A Comparative Study of Topic Modeling Methods for Document Retrieval

Ibrahim Reyad
Information Systems Department
Faculty of Computers & Artificial Intelligence, Benha University, Egypt
ibrahim.elsayed@fci.bu.edu.eg

Metwally Rashad
Computer Science Department
Faculty of Computers & Artificial Intelligence, Benha University, Egypt
Artificial Intelligence, Delta university for science and technology, Gamasa, Egypt
metwally.rashad@fci.bu.edu.eg

Mohamed Abdelfatah
Information Systems Department
Faculty of Computers & Artificial Intelligence, Benha University,Egypt
mohamed.abdo@fci.bu.edu.eg

***Abstract*—In information retrieval, topic modeling is used to predict hidden subjects from a text corpus. As a result, it offers a system for organizing, comprehending, and summarising vast amounts of text data automatically. Topic modeling can also be used to provide a document representation that may be interpreted in a broad spectrum of natural-language processing (NLP) applications Techniques for modeling span from probabilistic graphs models to neural models. This research looks at subject models from a variety of angles. The first element categorizes subject modeling strategies as algebraic, fuzzy, probabilistic, or neural. We investigate enormous number of handy models in every group, emphasize contrasts and similarities among models and model types from a unifying standpoint, examine the properties and limits of these models, and debate the correctness of the application. Another factor to consider is a review of datasets and benchmarks. We talk about studies done on datasets to evaluate topic models along stated measures. The study focuses on the contrast between both models and their rightness for miscellaneous applications.**

***Index Terms*—Topic Model, Neural, Probabilistic, LDA**

I. Introduction

The topic model (TM) has indeed been effectively used in the mining of large corpus of text. A topic model picks a group of textual data as input and then strives to identify the fundamental topics in this group [1]. Each subject described a linguistic concept that is understandable by humans. Following that, topic modeling generates a hidden interpretable text representation based on conceived subjects [2].

The topic model has been applied in a multitude of domains. Natural Language Processing (NLP) programs like summarization and text analytics have made use of the topic model. Furthermore, the topic model is being extended to new disciplines like medicine and economics. TM is also consumed with other sorts of data, like when the concepts of words and documents are substitute by similarly structured items. Entities might be physical items in online marketplaces, visual components, or genetics in gene sets.

Since the launch of Latent Semantic Indexing (LSI) in 1990 [3] researchers have created a variety of topic modelling strategies with varying modelling capacities. Furthermore, different models make various hypotheses about the corpora, the representation of documents, and the themes. The generated topic models cover a wide range of applications. They also have several estimation metrics scores. More models are still being developed. This ongoing work aims to fill present research gaps and broaden the scope of subject modelling applications.

Depending on their basic modeling methodologies, topic models are classified into four types: algebra, fuzzy, probability, and neural. Algebra topic models were expanded for the first time in the 1990s [3]. Following the coming of LDA in 2003, the vast majority of topic models developed were Bayesian Probability Topic models (BPTMs), like [1, 4, 5]. Until 2015, these BPTMs have proved their effectiveness and controlled research areas. They are easy to deploy, easy to explain, and flexible enough to be developed into more complex models. Because of the minimal number of parameters [1], they also are computationally more efficient. Topic models are gradual including neural, components [6]. Instances of neural topic models are [7, 9](NTMs). Most NTMs utilize contextualized representations to represent text documents rather than the traditional vector space model. They can work with pre-trained methods on a large corpora of data like BERT variations [13]. NTM, in general, suggests flexibility and scalability. It is critical to remember that various TM types provide distinct benefits and are best appropriate to a certain situation. Hence, they get along.

