International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June -2018 www.irjet.net p-ISSN: 2395-0072

A Review on Topic Detection and Term-Term Relation Analysis in Big Data

K Swanthana[1], K Swapnika[2]

[1,2] *Assistant professor, Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Telangana, India*

Abstract - *Topic models provide a way to aggregate vocabulary from a document corpus to form latent topics. In particular, Latent Dirichlet Allocation (LDA) is one of the most popular topic modeling approaches. Learning the meaningful topic models with massive document collections which contain millions of documents and billions of tokens is challenging. Topic detection is a tool to detect topics from media attracts much attention. Generally, a topic is characterized by a set of informative keywords/terms. Traditional approaches are usually based on various topic models, such as Latent Dirichlet Allocation (LDA). They cluster terms into a topic by mining semantic relations between terms. However, they neglect the co-occurrence relations found across the document, which leads to the detection of incomplete information. Furthermore, the inability to discover latent co-occurrence relations via the context or other bridge terms will prevent the important but rare topics being detected. To solve this problem we integrate semantic relations and implicit co-occurrence relations for topic detection. Specifically, the approach combines multiple relations into a term graph and detects topics from the graph using a graph analytical method. By combing mutually complementary relations it cannot detect the topics more effectively. This approach can mine important rare topics by leveraging latent co-occurrence relations in huge corpus.*

***Key Words*: Topic model, LDA, Semantic, Co-occurrence, Latent Co-occurrence, Corpus**

1. INTRODUCTION

Due to the rapid development of internet, the amount of online text is experiencing explosive growth by means of online news media and social media. The rich information within the text data can be utilized to reveal some meaningful trends/topics or the evolution of certain social phenomena, such as the presidential election of the USA. Besides, it can also be exploited for detecting some emergency events or natural disasters, such as 2014 Shanghai Stampede1 and 2014 Kangding earthquake.2 In general, an event is considered as something non-trivial happening at a specific date/time and in a specific location [1]. A topic can be considered as a kind of "abstract" event which consists of some "concrete" events with semantic relatedness, here "events" and "topics" are inter-changeable. Topic detection is a sub-task of Topic Detection and Tracking (TDT) [2]. It aims at detecting topics or trends from various text corpuses such as online media. Topic detection as a fundamental problem of information retrieval can help the decision makers to efficiently detect meaningful topics. Therefore, it has attracted much attention such as public opinion monitoring, decision supporting and emergency management.

Automatic identification of semantic content of documents has become increasingly important due to its effectiveness in many tasks, including information retrieval, information filtering and organization of documents collections in digital libraries. The identification of the topic(s) that a document addresses increases our understanding of that document, the characteristics of the collection as a whole and the interplay between distinct topics. In collections where the temporal ordering of documents is not of importance, studying a snapshot of the collection at any given time is sufficient to deduct as much information as possible about the various topics of interest in the collection. On the other hand, many document collections exhibit temporal relationships that are often times utilized to aid the topic discovery process. The analysis of the temporal dimension of collections has become an important field of study in many applications, including weblog topic mining [5], evolution of author and paper networks [4], news event analysis [6] and social interaction of researchers [7,3].

In topic detection, a topic is generally represented as a set of descriptive and collocated terms. Firstly, document clustering techniques are applied in topic detection to cluster content-similar documents and extract keywords from clustered document sets as the representation of the topics. Currently, most of the approaches use various topic models, a type of generative probabilistic model such as LDA [8], pLSA [9] and their extensions, to detect topics. Among them, LDA has been proved to be a powerful algorithm because of its ability on mining the semantic information from the text data. Terms having semantic relations with each other are collected as a topic.

Latent Dirichlet Allocation (LDA) [10] has been shown to be a highly effective unsupervised learning methodology for finding distinct topics in document collections. LDA is a generative process that models each document as a mixture of topics where each topic corresponds to a multinomial distribution over words. The learned document-topic and topic-word distributions are then used to identify the best topics for the documents and the most descriptive words for each topic. However, the original formulation of LDA focuses

on analyzing a snapshot of collections and views the collection as being generated at a single point in time.

However, there are two challenges to existing topic model based approaches. Firstly, as claimed by Sayyadi and Raschid [1], "Current topic modeling methods do not explicitly consider word co-occurrences". "Co-occurrence" means two terms co-occur in the same document. Unfortunately, due to the fact that "Extending topic modeling to include co-occurrence can be a computationally daunting challenge" [1], their proposed graph analytical approach only made an approximation to this extension: they merely took into account co-occurrence information alone while ignoring semantic information. Therefore, how to combine semantic relations and co-occurrence relations to complement each other remains to be a challenge. Secondly, existing approaches usually focus on detecting prominent or distinct topics by mining explicit semantic relations or frequent co-occurrence relations. However, they neglect to uncover latent co-occurrence relations. The inability to uncover latent relations prevents the important but rare topics, which are hidden in large scale and noisy data collections, from being detected. Such important rare topics have two attributes: significant for human decision making but rare that cannot be discovered easily [11]. In other words, their features are commonly implicit or latent. Here gives some examples: the latent omens such as the abnormal behaviors of some animals may reveal that the disasters such as earthquake will occur soon; the early incubations of the disease may trigger the subsequent cancer. The reason why such topics are commonly ignored is: they differ from the distinct topics indicating common patterns; besides, they are not outliers or noises in the sense of anomaly detection [11].

