International Journal of Advanced Science and Technology
Vol. 29, No. 3, (2020), pp. 14089 - 14096

Toward Representing Automatic Knowledge Discovery from Social Media Contents Based on Document Classification

Zeinab Shahbazi, Yung-Cheol Byun*, Dong Cheol Lee*

Jeju National University
Zeinab.sh@jejunu.ac.kr, ycb@jejunu.ac.kr, dchlee@jejunu.ac.kr

Abstract

Representing documents is one of the critical steps in natural language processing and text mining that focus on converting unstructured to structured documents with numeric vectors to get access to machine learning and data mining algorithms. Bag of word (BOW) model is an adopted text representation system in document classification. Based on BOW, document demonstrate as fixed-length. This process means word dimensions presented as a numerical value that is defined as TF-IDF or word frequency. In this paper, we analyze, the combination of Bag-of-Concept (BOC) and BOW demonstration applying attention mechanism to operate information of word-level and concept-level to achieve the optimal performance of document classification.

Keywords: *Bag of word, Latent Dirichlet Allocation, Latent Semantic Analysis, Doc2vec*

1. Introduction

In the information exploration area, there are various text type of data, e.g. news articles, customer's comments and feedback on social media, reports and logs on medical industries and etc. [1]. Regardless of individual industry or companies, using Natural Language Processing (NLP) on automatic text analysis requirements are continuously show the growth of decisions and also show the text data without cover or generally speaking uncovering data. Text analysis has a various task that among them, the most classic one is document classification which assigns documents in various categories [2].

A generic document classification process divided into three primary steps: document pre-processing, Document classification and document representation. The important part of system performance is related to document representation that transforms the raw data to numeric vectors [3]. In this case, the data is ready to run in machine learning and data mining algorithms. Documents representation quality estimated with sentiment and statistical quality. Feature vectors are evaluated based on semantic quality. i.e. how the features can explain the document contents [4].

Bag-of-Word (BOW) model is a simple and common document classification task which has a moderate to high accuracy, which is acceptable in process and also good interpretability. In another point of view, BOW has some issues from short incomes that prevents performance of BOW-based models [5]. To start the processing system demands the level of the word and decline the conceptual data and semantic relationship between terms and phrases. e.g. "Pyria is going for driving license test" and "Prince is buying a car". Comparing this sentence shows there are not any standard terms between sentences and BOW present this as entirely different vectors [6].

Based on the detected knowledge, document conceptualization proposed to determine the highest level of semantic texts and create Bag-of-Concepts (BOC) model by applying a novel implication score-inverse soft document frequency weighting scheme [7].

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

Comparing BOW and BOC effects of dimensionality and sparsity, which has lower result due to document probabilistic mapping space. In other words, BOC obtains semantic information and concept of knowledge in a document which is one of the essential parts in document classification tasks [8].

In this paper, we organize the process based on the following sections. Section 2 represents the literature review of the related document classification. Section 3 presents the proposed knowledge extraction approach. Section 4 presents the development environment of the proposed approach and finally conclude this paper in section conclusion.

2. Literature Review

This section is presenting the previous research and approaches in probabilistic knowledge extraction.

2.1 Data Representation

Most of the BOW based models contain some type of constraint which defines as sparsity, dimensionality, and lack of capability to obtain the conceptual text details [9]. Recently, there are many published articles related to improving the representation of BOW model. Latent semantic analysis (LSA) and topic modelling are famous in this area. Probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Analysis (LDA) are including in topic modelling subject too. Term-document statistical analysis matrix, detect the structure of the latent semantic system in the document and transfigure it to the low dimensional and dense vector of BOW model [10]. Doc2vec model first proposed by Le and Mikolov to exploit textual information in documents and words to get more information from contents in ongoing vector space [11,12,13]. Recently, with the development of deep learning, lots of deep compositional models applied in a document classification system. Deep compositional model posses with multilayer neural network in different forms e.g. convolutional, recursive, recurrent neural network similar to attention-based transforms for knowing text representation by accomplishing word embedding [14,15].