Because there are numerous models available, they must be summarised. In the literature, there have been attempts to summarise the vast array of extant topic models. [14] gave two perspectives. The first categorization categorizes models into four sub-types: Latent semantic indexing (LSI), Probabilistic latent semantic indexing (PLSI), Latent Dirichlet allocation (LDA), and Correlated topic model (CTM). The second type investigates several topic evolution models over time. In [15], the survey provided hierarchical classification criteria, with models categorized as LDA or non-LDA-based, bag-of-words or sequence-of-words technique, and finally, unsupervised or supervised learning models. [16] examined many models in terms of their strengths and disadvantages. [17]'s authors grouped topic models into three different kinds. The first is a traditional topic model such as LSI, PLSI, and LDA. The

second type address evolution models. Finally, the third class looks at topic modeling and various algorithm partnership tactics like LDA-VSM+K-means and LDA-Word2vec+SVM. [18]'s writers looked at previous studies involving LDA topic modeling. [19] provided a comprehensive assessment of topic modeling for both long and short text over the last decade, while [20] examined the particular class of neural topic models.

This study conducts a thorough survey of TM. It covers the four types of TM and organizes a broad number of ready for use models. The goal is to draw attention to research gaps, discrepancies, and similarities among models and model types. In addition, because assessment criteria differ amongst publications, we present a full set of evaluation measures that may be used in all models to assure model evaluation consistency. In addition, available topic modeling datasets are discussed, followed by a review of different topic modeling applications and tools. We next test a variety of models from diverse sectors and evaluate them using common datasets along the specified assessment criteria. This point of view considers the much more recent breakthroughs in numerous issue modeling methodologies. It is intended that it will pique the interest of the audience in TM and help the scientific community produce better models.

The rest of this work is arranged as follows. Section 2 contains background information, definitions, and notation, whereas Section 3 describes the approach and discusses model categorization, which is followed by an examination of several models from the groups. In Section 4, we evaluate numerous models from various types using the suggested metrics, and in Section 5, we discuss the results, noting trends in research and highlighting research breaks. Finally, the paper is finished.

II. Topic Models

A. Background and Definitions

In information retrieval, topic modeling is applied to discover invisible topics in documents, providing an automated mechanism to arrange, comprehend, and summarise vast volumes of textual material. Topic models can also be utilized to provide an interpretable presentation of text in various downstream Natural Language Processing (NLP) applications.

Three entities are attempted to be modeled by topic models: constructions, sets, and subjects. The constructions are the components that make up a collection. Constructs in text data are typically words that are put together to form a text or a set of words. A topic is a group of constructs that describe a single semantic meaning. It is an idealistic presentation of a document as pure as feasible. Furthermore, it is a homogeneous set of constructs that have a lot in common, usually conceptually. A topic is mathematically defined as a probability distribution over the constructs [22].

The majority of topic models operate by detecting the co-frequency of structures in collections. This corresponds to the distributional probability [22, 23], which claims that the distributional features of tokens determine their semantic meaning. In other words, the context in which a word appears defines it. Similarly, because to spatially local correlations between pixels, the co-occurrence of pixels in images defines segments (topics).

Topic Model can be applied to a variety of data types, however we will only discuss their use in text in the following sections. Textual data is assumed to be a diverse group of structures that span more than one topic. Topic sparsity is also assumed, based on the heuristic that any given document will typically cover a small number of topics.

Topic models are fed corpus D and attempt to generate two sets of distributions: The first set T is for topics: K distributions over V constructions (tokens), where K is the number of topics, which is frequently a hyperparameter, and V is the corpus vocabulary size. The second set of distributions Z is for documents, and it is a distribution over K topics for each document in the corpus, where z_k specifies the weight of the kth subject for each document. This second release provides a document-interpretable latent representation (Similar to disentangled Variational Auto Encoder, VAE [24, 25]). To best interpret the observed texts, these latent representations and T (TM parameters) are trained. This is referred to as the inference process. The generative process produces a document from a presumed Z and T.

TM can reveal polysemy, which occurs when a single concept has many semantic interpretations [26]. TM reflects this by providing the same construct appear in multiple themes. When distinct constructions exist together in the same subject in more than one distribution of T, TM can capture synonymy, where 2 distinct constructs have identical semantic meaning.

Document representation varies depending on the model. Some models regard the collections as an unordered collection of its constructs. These models are known as the bag of words models in literature [23]. In this scenario, the topic model's input is a vector of word counts.

