Categorizing Research Papers By Topics Using Latent Dirichlet Allocation Model

Mahesh Korlapati, Tejaswi Ravipati, Abhilash Kumar Jha, Kolla Bhanu Prakash

Abstract: Topic models are extensively used to classify documents into topics. There are many topic models in the field of text analysis which classify the documents efficiently. In this paper, we propose a method to categorize research papers to topics using LDA or Latent Dirichlet Allocation Model for text categorization. The research papers are selected randomly and based on variational parameters; every document is classified into a topic. It is proved that the LDA model among the topic models is efficient, reliable and simple for classifying textual data. By our observations, we infer that the LDA model is better than the conventional Naive Bayes and Support Vector Machines in the field of text classification. We use the Cora dataset to observe the distribution of ten topics over nine randomly selected scientific documents.

Index Terms: Text Analysis, topic modeling , Natural Language Processing, LDA, topic models, topic.

1. INTRODUCTION

Many models in text analysis follow the Bag of Words model, where all the words in a document are considered and the relationship among the words is neglected. Due to this drawback and poor representation of text, the conventional algorithms that follow the Bag of Words model yield unsatisfactory results. Recently, new and efficient methods of statistical models called Topic Model quickly became popular in the fields of text analysis[1] and text classification and some text-related tasks. As the name topic modeling suggests, it finds the patterns in the text and divides it into respective topics based on some predefined models. So, it helps to take better decisions. It is different conventional text classification methods and bag of words methods that classifies the text based on the frequency of a word . It can be classified as an unsupervised approach used for identifying the topic in document and classifying the large texts[8] to one topic.A topic can be defined as a distribution over the vocabulary of documents. [15] Every topic has a unique distribution over a set of documents. We can view a topic as a group of words which have similar meaning , and each word in the topic has a weight. [16] The same word can have different weights in different topics. [17] In our context, a word from a document can be in different documents with different weights. [18] Topic models process the text without any knowledge regarding the documents. [19] The topics emerge only through text analysis. [20] We can say that a topic is a repeating pattern of terms with relatable meaning in a set of vocabulary. [21] A good topic model results in such as – “exams”, “student”, “teachers”, “library” for a topic – Education, and “guns”, “politics”, “nations” for a topic – “Military”. [22]

- *Mahesh Korlapati, Tejaswi Ravipati, Abhilash Kumar Jha, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India*
- *Kolla Bhanu Prakash, Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India. drkbp@kluniversity.in*

Topic Models are extensively used in the areas of document clustering, summarizing huge data, extracting information from unstructured[6][10] or poorly structured text . [23] Topic models are being used to select a rightproduct. [24] A user provides his/herrequirements and the topic modellingmap therequirements to the products that match the requirements of the user. [25] They are being used to classify articles of newspapers[2], product reviews[14], huge data of messages and user profiles[9]. Topic models are can also be termed as probabilistic topic models, because of their statistical ability to discover the semantic structures of text. [26] In the modern age, the amount of written material we encounter each day is too high. [27] Topic models are useful in organizing and offering insights for us regarding the large unstructured text.[28]

1.1APPLICATIONS OF TOPIC MODELS:

1.Bioinformatics: Topic modeling is a useful method and enhances researchers’ ability to interpret biological information. The topic models are not developed widely to handle biological data. [29]

2.Opinion summarization[12][13]: Opinion summarization is an important task that helps to make the right decision for a government. Corporate firms use opinion summarization to optimize their profit. [30] Topic modeling helps to summarize huge data and provide the right opinion. [31]

3.Sentiment: With the increase in the amount of data produced by users in social media, product reviews, and blogs, research on sentiment analysis[3] was also increased extensively. With better analyzing of sentiments in text, the right decisions can be taken. [32]

4. Meeting summarization[7]: By recording the entire conversation of a meeting and summarizing it by providing the result of the meeting is an effective approach. [33] Using human resources for such tasks is valuable and sometimes can be an inefficient way.[34]

