kim_2020_word2vec_based_latent_semantic_analysis_w2v_lsa_for_topic_modeling_a_study_on_blockchain_technology_trend_analysis

Year

2020

Author(s)

Suhyeon Kim and Haecheong Park and Junghye Lee

Title

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

Venue

Expert Systems with Applications

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

Label generation

“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.”

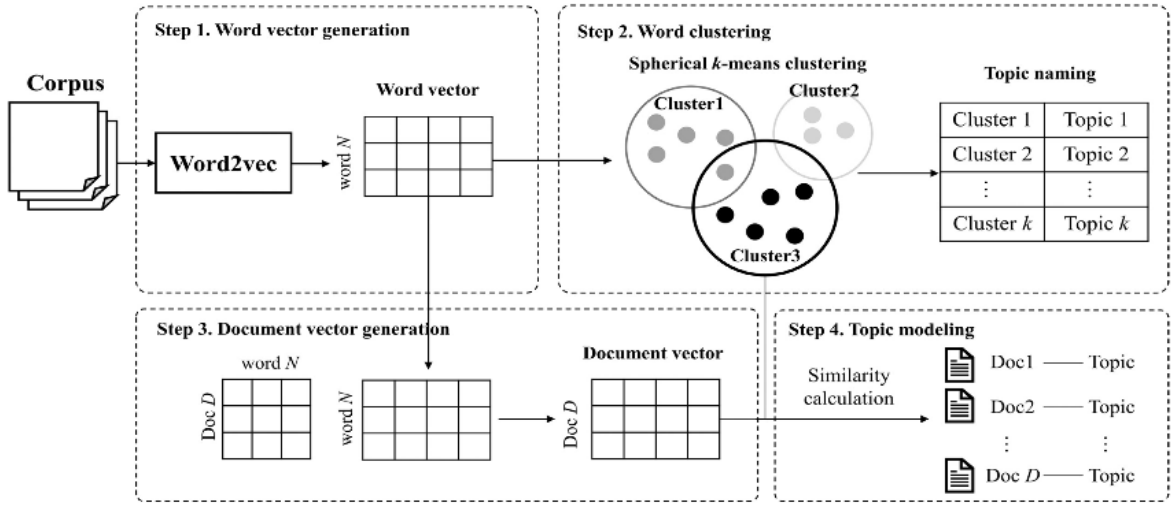

Fig. 3. Overall process of W2V-LSA.

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
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

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

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

Motivation

/

Topic modeling

Word2vec and Spherical k-means clustering based technique (Word2vec-based Latent Semantic Analysis (W2V-LSA))

Baseline: PLSA

Topic modeling parameters

W2V-LSA

Skip-gram method for W2V:

- m: 100
- δ: 12

Spherical k-means clustering:

- optimal number of clusters (i.e. topics): {6, 6, 7, 7} (nr of clusters per country of publication)

Words considered for similarity between cluster and document: 3
(determined by the average value of the cosine similarity with the top t words of each cluster)

PLSA

Nr of topics: {7, 9, 6, 9} (nr of topics per country of publication)

Nr. of topics

26

Label

Manually assigned single or multi-word labels

Label selection

/

Label quality evaluation

/

Assessors

/

Domain

Paper: Topic modeling
Dataset: Blockchain

Problem statement

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.

Corpus

Origin: Scopus, ScienceDirect, Web of Science, IEEE Xplore, Google Scholar, and Korean Citation Index.
Nr. of documents: 763 (231 after processing)
Details:

- keywords and abstracts contain the words such as‘Blockchain,’ ‘Block chain,’ and ‘Block-chain’ from 2014 to August 2018