Another document representation that is gaining popularity due to the prevalence of transformer-based language models such as BERT is the contextual representation, in which the document is represented as an embedding vector. Because they learn distinct representations for polysemous words, BERT models capture contextual word embeddings better than global embeddings.

B. Methodology

There are four types of models: algebraic, fuzzy, Bayesian probabilistic, and neural topic models. This classification.

The algebraic models are the first TM category. They are straightforward, but lack a solid statistical foundation and fail to construct a robust generative data model [20]. This linear algebra technique was first proposed in [3]. There are also more contemporary approaches that use spectrum decomposition for inference, such as [26]. Algebraic models are straightforward and generally efficient. Fuzzy topic models, on the other hand, are clustering approaches that aim to allocate words to related subjects. They have demonstrated their usefulness in short text documents by overcoming the sparsity challenge.

Prior to 2015, Bayesian probabilistic models were prevalent, but research efforts turned to building neural topic models (NTMs). Bayesian models are simple to use and adaptable. They define a fake generation mechanism for the observed papers. Then they go backward to deduce the issues that could have given rise to the documents. The parameters of Bayesian models have more specific meaning and are thus easier to interpret. This parameter interpretability is useful for troubleshooting, validating, and interpreting results, especially in cross-disciplinary use cases. It is also useful when we need to generate more data. In this group of models, however, the inference problem grows more challenging as the modelling becomes more complex.

Because the inference problem in NTMs is an optimization problem, they are more flexible and scalable. Their parameters, however, may not have a valid interpretation. It is frequently difficult to determine why a model works or does not work. NTMs train with various objectives in mind, such as lowering document reconstruction error or improving the model's forecast accuracy. This optimization aim does not always produce high-quality interpretable topics. Table 1 presents a comparison of various modelling methodologies.

III. Literature Review

Topic modelling has proven to be effective not just for information retrieval but also in text categorization and exploratory analysis of large corpora. It is hardly unusual, then, to find examples of topic modelling in a variety of academic subjects, like information science, social sciences, legal studies, and the humiliates. In the next paragraphs, I will mention a few papers in these areas. These research form the basis for my further investigation of the criteria required for successful topic modelling.

In the area of information science, there are numerous examples of using topic models to categorize scientific papers. For text economic stability report quantification and visualization in [32] suggested a novel approach called LDA in order to evaluate and visualize economic security. It is made up of several steps. To begin, it is acceptable to assume that the LDA model could be used to assess the Economic Stability Report for China. Second, by splitting the core words of each subject in essential terms and specific terms, We could create visuals of economic and topical entanglements rating of 5 years or for each and every year analyzed, resulting in a design matrix and the balance of the economy investigated trends. Finally, the word cloud may readily illustrate the macro-environment in finance. This technique fared well on the China Financial Stability Report [32]. Readers will discover that each clustered theme is tied to a specific chapter, and by splitting core terms of each topic into basic keywords and specific keywords, visuals of every financial embranchment may be formed. The LDA model could analyze a Document-Term Matrix (DTM) to generate a word cloud. A picture of the macro-environment can also be drawn from the figure from 2012 to 2016.

Because of the essence of LDA, it is a great tool for document organization and corpora summarization. One drawback of LDA would be that its success is strongly dependent on priors. Researchers demonstrate that priors matter in LDA and provide strategies for learning priors for improved modeling, in spite of whether symmetric priors are used. But, LDA execution does not necessarily correspond to LDA modelling capabilities in document presentation and topic elicitation. A fresh previous setting for LDA was proposed in [42]. The option increases LDA's performance in both document presentation and topic elicitation. Unlike the prior approach, which develops the priors when making posterior inferences, the recommended set of parameters teaches the parameter Prior to inference. Experiments were carried out on the Health Twitter Dataset [**?**] and the 20 Newsgroups dataset [61, 63] to validate the efficacy of the proposed previous setup. Evaluations on the topic's quality reveal that LDA with the previous setup gains superior quality subjects. Analyze the document presentation of LDA with suggested priors, LDA with past prior development, and LDA with symmetrical priors by performing standard tasks like document clustering and document categorization to illustrate the improvement in document representation quality. The findings confirm that the proposed previous settings improve LDA's document representation power.