1.2 TYPES OF TOPIC MODELLING:

1. Latent Dirichlet Allocation: It is a statistical model of topic modeling that forms imaginary groups and categorizes the large cluster of text to those groups. [35] Each document is made up of topics where each topic has a proportion and that each word has a weight on that topic. [36]

2. Latent Semantic Analysis or Latent Semantic Indexing: It is a widely used technique in natural language processing and distributional semantics. It is used to identify the relation between the corpus and the documents by developing concepts about those documents and the corpus. In LSA words with related meaning occur frequently in the document.
3. Non-Negative Matrix Factorization: In Non-Negative Matrix Factorization a matrix is factorized into several matrices(2 in general) where the resultant matrices and the original matrix have only positive elements. It helps to observe the obtained matrices easily. It is widely used in multivariate analysis and linear algebra.

2 LATENT DIRICHLET ALLOCATION

Topic modeling refers to the task of identifying topics that best describes a set of documents. The technique is latent because the topics emerge during the topic modeling process. And one popular topic modeling technique is Latent Dirichlet Allocation (LDA). In natural language processing, Latent Dirichlet Allocation (LDA) is a topic modeling technique that automatically discovers topics in text documents. LDA considers a document as a mix of various topics and each word belongs to one of the document's topics. This algorithm was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael Jordan in 2003. LDA imagines a fixed set of topics. Each topic represents a set of words. The goal of LDA is to map all the documents to the topics in a way, such that the words in each document are mostly related to the imaginary topics. When classifying newspaper articles[5], Story A may contain a topic with the words “catch,” “goal,” “referee,” and “won.” It’d be reasonable to assume that Story A is about Sports. Whereas Story B may return a topic with the words “forecast,” “economy,” “shares,” and “profits.” Story B is clearly about Business. LDA calculates the probability that a word belongs to a topic and processes the text. For instance, in Story B, the word “movie” would have a higher probability than the word “rated.” This makes intuitive sense because “movie” is more closely related to the topic Entertainment than the word “rated.” LDA is useful when there are a set of documents, and the goal is to discover patterns within but without knowing about the documents. LDA is used to generate topics of a document, recommendation systems, document classification, data exploration, and document summarization. Further, LDA is useful in training linear regression models with the topics and their occurrences.