2. RELATED WORK

Numerous statistical approaches for modeling text documents have been proposed to model text documents. Probabilistic latent semantic analysis (pLSA) [12] models the generation of each document through activating multiple topics over words. This model improves upon the singular value decomposition based latent semantic analysis (LSA) [13] which cannot handle polysemy. pLSA, on the other hand, uses a distribution indexed by the training documents, leading to the fact that the number of parameters to be estimated grows linearly with the number of training documents in the collection. Thus, practical applications with large training documents are susceptible to over fitting with this model. Latent Dirichlet Allocation (LDA) [10] is another generative model that has become popular in recent years that overcomes the over fitting problem of pLSA by using a Dirichlet for modeling the distribution of topics for each document.

To uncover latent co-occurrence relations and discover important rare topics Chance Discovery (CD) theory and Idea Discovery (ID) theory are used. These are defined as a rare but important event or situation which has a strong impact on human decision making [11,14]. CD and ID as extensions of Knowledge Discovery have been used to detect chances by uncovering latent co-occurrence relations among terms. In the ID process, text data is analyzed and converted into a term graph by mining co-occurrence relations, where latent co-occurrence relations are uncovered and visualized to capture chances. Latent co-occurrence relations means there may be no frequent co-occurrence relations between two terms (the terms do not frequently co-occur in the same documents); but, the two terms can be implicitly related/linked by considering the "context" of one of the terms or other bridge terms. Here "context" denotes neighbors of the term, which are strongly interconnected in the form of a community (cluster). In other words, latent co-occurrence relations between two terms cannot be measured in an isolated term-term view; the context of the term should be taken into account. To address these challenges, We use an integrated approach to integrate semantic information and co-occurrence information among terms for topic detection. Specifically, the approach fuses multiple types of relations into a uniform term graph by incorporating ID theory with topic modeling method.

Firstly, an Idea Discovery algorithm called Idea Graph is adopted to mine co-occurrence relations (especially latent co-occurrence relations) for converting the corpus into a term graph. Then, a semantic relations extraction approach is proposed based on LDA to enrich the graph with semantic information. Lastly, a graph analytical method is presented to exploit the graph for detecting topics.

To the best of our knowledge, the coupling of ID and topic model for topic detection has not been seen as follows:
(1) It can detect topics more effectively to support human decision making by combing mutually complementary relations: semantic relations and co-occurrence relations.
(2) It can mine important rare topics by leveraging latent co-occurrence relations, which may aid human to perceive the topics with great significance.

3. TOPIC MODELING APPROACHES

Topic Detection and Tracking (TDT) is an integral part of the Translingual Information Detection, Extraction, and Summarization (TIDES) program [2]. TDT mainly contains two sub-task: Topic Detection and Topic Tracking. Topic detection (Event detection) aims at detecting novel topics/events from text corpus while topic tracking is dedicated to tracking the evolution of existing topics over temporal dimension. Topic detection has attracted much attention in machine learning, information retrieval and social media modeling [1,11,15]. Specifically, topic detection can be classified into two types: New Event Detection (NED) and Retrospective Event Detection (RED).

NED aims at detecting newly encountered topics/events from online text streams [15]. Rill et al. [16] proposed a system to detect emerging political topics in Twitter. The detected topics can be used to extend existing knowledge

bases for better concept-level sentiment analysis. Hou et al. [17] proposed a multifaceted news analysis approach to detect events from online news. They represented news as a link-centric heterogeneous network and formalized news analysis and mining task as link discovery problem. Based on that, they presented a unified probabilistic model for topic extraction and inner relationship discovery within events.

RED is dedicated to discovering the events from the historical corpus in an offline way [18]. Yang et al. [18] proposed an agglomerative clustering algorithm, named augmented Group Average Clustering, to cluster articles into events. They also employed an iterative bucketing and re-clustering model proposed by Cutting et al. [19] to control the tradeoff between cluster quality and computational efficiency. Zeng and Zhang [20] presented a variable space Markov model for topic detection, where several steps based on space computation and a hierarchal clustering algorithm are proposed to tackle the issues of document imbalance and topic transition. We note our paper only addresses RED problem.