An developed LDA algorithm for text categorization termed gLDA was developed in [43], in which topic-category distribution parameters were incorporated in the structure of LDA, a flexible generative probabilistic model for discrete datasets. gLDA adds document categories based on a simple interchangeability hypothesis for words and subjects in a text. The subjects are constructed by assembling papers for specific groups, and the probability distributions are calculated by normalizing the word frequency of the linked texts. Text documents in this model can be automatically classified into many groups. In our model, Gibbs sampling was used for parameter evaluation instead of the EM technique, which tested to have a reduced perplexity consumption on parameter estimation. gLDA offers a more successful modeling method than LDA for word-sense disambiguation since it uses a Dirichlet forest prior over subjects in a group. Our framework is a little more broad. We increase the goodness of make a forecast on text data by stating which words should have a high likelihood in a subject. The model describes the subjects that should have a high likelihood in a group. Text documents in the same area of knowledge could be considered in the iteration of the topic modeling process by adding the multinomial distribution over subjects in a category; this is significantly different from an iterative topic modeling approach that mostly relies on human inspection. Furthermore, the topic model did not require label for data. Furthermore, there are a number of potentially helpful avenues in which the modified structure of topic model might be extended. The use of gLDA in text retrieval is an important direction that should be pursued. To test the model's validity, The evaluation of gLDA was evaluated by assessing its ability to categorize new documents on which the model had not been trained. Documents from the same genre should be generalized in models trained from that genre.

A text retrieval technique that can compute test similarity relying on the hybrid LDA and word2vec is presented in [44]. Word2Vec refers to a set of interrelated models that are made to generate word embeddings. These are shallow, 2-layer neural networks that have been taught to construct the linguistic context of words. A vast corpora of text is fed into the Word2Vec model, which subsequently generates a vector space with hundreds of vectors. Furthermore, every distinct word in the corpora is appointed to a corresponding dimension space; hence, words in the corpus that have public vocabulary are placed in close adjacency to one another in the vector space. The projecting layer is then utilized to create a contextual vector. after that, In terms of the occurrence in the corpus, the word in the dictionary is treated as a leaf node in the output layer. Following that, each text may be characterized as a distance distribution from a text to all of the distinct topics in the same space. The 20Newsgroups dataset [61, 63] was used to test the performance of the Hybrid LDA and Word2Vec model, and the findings show that the technique is effective.

For text document clustering, an updated ant method with LDA-based representation is used in [46], two new heaps combing algorithms depended on the most diverse objects in heaps and objects-center distances were introduced in order to improve the clustering quality of ant-based clustering. The suggested heaps combining procedures address the problem of the Class algorithm producing an excessive number of groups (heaps). Furthermore, the LDA is used to present text sets as a group of latent topics. The algorithm is made up of four steps. The algorithm begins with an ant-based technique for item grouping. The K-means algorithm is then applied. The ant-based clustering process is then done again, and the algorithm concludes by performing the K-means algorithm on items once more. Each data in this system comprises a vector of n real values. A distance measure is used to calculate the distance between two data objects. Like to other ant-based clustering techniques, data objects are randomly distributed on a grid. Ants can generate, build, and destroy piles, which is a fundamental difference in the algorithm. A heap is a type of data structure that holds two or more things. A heap can exist on a single grid. On 25 text benchmarks, the suggested clustering techniques are compared to conventional and metaheuristic clustering algorithms in terms of F-measure. The experimental outcomes show that the merging strategy in heaps based on the most dissimilar objects surpasses the other clustering approaches used in the empirical research. Furthermore, The experimental results reveal that the suggested object-center distance-based heap merging strategy outperforms the K-means and AntClass algorithms.

