RESEARCH ARTICLE OPEN ACCESS

Labeling of Topics Generated By Topic Modeling Algorithms – A Study

Ms Anjana C Rajan, Dr R Kalaiselvi
Department of Computer Applications
Noorul Islam University
Kumaracoil – TamiNadu
India

ABSTRACT

Text mining also known as text analytics is an emerging area in Artificial Intelligence that uses techniques of natural language processing to transform unstructured text into a structured format for better understanding. It is emerging as an important area in almost all fields as it helps the user to extract information from huge corpus of text or unstructured data. One of the main tools used in text mining is topic modeling Topic modelling helps to unearth hidden topics which can be said to be recurring patterns of co-occurring words. Its goal is to find the latent topics from large volumes of unstructured data. But the topics discovered may not present to the user a very coherent picture of the document as it is a list of the top n words in a topic. So in order to make the understanding of the topics better, automatic labelling of the topics has been explored. This paper attempts to explore the work done for labeling the topics generated by topic modelling algorithms

Keywords:- Text mining, topic modelling, LDA, latent topics, labels, concepts, candidate generation, candidate ranking

I. INTRODUCTION

Topic modeling is one of the most popular techniques to unearth the hidden or latent topics from a large corpus of text documents. The generic method to represent the discovered topics is to represent it using the top ‘ n ’ number of terms with the highest marginal probabilities. But in some cases the top words returned by topic modeling algorithms may not give the user a completely coherent picture of the topic it is trying to represent.So it would be very useful if a label representing the top topic words can be used to label the topics returned. These labels will be very useful as it will help the user to get an idea about what are the major topics in a given corpus of data rather than looking up the top words in each topic list. Another issue with not labeling topics is that each person may interpret the top n topic words in different perspectives but if we can label the topics it will be more useful for comprehending the topics in a universal manner.

Most of the research work conducted to automatically label topics has divided the task into two major areas – candidate label generation and candidate label ranking. The candidate label generation area focuses on generating candidate words for representing the labels and the candidate ranking area focuses on ranking the generated candidate words based on some semantic ranking method. It helps to identify the top or most highly suitable label word.

II. LABELING OF TOPICS

In this paper we have attempted to study all the previous work that has been done in order to do topic labeling. Jeyhan Lau et al [1] proposed a method for automatically labeling topics generated by LDA by generating candidate label set by combining top ranking topic terms, Wikipedia title containing the top topic terms and sub phrases extracted from Wikipedia articles. The ranking method they use is a mixture of association measures and lexical features, fed into a supervised ranking model.

Qiaozhu Mei et al [2] have explored a probabilistic approach in automatically labeling topics. In the paper they propose probabilistic approaches for labeling multinomial topic models in an objective way. They have approached the labeling problem as an optimization problem involving minimizing Kullback-Leibler divergence that occurs between word distributions and which involves maximizing mutual information between a label and a topic model.The results generated by them are effective to generate labels which are meaningful and can be used for interpreting the discovered topic models. This method can also be used on different types of topic models such as LDA, PLSA and their variations.

Mehdi Allahyari et al [3] have created a topic model that combines ontological concepts with topic models in single framework where topics and concepts are represented as a multinomial distribution over concepts and over words. In selecting the most apt topic labels they depended on the semantic relatedness of the concepts and their ontological classifications. Their method has (1) onto/LDA, an ontology-dependent topic model which incorporates an ontology into the topic model. (2) A topic labeling method which is based on the semantics of the concepts in the discovered topics, and it also explores the ontological relationships that exist among the concepts in the ontology. They have shown that their model improves the accuracy of labeling by making use of the topic-concept relations and it can automatically generate labels that will be meaningful to the users while interpreting the topics

.

Xian-Ling Mao et al [4] has proposed to overcome the challenge of applying topic models to other knowledge management problem as it is difficult to accurately interpret the meaning of each topic. . In their paper they propose two algorithms that automatically allocate precise labels to each topic in a hierarchy by making use of the sibling and parent-child relationship that exist among topics. They have shown that by using their method the inter-topic relation is very useful in enhancing the topic labeling accuracy and they are capable of generating meaningful topic labels that will be useful for interpreting the hierarchical topics.

Davide Magatti et al [5] proposed an algorithm which makes use of a hierarchy for the automatic labeling of topics. Its main measures are a set of similarity measures and a set of topic labeling rules. These labeling rules are designed to extract the most agreed labels between the given topic and the hierarchy. The Google Directory service together with topics extracted via an ad-hoc developed software procedure and expanded through the use of the Open Office English Thesaurus is used in order to build the hierarchy.

Xiaojun Wan et al [6] proposed to make use of text summaries for topic labeling. Sentences are extracted from the most related documents to form the summary for each topic. The work carried out was based on sub modular optimization in order to extract high relevance summaries. They have demonstrated that the summaries extracted by them are better than the ones discovered by existing summarization methods and that summaries are better at labeling topics that words or phrases.