2.2 Knowledge-based Representation

Word-Net is one of the popular knowledge-based system used in the recent research area. To improve the representation of the document, many examples utilized in various researches [16]. Extracted knowledge from word-net used to accomplish word sense disambiguation (WSD) by using the LSA modelling technique and PageRank system to execute the reviews sentiment classification. To extract the casual relationship in document combination of word-net and Frame-Net applied in CNN network. Regardless of efficiency in lexical knowledge for text representation, dictionaries manually selected to cover the word issues and perceptual knowledge [17]. There are a few studies related to conceptual knowledge for discovering short documents semantics. Similarly, learned representations sometimes are not authentic for a particular domain. Many researches presented to combine the external and existing knowledge based on the machine learning prior knowledge to decrease the dependence of training data in term of providing useful information for document representation [18,19,20].

3. Proposed Approach

This section is to present the overall architecture of the proposed knowledge extraction from probabilistic documents. This model includes three layers: knowledge extraction

Copyright ⓒ 2020 SERSC

layer, pre-processing layer, and output layer.
Figure 1 shows the total overview of the proposed system. The first layer contains the information related to extracting dataset from web servers and social media contents. To prepare the data, necessary and useful information need to extract from collected data. The second layer is the pre-processing data layer. To pre-process the data and make it suitable for further steps in the proposed system, we need to follow word filtering, document stemming, feature, selection, lost contracting, feature weighing, and finally map all the information in the dictionary. Based on the dictionary or extracted BOW, we build the classifier system, categorize the document and change the BOW to vectors. The process of automatic knowledge extraction contains concepts and entities which used a mapping dictionary to recognize the meaning of the document. The algorithms which applied in this procedure named as Backward Maximum Matching (BMM) and mainly presented for Chinese word segmentation which contains various characters. In the same way, with maxlen words from right to left of the text and similarly mapping, the concept and entities of text in Pro base. Figure 2 shows the BMM process in details.

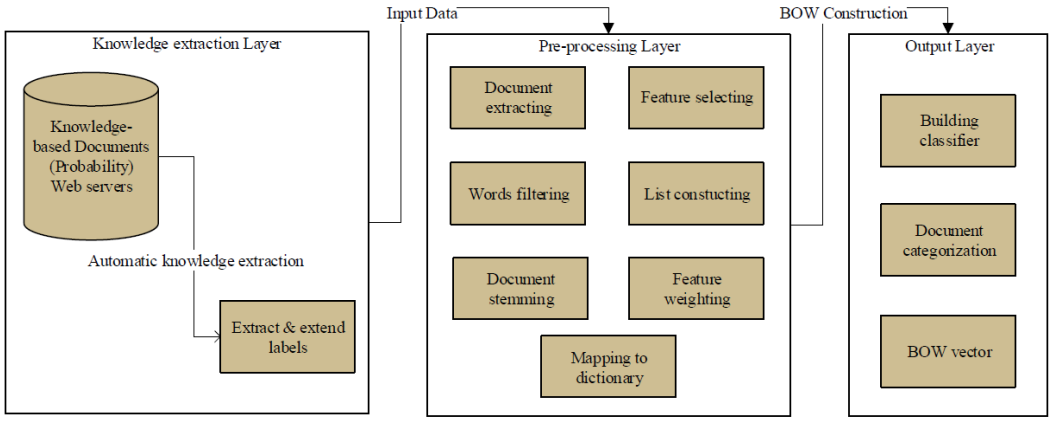

Figure 1. Overview of the proposed automatic knowledge extraction approach

Before applying machine learning algorithms, data pre-processing, sentence segmentation and word tokenization processed in the dataset. In the next step, sentence concept recognition and entity processed based on the designed flowchart. In this system, two mapping dictionaries defined. The proposed defining two dictionaries are to identify the concepts and entities easier. One dictionary defines entities and in the same way, another for concepts. Single words don't contain in any of mentioned dictionaries.