As proposed in [46], Topic Modeling was employed as a classification approach, standed for as Self-training with Latent Dirichlet Allocation (ST LDA), that is experienced semi-supervised. The self-training method was utilized for this goal, and the documents were represented with topic allocation depended on the LDA topic model. The ST LDA method accepts as input a lean labeled documents that make an introductory labeled set and many more unlabeled documents; the unlabeled documents might be formed of diverse sources with a similar subject distribution as the first labeled set. The method generates a classification model using a supervised classifier on a final labeled set - the introductory labeled set augmented with labeled instances from an unlabeled set. As a result, ST LDA is an inductive classification algorithm. Experiments were run on six different datasets: 20 Newsgroups [63], Reuters R8, Reuters R52 [61], WebKB [64] with four classes, Ohscal [46], and Google snippets [46]. According to the results, the suggested ST LDA approach beats other methods in 9 of 11 initial labeled sets.

IV. Conclusion

This survey provided a thorough overview of current developments, prospective research trends, potential new research fields, and a new classification of state-of-the-art methodologies in the area of Quality Estimation (QE). The investigation focused on four important areas: (1) data source, which is a group of documents used to expand the user's beginning query, (2) working approach, that describes the procedure for widening the query, (3) significance and application, which examines the significance of QE in IR and its use in the modern trend beyond the key field of IR, and (4) Basic approaches, which explain several QE methods based on various features of data sources. Furthermore, we have supplied several datasets for textual data from diverse sources in this study. We described how cutting-edge text document retrieval methods were used on these datasets in order to obtain high accuracy and narrow the semantic gap. We also discussed the relationship of approaches to one another and how one technique increased retrieval better than others, as well as the benefits and drawbacks of each strategy. As we can see, the process of retrieving text content from big repositories is fraught with difficulties. So, we hope that our survey can assist researchers in examining the drawbacks of current text document retrieval strategies in order to enhance them and offer the best way.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[2] S.P. Crain, K. Zhou, S.-H. Yang, H. Zha, Dimensionality reduction and topic modeling: From latent semantic indexing to latent Dirichlet allocation and beyond, in: C.C. Aggarwal, C. Zhai (Eds.), Mining Text Data, Springer US, Boston, MA, 2012, pp. 129–161.

[3] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.

[4] J. Lafferty, D. Blei, Correlated topic models, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, Vol. 18, MIT Press, 2005, p. 8.