3 WORKING OF LDA

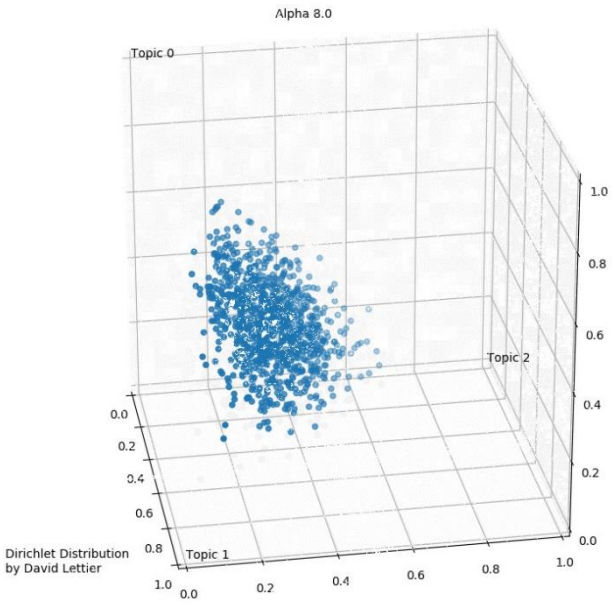

Fig. 1*(a)Graph with high Alpha value.*

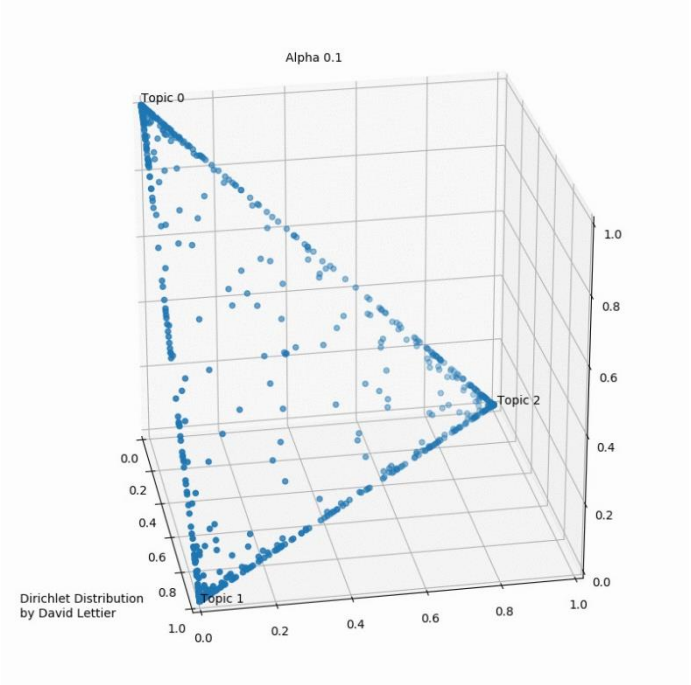

Fig. 1*(b) Graph with low alpha value*

In the Dirichlet distribution, there are two constants alpha and beta. The LDA takes a value for alpha for each imaginary topic. In the above image, we decided to divide the corpus into three topics and every topic is given the same alpha value for convenience purposes. Each point represents the proportion of the mixture of the selected topics(3 in our example). As we have selected 3 topics, the vectors look like (0, 0.4, 0.6) or (0.7, 0.3, 0). The sum of the values in each vector must be one. When the value of alpha is low, the topic distribution samples are near the topics. If the values of vector are either (1,0,0) or (0,1,0) or (0,0,1) we can infer that vectors are sampled. In such a case, a document is made up of only one topic and the rest of the topics have zero proportion in that document. If the value of

alpha is one, multiple conclusions can be drawn from the possible distributions. In such a case the distribution might either support a single topic or an even proportion of all the three topics or a random proportion of the three topics. If the value of alpha is greater than one, the vectors start to converge in the center of the triangle as the selected topics were 3. From the image(a) we can infer that as alpha gets bigger, the samples are more likely to have a uniform mixture of all the topics. The above image depicts the distribution of samples of topic mixtures for the documents. We used three topics as it can be represented in three dimensions. In general, it is better to use several numbers of topics, depending on the number of research papers we consider. There are two components in an LDA tool, an alpha value, and a beta slider. These two parameters are required by LDA to perform the process of document classification. The alpha value is used to control the proportion of topics in a particular document. From the image(b) we can infer that, if the value of alpha decreases, the documents are likely to be made up of a single topic rather than a mixture. If the value increases, the documents will have a uniform proportion of all the topics. The beta parameter is used to control the presence of words in a topic. If its value is decreased the topics will have fewer words. If the value of beta is increased the topics will have more words. To get accurate results, the number of topics must be less. By considering the above statements, alpha and beta are usually set below one. In LDA, there are two terms phi and theta. Phi is typically the relation between words and topics, and theta is the relation between a document and the number of topics.

4 IMPLEMENTATION

4.1 Loading Required Packages

We use R Studio to generate a graph that depicts the distribution of topics over the documents. We are going to use add-on packages present in R such as ggplot2, reshape2, and lda. lda package implements latent Dirichlet allocation (LDA)and other related models such as sLDA, corrLDA, and the mixed-membership stochastic block model. The reasoning for all these models is provided by a quick and efficient Collapsed Gibbs sampler written in C. R studio is equipped with functions for reading and writing data that are used in topic models. In the lda package, the tools required for examining posterior distributions are also included. ggplot() works by creating a ggplot object. It is used to visualize data by declaring the input data frame. ggplot stands for the grammar of graphics. It breaks a graph into different parts such as layers, scales, and other semantics. qplot is a function in ggplot2 package. It is analogous to the plot but is used extensively to plot the given data quickly.