Copyright ⓒ 2020 SERSC

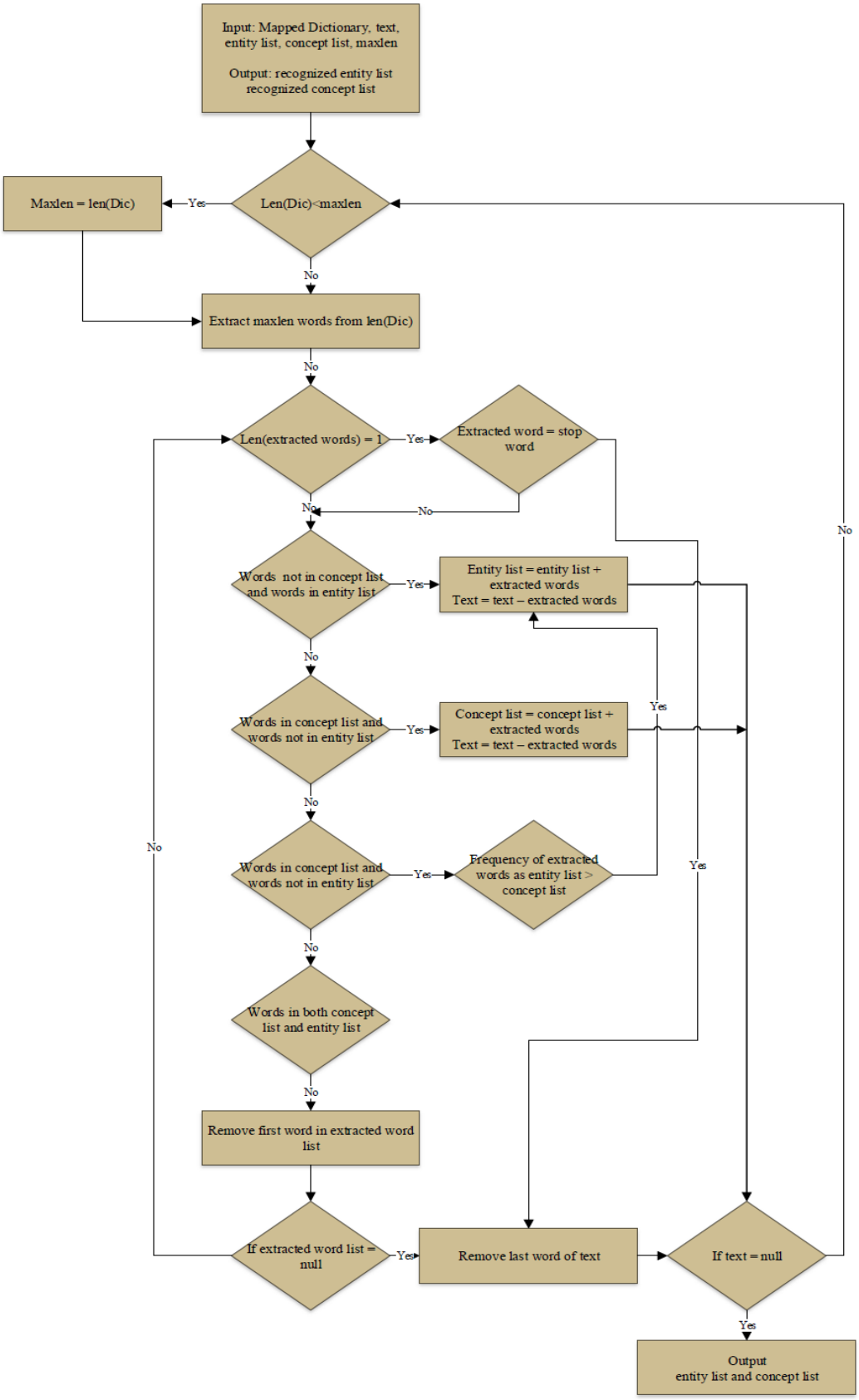

Figure 2. Backward Maximum Matching procedure flowchart

Copyright ⓒ 2020 SERSC

4. Implementation and Results

This section presents the development environment in detail. Experimental setup shows in Table 1. All experiments and results of the system are carried out using Intel(R) Core (TM) i7-8700 CPU @3.20GHz processor with 32 GB memory. Library and framework used in the proposed system are Jupyter notebooks. The programming language used in the designing of this system is WinPython–3.6.2.