[5] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine

Learning - ICML '06, ACM Press, Pittsburgh, Pennsylvania, 2006, pp. 113–120.
[6] Z. Cao, S. Li, Y. Liu, W. Li, H. Ji, A novel neural topic model and its supervised extension, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[7] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TFIDF procedure, 2022.
[8] A.B. Dieng, F.J.R. Ruiz, D.M. Blei, Topic modeling in embedding spaces, Trans. Assoc. Comput. Linguist. 8 (2020) 439–453,.
[9] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, 2017, arXiv:1703.01488 [stat].
[10] Y. Miao, E. Grefenstette, P. Blunsom, Discovering discrete latent topics with neural variational inference, 2018, arXiv:1706.00359 [cs].
[11] F. Bianchi, S. Terragni, D. Hovy, Pre-training is a hot topic: Contextualized document embeddings improve topic coherence, 2021, arXiv 2004.03974 [cs].
[12] F. Bianchi, S. Terragni, D. Hovy, Pre-training is a hot topic: Contextualized document embeddings improve topic coherence, 2021, arXiv 2004.03974 [cs].
[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, arXiv:1810. 04805 [cs]
[14] R. Alghamdi, K. Alfalqi, A survey of topic modeling in text mining, Int. J. Adv. Comput. Sci. Appl. 6 (1) (2015).
[15] D. Sharma, A survey on journey of topic modeling techniques from SVD to deep learning, Int. J. Mod. Educ. Comput. Sci. 9 (2017) 50–62.
[16] B.V. Barde, A.M. Bainwad, An overview of topic modeling methods and tools, in: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, Madurai, 2017, pp. 745–750.
[17] L. Xia, D. Luo, C. Zhang, Z. Wu, A survey of topic models in text classification, in: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2019, pp. 244–250.
[18] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimedia Tools Appl. 78 (11) (2019) 15169–15211.
[19] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimedia Tools Appl. 78 (11) (2019) 15169–15211.
[20] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimedia Tools Appl. 78 (11) (2019) 15169–15211.
[21] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, Association for Computing Machinery, New York, NY, USA, 1999, pp. 50–57.
[22] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, Association for Computing Machinery, New York, NY, USA, 1999, pp. 50–57.
[23] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, Commun. ACM 8 (10) (1965) 627–633.
[24] Z.S. Harris, Distributional structure, WORD 10 (2–3) (1954) 146–162.
[25] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[26] C.P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, Understanding disentangling, 2018.
[27] D.E. Klein, G.L. Murphy, The representation of polysemous words, J. Memory Lang. 45 (2001) 259–282.
[28] A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade, M. Telgarsky, Tensor Decompositions for Learning Latent Variable Models, Tech. Rep., Defense Technical Information Center, Fort Belvoir, VA, 2012.
[29] M. Rashad, I. Reyad and M. Abdelfatah, "Topic Modelling with Bag-of-concepts Document Representation," 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2022, pp. 216-220, doi: 10.1109/NILES56402.2022.9942412.
[30] H. Kyul Kim, H. Kim and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation", Neurocomputing, 2017 Nov 29;266:336-52.
[31] L. Li, B.Qin, W. Ren and Liu T. Document representation and feature combination for deceptive spam review detection. Neurocomputing. 2017 Sep 6;254:33-41..
[32] G. Li, X. Zhu, J. Wang, D. Wu and J. Li, "Using lda model to quantify and visualize textual financial stability report", Procedia computer science, 2017 Jan 1;122:370-6.
[33] Y. Djenouri , A. Belhadi and R. Belkebir ,"Bees swarm optimization guided by data mining techniques for document information retrieval", Expert Systems with Applications, 2018 Mar 15;94:126-36.
[34] S. Hao, C. Shi, Z. Niu and L. Cao, "Modeling positive and negative feedback for improving document retrieval", Expert Systems with Applications, 2019 Apr 15;120:253-61.
[35] G. Rao, W. Huang, Z. Feng and Q. Cong, "LSTM with sentence representations for document-level sentiment classification", Neurocomputing, 2018 Sep 25;308:49-57.
[36] W. Zhang, Y. Li and S. Wang, "Learning document representation via topic-enhanced LSTM model", Knowledge-Based Systems, 2019 Jun 15;174:194-204.
[37] K. Kowsari, DE. Brown, M. Heidarysafa, KJ. Meimandi, MS. Gerber and LE. Barnes, "Hdltex: Hierarchical deep learning for text classification", In2017 16th IEEE international conference on machine learning and applications (ICMLA), 2017 Dec 18 (pp. 364-371). IEEE.
[38] Z. Zhang, L. Wang, X. Xie and H. Pan, "A graph based document retrieval method", In 2018 IEEE 22nd International Conference on Computer Supported Cooperative

Work in Design ((CSCWD)) 2018 May 9 (pp. 426-432). IEEE.

[39] F. Ensan and E. Bagheri, "Document retrieval model through semantic linking", In Proceedings of the tenth ACM international conference on web search and data mining, 2017 Feb 2 (pp. 181-190).

[40] B. Mitra, F. Diaz and N. Craswell, "Learning to match using local and distributed representations of text for web search", In Proceedings of the 26th International Conference on World Wide Web 2017 Apr 3 (pp. 1291-1299).

[41] R. Syed, K and Collins-Thompson, "Retrieval algorithms optimized for human learning", InProceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval 2017 Aug 7 (pp. 555-564).

[42] J. Ding, W. Jin, "A Prior Setting that Improves LDA in both Document Representation and Topic Extraction", In 2019 International Joint Conference on Neural Networks (IJCNN) 2019 Jul 14 (pp. 1-8). IEEE.

