.-j., Jo, N.-o., & Shin, K.-s. (2015). Text mining-based emerging trend analysis for the aviation industry. *Journal of intelligence and information systems*, *21*, 65–82.

Kim, Y.-M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. *Health informatics journal*, *24*, 432–452.

Kivikunnas, S. (1998). Overview of process trend analysis methods and applications. In *ERUDIT Workshop on Applications in Pulp and Paper Industry* (pp. 395–408). Citeseer.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers online reviews. *Expert Systems with Applications*, *116*, 472–486.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539).

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*, 481–497.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 165–174). ACM.

Lu, Y. (2019). The blockchain: State-of-the-art and research challenges. *Journal of Industrial Information Integration*, .

Lu, Y., & Zhai, C. (2008). Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web* (pp. 121–130).

McGhin, T., Choo, K.-K. R., Liu, C. Z., & He, D. (2019). Blockchain in healthcare applications: Research challenges and opportunities. *Journal of Network and Computer Applications*, .

Miau, S., & Yang, J.-M. (2018). Bibliometrics-based evaluation of the blockchain research trend: 2008–march 2017. *Technology Analysis & Strategic Management*, *30*, 1029–1045.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, .

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics.

Miraz, M. H., & Ali, M. (2018). Blockchain enabled enhanced iot ecosystem security. In *International Conference for Emerging Technologies in Computing* (pp. 38–46). Springer.

Nakamoto, S. et al. (2008). Bitcoin: A peer-to-peer electronic cash system, .

Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, *57*, 753–767.

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, *42*, 4348–4360.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, G., Panayi, E., & Chapelle, A. (2015). Trends in cryptocurrencies and blockchain technologies: a monetary theory and regulation perspective. *Journal of Financial Perspectives*, *3*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

dos Santos, F. F., Domingues, M. A., Sundermann, C. V., de Carvalho, V. O., Moura, M. F., & Rezende, S. O. (2018). Latent association rule cluster based model to extract topics for classification and recommendation applications. *Expert Systems with Applications*, *112*, 34–60.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 298–307).

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*, 461–464.

Sharma, I., Anand, S., Goyal, R., & Misra, S. (2017). Representing contexual relations with sanskrit word embeddings. In *International Conference on Computational Science and Its Applications* (pp. 262–273). Springer.

Swan, M. (2015). *Blockchain: Blueprint for a new economy*. " O'Reilly Media, Inc.".

Terachi, M., Saga, R., & Tsuji, H. (2006). Trends recognition in journal papers by text mining. In *2006 IEEE International Conference on Systems, Man and Cybernetics* (pp. 4784–4789). IEEE volume 6.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*, 1216–1247.

Van Hooland, S., Coeckelbergs, M., Hengchen, S., & Rizza, E. (2017). Scrambling for metadata: Using topic modeling and word2vec to explore the archives of the european commission. In *Digital approaches towards serial publications (18th–20th centuries)*.

Xie, Z., & Miyazaki, K. (2013). Evaluating the effectiveness of keyword search strategy for patent identification. *World Patent Information*, *35*, 20–30.

Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology?a systematic review. *PloS one*, *11*, e0163477.

Zeng, S., & Ni, X. (2018). A bibliometric analysis of blockchain research. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 102–107). IEEE.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and svmperf. *Expert Systems with Applications*, *42*, 1857–1863.

Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. In *2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 557–564).

IEEE.

AUTHOR DECLARATION

We wish to confirm that there are no conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from junghyelee@unist.ac.kr

Signed by all authors as follows:

Suhyeon Kim. Date: November 24, 2019

Haecheong Park. Date: November 24, 2019

Junghye Lee. Date: November 24, 2019