Document

Abstract of blockchain-related paper

Pre-processing

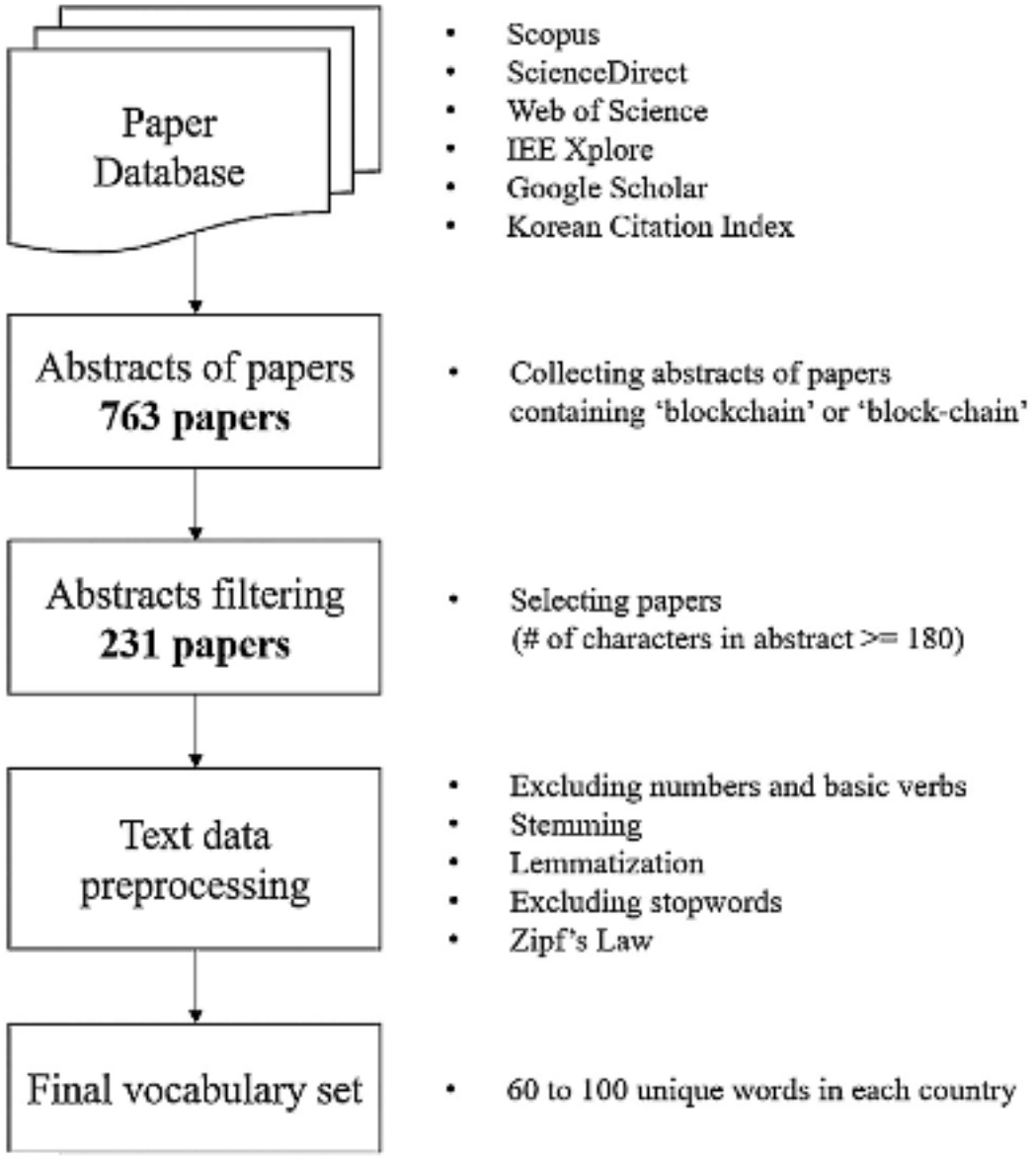

Fig. 1. Process of data collection and preprocessing.

```
@article{kim_2020_word2vec_based_latent_semantic_analysis_w2v_lsa_for_topic_modeling_a_study_on_blockchain_technology_trend_analysis,
  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.},
author = {Suhyeon Kim and Haecheong Park and Junghye Lee},
date-added = {2023-03-23 19:14:07 +0100},
date-modified = {2023-03-23 19:14:07 +0100},
doi = {https://doi.org/10.1016/j.eswa.2020.113401},
issn = {0957-4174},
journal = {Expert Systems with Applications},
keywords = {Trend analysis, Topic modeling, Word2vec, Probabilistic latent semantic analysis, Blockchain},
pages = {113401},
title = {Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis},
url = {https://www.sciencedirect.com/science/article/pii/S0957417420302256},
volume = {152},
year = {2020}}

#Thesis/Papers/Initial