[43] D. Zhao, J. He, J. Liu, "An improved LDA algorithm for text classification", In 2014 International Conference on Information Science, Electronics and Electrical Engineering 2014 Apr 26 (Vol. 1, pp. 217-221). IEEE.

[44] M. Xue, "A text retrieval algorithm based on the hybrid LDA and Word2Vec model", In2019 International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS) 2019 Jan 12 (pp. 373-376). IEEE.

[45] A. Onan, H. Bulut, S. Korukoglu, "An improved ant algorithm with LDA-based representation for text document clustering", Journal of Information Science. 2017 Apr;43(2):275-92.

[46] M. Pavlinek, V. Podgorelec, "Text classification method based on self-training and LDA topic models". Expert Systems with Applications. 2017 Sep 1;80:83-93.

[47] Z. Chao, "Research on English translation long text filtering based on LSTM semantic relevance", Microprocessors and Microsystems. 2021 Feb 1;80:103574.

[48] A. Berger, J. Lafferty. "Information retrieval as statistical translation", InACM SIGIR Forum 2017 Aug 2 (Vol. 51, No. 2, pp. 219-226). New York, NY, USA: ACM.

[49] Z. Dai, C. Xiong, J. Callan,and Z. Liu, "Convolutional neural networks for soft-matching n-grams in ad-hoc search", In Proceedings of the eleventh ACM international conference on web search and data mining 2018 Feb 2 (pp. 126-134).

[50] MH. Al-Bayed, SS. Abu-Naser, "Intelligent Multi-Language Plagiarism Detection System",(2018).

[51] H. Zamani, B. Mitra, X. Song, N. Craswell,and S. Tiwary, "Neural ranking models with multiple document fields", In Proceedings of the eleventh ACM international conference on web search and data mining 2018 Feb 2 (pp. 700-708).

[52] R. Zhao, K. Mao, "Fuzzy bag-of-words model for document representation", IEEE transactions on fuzzy systems. 2017 Mar 31;26(2):794-804.

[53] V. Lavrenko, WB. Croft, "Relevance-based language models", In ACM SIGIR Forum 2017 Aug 2 (Vol. 51, No. 2, pp. 260-267). New York, NY, USA: ACM.

[54] A. Curiel, C. Gutiérrez-Soto, JR. Rojano-Cáceres, "An online multi-source summarization algorithm for text readability in topic-based search", Computer Speech and Language. 2021 Mar 1;66:101143.

[55] E. Delasalles, S. Lamprier, L. Denoyer, "Deep dynamic neural networks for temporal language modeling in author communities", Knowledge and Information Systems. 2021 Mar;63(3):733-57.

[56] A. Radhika, MS. Masood, "Effective dimensionality reduction by using soft computing method in data mining techniques", Soft Computing. 2021 Mar;25(6):4643-51.

[57] M. Aman, SJ. Abdulkadir, IA. Aziz, H. Alhussian, I. Ullah, "KP-Rank: a semantic-based unsupervised approach for keyphrase extraction from text data", Multimedia Tools and Applications. 2021 Jan 11:1-38.

[58] R. Aragao, TE. El-Diraby,"Network analytics and social BIM for managing project unstructured data", Automation in Construction. 2021 Feb 1;122:103512.

[59] A. Mandal, K. Ghosh, S. Ghosh, S. Mandal, "Unsupervised approaches for measuring textual similarity between legal court case reports", Artificial Intelligence and Law. 2021 Jan 4:1-35.

[60] YA. Al-Lahham, "Index Term Selection Heuristics for Arabic Text Retrieval", Arabian Journal for Science and Engineering. 2021 Apr;46(4):3345-55.

[61] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[62] J. Li, M. Ott, C. Cardie, E. Hovy. "Towards a general rule for identifying deceptive opinion spam", InProceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2014 Jun (pp. 1566-1576).

[63] http://qwone.com/~jason/20Newsgroups/

[64] http://www.cs.cmu.edu/~webkb/

[65] L. Liu, MT. Özsu, editors. Encyclopedia of database systems. New York, NY, USA:: Springer; 2009 Sep 29.