4.2 Dataset

The Cora data contains research papers and records of machine learning that have been manually clustered into groups. All these research papers belong to the same publication. The Cora dataset has 2410 scientific publications where every document falls into one of seven classes. Every publication in the dataset has a binary word vector(i.e. it takes values of either 0 or 1). 0 indicates absence whereas 1 indicates the presence of the corresponding word from the class.

4.3Selection of Clusters and Topics

We decided to select 9 documents randomly from the Cora dataset and divide the vocabulary in those documents into 10 topics or categories.

4.4LDA.COLLAPSED.GIBBS.SAMPLER

These functions use a collapsed Gibbs sampler[4] to fit three different models that are latent Dirichlet allocation, the mixed-Membership Stochastic block Model and supervised LDA (sLDA). The poorly represented documents are taken as inputs. It performs inference and returns estimates of topics using the last iteration of Gibbs Sampling. The inputs that we provide to the function are the number of topic clusters, documents that we selected from the Cora dataset, the number of iterations that needed to be performed.

4.5 TOP.TOPIC.WORDS

This function is used to get top words and documents in each topic. It takes a model fitted by the collapsed Gibbs sampler function and returns a matrix of the top words in each topic. For this function, a 2-dimensional matrix is selected. In this matrix, each entry is proportional to the probability of seeing the word of a topic. The column names should match the words in the vocabulary. The output of collapsed Gibbs sampler function is used in top.topic.words.

5 RESULTS

topic

- markov.distribution.state.convergence.algorithm
- feature.genetic.features.data.classification
- research.grant.university.report.technical
- genetic.search.fitness.population.evolutionary
- learning.knowledge.rules.queries.revision
- neural.network.system.robot.dynamics
- decision.trees.tree.classification.error
- model.bayesian.models.network.recurrent
- learning.reinforcement.network.system.control
- system.design.reasoning.knowledge.planning