Table 1. Development Environment of Proposed System

Component	Description
Programming language	WinPython 3.6.2
Operating system	Windows 10 64bit
Browser	Google Chrome, opera
Library and framework	Jupyter notebook
CPU	Intel(R) Core (TM) i7-8700 CPU @3.20GHz
Memory	32GB

4.1 Dataset

The dataset in the proposed system collected from six real-life information web pages. 20newsgroups, two categories from Reuters-21578 (R8) and (R52), B.B.C., B.B.C. Sport and Yahoo are the mentioned six real-life datasets. The collected datasets used for document classification and calculate various ways for document representation. Table 2 shows the dataset statistics.

Table 2. Dataset

Statistic	20NG	R8	R52	BBC	BBC Sport	Yahoo
Vocab size	228271	20067	33359	32025	9748	992905
Class No	20	8	52	5	5	10
Sample No	18845	7674	9100	2225	737	1460k
Avg. words	352	98	215	481	434	90
Train	22425	6596	7643	2223	478	1400k
Test	8643	3291	3679	2224	479	60k

4.2 Experimental Setting

In the presented BOC system, we compare the algorithm with other benchmarks, e.g. BOW, LDA, LSA and Doc2vec. 1) BOW: Bag of Words model is the baseline model in this system based on sklearn. Stop words and punctuations cleared and in the same way, the words with a frequency less than two also cleared. 2) LSA: Latent Semantic Analysis is one of the implementation processes based on sklearn. 3) LDA: Latent Dirichlet Allocation is another implementation process based on sklearn. LDA is extracting the probability of topics in the document. 4) Doc2vec: is famous as paragraph vector model. The implementation is based on gensim. Table 3 shows the classification accuracy based on the presented benchmarks and dataset.

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

Table 3. Different document representation dimensionality

Datasets	BOW(%)	LSA(%)	LDA(%)	Doc2vec(%)
20NG	90.43	90.35	86.36	79.47
R8	97.64	97.51	96.31	95.96
R52	96.44	96.31	91.32	92.18
BBC	97.97	98.32	96.26	97.48
BBC Sport	98.48	98.75	93.52	97.48
Yahoo	83.21	83.24	80.92	77.57

4.3 Model Semantic Analysis

Based on the mentioned process in the previous step, document dimensionality evaluated. The high dimensional feature vector is defined to use a large number of features for document representation. Among the compared features, Doc2vec has the lowest dimensionality. The features extracted from words, distributed representation, and it's not easy to explain. Similarly, it's difficult for a human to discover the meaning of semantic from representations. Table 3 shows the BOW, LSA and LDA representation dimensionality from various document classification.

Table 4. Top three features for various document representation

Model	Document 1
BOW	House, rutabaga, Leary
LSA	ibm, father, mother, house, price, students kid, lee, Holland, spider, insect kid, house, America, radio, gift
LDA	Ankle, hose, rutabaga, injury, five House, Leary, we, four, Southampton Leg, Robben, days, injury, five
Model	Document 2
BOW	Wars, Timothy, rating
LSA	Say, Mrs, will, people, month Yellow, Galloway, hybrid, music, Glaser Movie, worth, gift, economy, present
LDA	Movie, gift, tell, star, candidacy Movie, Stefan, tell, life, various Morrison, album, save, sign, record

Table 4 records the LSA and LDA top-weighted topics. Each topic meaning is based on the extracted top-weighted words. LSA accomplish the combination of the linear words that has a positive effect in statistical identification. LDA, extracts the topics out of the document and distribute words co-occurrences to process better interpretability than LSA Table 5 shows the classification accuracy of various algorithms based on the human being.

Table 5. Top three features for various document representation

#	BOW (%)	LSA (%)	LDA (%)
1	0.77	0.59	0.85
2	0.57	0.53	0.69
3	0.73	0.73	0.79
Avg	0.691	0.617	0.778

Copyright ⓒ 2020 SERSC

Conclusion

Text representation is one of the definite steps in various computer science topics. In this paper, we present the concept of knowledge representation based on document classification using topic modelling techniques and BOW. The decisive point of the proposed system is automatically extracting the perceptual knowledge from the document based on the probabilistic matters and obtain the semantic relationship in the highest meaning. By combining the concepts and entities in the document, distributed vectors become as a concept space for document frequency and weight. Moreover, BOW and BOC provide relatively high interpretability which is effective for deep human understanding. The experimental result shows the semantic quality of BOW and BOC that obtain the document information level and similarly improving the performance of context.