3.1 Probabilistic approach

As mentioned previously, some RED approaches employ topic modeling, a type of the probabilistic modeling, to detect topics from the text data. A famous topic model named Latent Dirichlet Allocation (LDA) [8] is a three-level hierarchical Bayesian model. Each document is represented as a finite mixture over an underlying set of latent topics, where each topic is characterized by a distribution over terms. Terms having strong semantic relations with each other are clustered as a topic's features or representation. There are plenty of improved versions of LDA. For example, several knowledge-based topic models have been proposed to incorporate prior domain knowledge from the user to generate coherent topics [21–26]. Among which, Chenand Liu [22] proposed the AMC algorithm, topic modeling with automatically generated Must-links and Cannot-links, to incorporate the knowledge automatically mined from the past learning/modeling results, which can help future learning. Xu et al. [26] used a knowledge based topic model to extract implicit features of product reviews for opinion mining task.3

3.3 Graph analytical approach

Other approaches on topic detection leverage the graph analytical method to detect the topics within the graph or network. Sayyadi and Raschid [1] proposed a graph analytical approach for topic detection. They used a Key Graph algorithm [27] to convert text data into a term graph based on co-occurrence relations between terms. Then they employed a community detection approach to partition the graph. Eventually, each community is regarded as a topic and terms within the community are considered as the topic's features.

4. CONCLUSIONS

In this paper, we explored the solution for the problem of ignoring the latent i.e implicit meaning of the terms which co-occur in the documents by integrating semantic relations with implicit or frequent co-occurrence relations and latent co-occurrence relations for important topics detection. In this approach topic modeling is incorporated with Chance Discovery (CD) to capture the semantic relations and co-occurrence relations, which facilitates effective topic detection. Idea Discovery (ID) is to convert the analyzed data into a term graph by mining co-occurrence relations. In future there is a need and extension for this approach for new event detection (NED) to satisfy the need of Big Data era. Finally we can extend this approach to represent the topic hierarchy, where the events as the basic elements of the topic to be detected.

REFERENCES

[1] H. Sayyadi, L. Raschid , A graph analytical approach for topic detection, ACMTrans. Internet Technol 13 (2) (2013) 4, 23.

[2] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study, in: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194–218.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 44–54, New York, NY, USA, 2006. ACM Press.

[4] K. Brner, J. T. Maru, and R. L. Goldstone. The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences, 101 Suppl 1:5266–5273, April 2004.

[5] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In WWW '06: Proceedings of the 15th international conference on World Wide Web, pages 533–542, New York, NY, USA, 2006. ACM Press.

[6] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 198–207, New York, NY, USA, 2005. ACM Press.

[7] D. Zhou, X. Ji, H. Zha, and C. L. Giles. Topic evolution and social interactions: how authors effect research. In CIKM '06: Proceedings of the 15th ACM

international conference on Information and knowledge management, pages 248–257, New York, NY, USA, 2006. ACM Press.

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

[9] T. Hofmann, Probabilistic latent semantic analysis, UAI (1999) 289–296.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[11] Y. Ohsawa, Chance discoveries for making decisions in complex real world, New Gen. Comput. 20 (2) (2002) 143–163.

[12] T. Hofmann. Probabilistic latent semantic indexing. In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and 25 development in information retrieval, pages 50–57, New York, NY, USA, 1999. ACM Press.

[13] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.

[14] H.Wang, Y. Ohsawa, Y. Nishihara, Innovation support system for creative product design based on chance discovery, Expert Syst. Appl. 39 (5) (2012) 4890–4897.

[15] J. Allan, R. Papka, V. Lvrenko , On-line new event detection and tracking, SIGIR (1998) 37–45.

[16] S Rill, D Reinel, J. Scheidt, et al., Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. 69 (2014) 24–33.

[17] L. Hou, J. Li, Z. Wang, et al., NewsMiner: Multifaceted news analysis for event search, Knowl. Based Syst. 76 (2015) 17–29.

[18] Y. Yang, T. Pierce, J.G. Carbonell, A study on retrospective and on-line event detection, SIGIR (1998) 28–36.

[19] D.R. Cutting, D.R. Karger, J.O. Pedersen, 803 J.W. Tukey, Scatter/gather: A cluster-based approach to browsing large document collections, SIGIR (1992) 318–329.

[20] J. Zeng, S. Zhang, Variable space hidden Markov model for topic detection and analysis, Knowl. Based Syst. 20 (7) (2007) 607–613.

[21] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, ICML (2009) 25–32.

[22] Z. Chen, B. Liu,Mining topics in documents: standing on the shoulders of big data, KDD (2014) 1116–1125.

[23] Z. Chen, A.Mukherjee, B. Liu,M. Hsu,M. Castellanos, R. Ghosh, Exploiting domain knowledge in aspect extraction, EMNLP (2013) 1655–1667.

[24] A. Mukherjee, B. Liu, Aspect extraction through semi-supervised modeling, ACL (2012) 339–348.

[25] X. Fu, K. Yang, J.Z. Huang, L. Cui, Dynamic non-parametric joint sentiment topic mixture model, Knowl. Based Syst. 82 (2015) 102–114.

[26] H. Xu, F. Zhang, W. Wang, Implicit feature identification in Chinese reviews using explicit topic mining model, Knowl. Based Syst. 76 (2015) 166–175.

[27] Y. Ohsawa, N.E. Benson, Y. Masahiko, KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor, ADL (1998) 12–18.