Fig. 2*(a) Imaginary Topics*

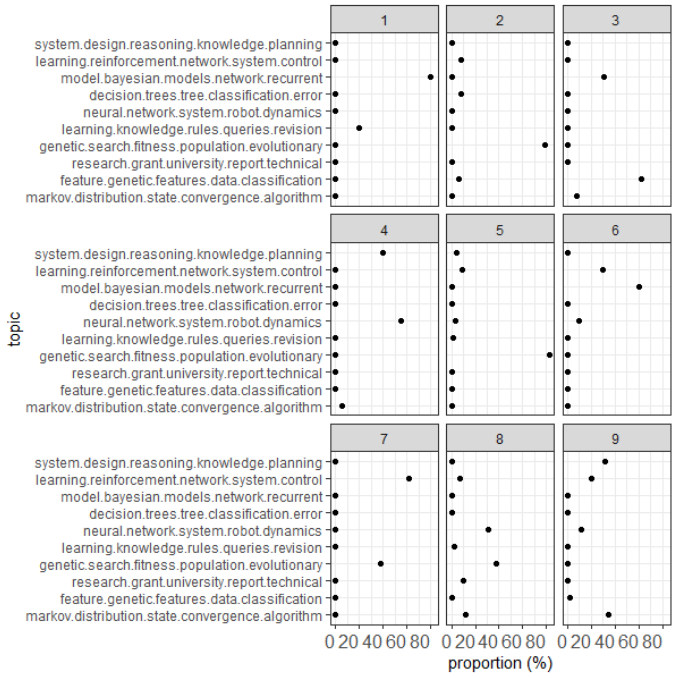

Fig. 2(b)*Distribution of ten topics over 9 scientific documents*

In this experiment, we used the Cora dataset. It is a collection of 2410 scientific documents. R Studio has an add-on package that enables the user to perform text analysis on the Cora dataset. Nine documents are selected randomly from the Cora dataset. Based on the vocabulary or words in the documents, we decided to divide the document into ten topics. So, now a document can be viewed as the sum of the distribution of the ten topics. Here, a topic can be termed as a distribution over a fixed vocabulary of words. A document is made up of different topics with varying proportions. A topic is an imaginary cluster and each topic is made of words with similar meaning where each word has a weight within that topic. By using the latent Dirichlet allocation on the fixed vocabulary, we have divided the vocabulary into 10 latent topics that remain constant. In LDA, each latent topic follows a Dirichlet distribution over the corpus and each document is represented as a mixture of these topics. The goal of LDA is to infer the underlying topics, topic proportions and topic assignments for every document. From the first research paper, we can infer that the prime focus is on Bayesian belief networks and recurrent models. The efficacy of this model can be checked by viewing the document which is used from the Cora dataset. We can analyze it manually and conclude the topic of the research paper. The result can then be compared with the LDA topic modeling result and hence the accuracy of the model can be proved.

6 CONCLUSION

Topic models can be used in document summarization, document classification, information retrieval, and collaborative filtering. Topic models are not limited to the textual pattern, they can also elaborate visual data. A document is made up of topics and words. In a similar way, images are made up of image features and pixels.We can achieve the task of textual classification by methods like k-means clustering or classification techniques like naïve Bayes or KNN. But as we observed the techniques of topic modeling are easier and can do the tasks of organizing, summarizing and searching more efficiently than other techniques. Despite there are several models available in topic modeling, Latent Dirichlet Allocation is the simplest topic model and can perform the tasks with high efficiency.

REFERENCES

[1] Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1348–1353,Seattle, Washington, USA, 18-21 October g2013.c©2013 Association for Computational Linguistics.

[2] Carina Jacobi, Wouter van Atteveldt& Kasper Welbers (2016) Quantitative analysis of large amounts of journalistic texts using topic modelling, Digital Journalism, 4:1, 89-106, DOI: 10.1080/21670811.2015.1093271.

[3] Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect Sentiment Analysis with Topic Models. 2011 IEEE 11th International Conference on Data Mining Workshops. doi:10.1109/icdmw.2011.125.

[4] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08. doi:10.1145/1401890.1401960.

[5] De Smet, W., &Moens, M.-F. (2009). Cross-language linking of news stories on the web using interlingual topic modelling. Proceeding of the 2nd ACM Workshop on Social Web Search and Mining - SWSM '09. doi:10.1145/1651437.1651447.

[6] Uys, J. W., du Preez, N. D., &Uys, E. W. (2008). Leveraging unstructured information using topic modelling. PICMET '08 - 2008 Portland International Conference on Management of Engineering & Technology. doi:10.1109/picmet.2008.4599703.

[7] David Newman , Jey Han Lau , Karl Grieser , Timothy Baldwin, Automatic evaluation of topic coherence, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, p.100-108, June 02-04, 2010, Los Angeles, California.

[8] Lan Du , Wray Buntine , HuidongJin, Modelling sequential text with an adaptive topic model, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, July 12-14, 2012, Jeju Island, Korea.

[9] Dong-mei Yang, Hui Zheng, Ji-kun Yan, & Ye Jin. (2012). Persona analysis with text topic modelling. International Conference on Automatic Control and Artificial Intelligence (ACAI 2012). doi:10.1049/cp.2012.1280.

[10] Eickhoff, Matthias and Neuss, Nicole, (2017). "TOPIC MODELLING METHODOLOGY: ITS USE IN INFORMATION SYSTEMS AND OTHER MANAGERIAL DISCIPLINES". In Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal, June 5-10, 2017 (pp. 1327-1347). ISBN 978-989-20-7655-3 Research Papers.