A method has been put forward by Lau et al. [7]in which they use phrases as topic labels and then use a supervised learning techniques for ranking candidate phrases which will be the most suitable label. The candidate labels are the top-5 topic terms and some noun from related Wikipedia articles.

Two algorithms have been used by Mao et al. [8] for automatically assigning labels to each topic in a hierarchy by making use of the sibling and parent-child relations among topics. Another novel method was proposed by Kou et al. [9] where they used a method to relate topics and phrases to word vectors and letter trigram vectors in order to discover which label is semantically more similar to the topics generated. Hulpus et al. [10] proposed a new approach which uses graph centrality measures to topic labeling by making use of data from DBpedia. A method has also been used to make use of images for representing topics. Here instead of labels candidate images are extracted and they are used to label the topics [11].Candidate images are selected by using a graph-based algorithm

III. CONCLUSION

Topic modelling is widely used and is emerging as one of the prominent areas where a lot of research activity is taking place as it of utmost importance in the context of knowledge discovery. But a common problem is the understandability of the topics generated by the topic modeling algorithms. This paper has attempted to provide an overview of some of the research work done in the area of labeling the topics so that it is easier for the user to understand the latent topics from a large corpus of text.

REFERENCES

[1] J.H. Lau, K. Grieser, D. Newman, and T. Baldwin. 2011. Automatic labelling of topic models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies Volume 1, pages 1536–1545. Association for Computational Linguistics.

[2] Mei, X. Shen, and C.X. Zhai. 2007. Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 490–499. ACM.

[3] Mehdi Allahyari, Seyedamin Pouriyeh, Krys Kochut, and Hamid R Arabnia. 2017b. A knowledge based topic modeling approach for automatic topic labeling. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 8(9):335–349.

[4] Xian-Li Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. 2012. Automatic labeling hierarchical topics. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM ’12), Sheraton, Maui Hawai.

[5] D. Magatti, S. Calegari, D. Ciucci, and F. Stella. 2009. Automatic labeling of topics. In Intelligent Systems Design and Applications, 2009. ISDA’09. Ninth International Conference on, pages 1227–1232. IEEE.

[6] Xiaojun Wan, Manifold-Ranking Based Topic-Focused Multi-Document Summarization,2007, International Joint Conferences on Artificial Intelligence Organization.

[7] J.H. Lau, D. Newman, S. Karimi, and T. Baldwin. 2010. Best topic word selection for topic labelling. In Proceedings of the 23rd International Conference on

Computational Linguistics: Posters, pp 605–613. Association for Computational Linguistics.

[8] Xianling Mao, Yi-Jing Hao, Qiang Zhou, Wenqing Yuan, Liner Yang, Heyan Huang: A Novel Fast Framework for Topic Labeling Based on Similarity-preserved Hashing. COLING 2016: 3339–3348

[9] Wanqiu Kou, Li Fang, and Timothy Baldwin. 2015. Automatic labelling of topic models using word vectors and letter trigram vectors. In Proceedings of the 11th Asian Information Retrieval Societies Conference (AIRS 2015), pages 229–240, Brisbane, Australia

[10] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. An eigenvalue based measure for word-sense disambiguation. In FLAIRS 2012, 2012.

[11] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. In Proceedings of the sixth ACM international conference on Web search and data mining, pp 465–474. ACM

[10] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. In Proceedings of the sixth ACM international conference on Web search and data mining, pages 465–474. ACM

[11] Nikolaos Aletras and Mark Stevenson. 2013. Representing topics using images. In Proceedings of NAACL-HLT, pages 158–167.

[12] Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. 2003. An analytical comparison of approaches to personalizing PageRank. Technical Report 2003–35, Stanford InfoLab.

[13] Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Shafiq Joty. 2013. Towards topic labeling with phrase entailment and aggregation. In Proceedings of NAACL-HLT, pages 179–189.

[14] Ehud Reiter and Robert Dale. 2000. Building natural language generation systems.

[15] S. Bhatia, J. H. Lau, and T. Baldwin, “Automatic labelling of topics with neural embeddings,” in 26th COLING International Conference on Computational Linguistics, 2016, pp. 953–963.

[16] Wanqiu Kou, Li Fang, and Timothy Baldwin. 2015. Automatic labelling of topic models using word vectors and letter trigram vectors. In Proceedings of the 11th Asian Information Retrieval Societies Conference (AIRS 2015), pages 229–240, Brisbane, Australia.

[17] A.Herzog, P. John, S. J. Mikhaylov. Transfer Topic Labeling with Domain-Specific Knowledge Base: An Analysis of UK House of Commons Speeches 1935–2014.

[18] A. Smith, T. Y. Lee, F. Poursabzi-Sangdeh, J. Boyd-Graber, N. Elmqvist, L. Findlater. Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic. In ACL, 2017

[19] D. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with Pachinko allocation. In Proc. 24th International Conference on Machine Learning (ICML), Corvallis, OR, 2007