Conflict of interest

"The authors declare no conflict of interest".

Acknowledgements

This research was supported by the 2019 scientific promotion program funded by Jeju national University.

References

[1] T.A. Almeida, T.P. Silva, I. Santos, J.M.G. Hidalgo, Text normalization and semantic indexing to enhance instant messaging and sms spam filtering, Knowl.-Based Syst. 108 (2016) 25–32.
[2] J. Dang, M. Kalender, C. Toklu, K. Hampel, Semantic search tool for document tagging, indexing and search, Google Patents, 2017, U.S. Patent 9, 684, 683.
[3] Y. Xiao, B. Liu, J. Yin, Z. Hao, A multiple-instance stream learning framework for adaptive document categorization, Knowl.-Based Syst. 120 (2017) 198–210.
[4] S. Joty, G. Carenini, R.T. Ng, Topic segmentation and labeling in asynchronous conversations, J. Artificial Intelligence Res. 47 (2013) 521–573.
[5] G. Lee, J. Jeong, S. Seo, C. Kim, P. Kang, Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network, Knowl.-Based Syst. 152 (2018) 70–82.
[6] Y. Li, H. Guo, Q. Zhang, M. Gu, J. Yang, Imbalanced text sentiment classification using universal and domain-specific knowledge, Knowl.-Based Syst. 160 (2018) 1–15.
[7] W. Zhang, T. Yoshida, X. Tang, A comparative study of tf* idf, lsi and multi-words for text classification, Expert Syst. Appl. 38 (3) (2011) 2758–2765.
[8] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 655–665.
[9] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1422–1432.
[10] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, Technical report, OpenAI, 2018.
[13] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.
[14] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

Copyright ⓒ 2020 SERSC

[15] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, 2014, pp. 1188–1196.
[16] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Vol. 1, 2018, pp. 2227–2237.
[17] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Amer. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[18] Y. Song, H. Wang, Z. Wang, H. Li, W. Chen, Short text conceptualization using a probabilistic knowledgebase, in: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Volume Three, AAAI Press, 2011, pp. 2330–2336.
[19] F. Wang, Z. Wang, Z. Li, J.-R. Wen, Concept-based short text classification and ranking, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, 2014, pp. 1069–1078.
[20] W. Hua, Z. Wang, H. Wang, K. Zheng, X. Zhou, Short text understanding through lexical-semantic analysis, in: Data Engineering (ICDE), 2015 IEEE 31st International Conference on, IEEE, 2015, pp. 495–506.

Authors

Zeinab Shahbazi received her B.S. in software engineering from Pooyesh University, IRAN. In March 2017, she moved to Republic of Korea for M.S studies and started working in internet laboratory, Chonbuk National University (CBNU). After completing her master in 2018, she moved to Jeju-do in March 2019 and started working as a Ph.D. research fellow in Machine Learning Laboratory (MLL), Jeju National University. Research interests include artificial intelligence and machine learning, natural language processing, deep learning and data mining.

Yung Cheol Byun received his B.S. from Jeju National University, Korea in 1993, M.S and Ph.D degrees from Yonsei University in 1995 and 2001. He worked as a special lecturer in SAMSUNG Electronics in 2000 and 2001. From 2001 to 2003, he was a senior researcher of Electronics and Telecommunications Research Institute and he promoted to join Jeju National University as an assistant professor in 2003, where he is currently a professor of Department of Computer Engineering. From 2012 to 2014, he had research activities at University of Florida as a visiting professor. His research interests include the areas of pattern recognition & image processing, artificial intelligence & machine learning, security based on pattern recognition, home network and ubiquitous computing, u-Healthcare and RFID & IoT middleware system (Corresponding Author).

Dong Cheol Lee received his B.S. in electrical engineering education from Chungnam National University, Korea in 1986. M.S. in MIS degrees from Kookmin University, and Ph.D. industrial engineering degrees from Sungkyunkwan University in 2002. He worked as a professor at the Tourism Information Processing Department of Jeju Tourism College from March 1993 to April 2003. He has been a professor at Jeju National University's Department of Management Information systems since April 2003. He is interested in e-commerce, intelligent agent, and digital cultural content (Corresponding Author).

Copyright ⓒ 2020 SERSC