[11] P. D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proceedings of the Association for Computational Linguistics, pp. 417–424, 2002.
[12] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up? Sentiment Classification Using Machine Learning Techniques,” Proceedings of EMNLP, pp. 79–86, 2002.
[13] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177, 200.
[14] Kavuri, M. & Prakash, K.B. 2019, "Performance comparison of detection, recognition and tracking rates of the different algorithms", International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 153-158.
[15] Kolla, B.P., Dorairangaswamy, M.A. & Rajaraman, A. 2010, "A neuron model for documents containing multilingual Indian texts", 2010 International Conference on Computer and Communication Technology, ICCCT-2010, pp. 451.
[16] Kolla, B.P. & Raman, A.R. 2019, Data Engineered Content Extraction Studies for Indian Web Pages, Advances in Intelligent Systems and Computing, 711, pp. 505-512.
[17] Naga Pawan, Y.V.R. & Prakash, K.B. 2019, "Variants of particle swarm optimization and onus of acceleration coefficients", International Journal of Engineering and Advanced Technology, vol. 8, no. 5, pp. 1527-1538.
[18] Pradeep Kumar, V. & Prakash, K.B. 2019, "QoS aware resource provisioning in federated cloud and analyzing maximum resource utilization in agent based model", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 8, pp. 2689-2697.
[19] Prakash, K.B. 2018, "Information extraction in current Indian web documents", International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 68-71.
[20] Prakash, K.B. 2017, "Content extraction studies using total distance algorithm", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 673.
[21] Prakash, K.B. 2015, "Mining issues in traditional indian web documents", Indian Journal of Science and Technology, vol. 8, no. 32, pp. 1-11.
[22] Prakash, K.B., Ananthan, T.V. & Rajavarman, V.N. 2014, "Neural network framework for multilingual web documents", Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014, pp. 392.
[23] Prakash, K.B. & Dorai Rangaswamy, M.A. 2019, Content extraction studies for multilingual unstructured web documents, Advances in Intelligent Systems and Computing, 749, pp. 653-664.
[24] Prakash, K.B. & Dorai Rangaswamy, M.A. 2016, "Content extraction studies using neural network and attribute generation", Indian Journal of Science and Technology, vol. 9, no. 22, pp. 1-10.
[25] Prakash, K.B., Dorai Rangaswamy, M.A. & Ananthan, T.V. 2014, "Feature extraction studies in a heterogeneous web world", International Journal of Applied Engineering Research, vol. 9, no. 22, pp. 16571-16579.
[26] Prakash, K.B., Dorai Rangaswamy, M.A., Ananthan, T.V. & Rajavarman, V.N. 2015, "Information extraction in unstructured multilingual web documents", Indian Journal of Science and Technology, vol. 8, no. 16.
[27] Prakash, K.B., Dorai Rangaswamy, M.A. & Raman, A.R. 2010, "Text studies towards multi-lingual content mining for web communication", Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, pp. 28.
[28] Prakash, K.B., Kumar, K.S. & Rao, S.U.M. 2017, "Content extraction issues in online web education", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 680.
[29] Prakash, K.B. & Rajaraman, A. 2016, "Mining of Bilingual Indian Web Documents", Procedia Computer Science, 89, pp. 514-520.
[30] Prakash, K.B., Rajaraman, A. & Lakshmi, M. 2017, "Complexities in developing multilingual on-line courses in the Indian context", Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017, pp. 339.
[31] Prakash, K.B., Rajaraman, A., Perumal, T. & Kolla, P. 2016, "Foundations to frontiers of big data analytics", Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016, pp. 242.
[32] Prakash, K.B. & Rangaswamy, M.A.D. 2016, "Content extraction of biological datasets using soft computing techniques", Journal of Medical Imaging and Health Informatics, vol. 6, no. 4, pp. 932-936.
[33] Prakash, K.B., Rangaswamy, M.A.D. & Raja Raman, A. 2012, ANN for multi-lingual regional web communication, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7667 LNCS(PART 5), pp. 473-478.
[34] Prakash, K.B., Rangaswamy, M.A.D. & Raman, A.R. 2013, "Attribute based content mining for regional web documents", IET Seminar Digest, pp. 368.
[35] Prakash, K.B., Rangaswamy, M.A.D. & Raman, A.R. 2012, Statistical interpretation for mining hybrid regional web documents, Communications in Computer and Information Science, 292 CCIS, pp. 503-512.
[36] Ismail, M., Prakash, K.B. & Rao, M.N. 2018, "Collaborative filtering-based recommendation of online social voting", International Journal of Engineering and Technology(UAE), vol. 7, no. 3, pp. 1504-1507.