International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 19
August 2018

Using Hashtag Graph without Co-Occurrence in Microblogs and Connect Semantically-Related Words

Katuri Lokesh[1], M.Mallikarjuna Rao[2], Sk.Khaleelullah[3]

[1]student, M.Tech (Cse), Lingayas Institute Of Management & Technology, A.P., India.

[2]assistant Professor, Dept. Of Computer Science & Engineering, Lingayas Institute Of Management & Technology, A.P., India.

[3]assistant Professor & Hod, Dept. Of Computer Science & Engineering, Lingayas Institute Of Management & Technology, A.P., India.

Abstract *—With billions of active users, Twitter is popular because of its massive spreading of instant messages (i.e. tweets), bursts of world news, entertainment gossip about celebrities, and discussions over recently released products are all spreading on Twitter vividly. The shortness and informality of tweets leads to extreme sparse vector representations with a large vocabulary. This makes the conventional topic models (e.g., Latent Dirichlet Allocation and Latent Semantic Analysis) fail to learn high quality topic structures. Tweets are always showing up with rich user-generated hashtags. The hashtags make tweets semi-structured inside and semantically related to each other. Since hashtags are utilized as keywords in tweets to mark messages or to form conversations, they provide an additional path to connect semantically related words. treating tweets as semi-structured texts, we propose a novel topic model, denoted as Hashtag Graph-based Topic Model (HGTM) to discover topics of tweets. By utilizing hashtag relation information in hashtag graphs, HGTM is able to discover word semantic relations even if words are not co-occurred within a specific tweet. With this method, HGTM successfully alleviates the sparsity problem. Our investigation illustrates that the user-contributed hashtags could serve as weakly-supervised information for topic modeling, and the relation between hashtags could reveal latent semantic relation between words. We evaluate the effectiveness of HGTM on tweet (hashtag) clustering and hashtag classification problems. Experiments on two real-world tweet data sets show that HGTM has strong capability to handle sparseness and noise problem in tweets.*

Keywords: Shortness, Graphs, HGTM, Tweets.

INTRODUCTION

MICROBLOGGING platforms such as Twitter have gone global. In this paper, introduce a new topic model to understand the chaotic microblogging environment by using hashtag graphs. Text content is one of the most important elements of social networks. It has been well recognized that uncovering topics of these user-generated contents is crucial for a wide range of content analysis tasks, such as natural disaster awareness , emerging topic

detecting , interesting content identification , user interest profiling, real time web search , et al. Characterizing contents of documents is a standard problem addressed in information retrieval and statistical natural language processing. Achieving good representations of documents could benefit tasks of organizing, classifying and searching a collection of documents. In recent years, topic models such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) , have been recognized as powerful methods of learning semantic representations for a corpus. According to the assumption that each document has a multinomial distribution over topics and each topic is a mixture distribution over words. Although traditional methods have achieved success in uncovering topics for normal documents (e.g., news articles, technical papers), the characteristics of tweets bring new challenges and opportunities to them. There are three key reasons. First, the severe sparsity problem of tweet corpora invalidates traditional topic modeling techniques. Typically, LDA and PLSA both reveal the latent topics by capturing the document-level word cooccurrence patterns. Compared with normal texts, tweets usually contain only a few words. Furthermore, the usage of informal language enlarges the size of the dictionary. Second, conventional topic models are designed for flat texts without structure. On Twitter, hashtags, prefixing one or more characters with a hash symbol as "#hashtag", are a community-driven convention for adding both additional context and metadata to tweets, making tweets semi-structured texts. Hashtags are created or selected by users to categorize messages and highlight topics. They provide a crowd sourcing way for tagging short texts, which is usually ignored by Bayesian statistics and machine learning methods. Last but not least, such crowd wisdom information clashes with the assumption of Independent Identical Distribution (i.i.d) of documents. The weakly-supervised information provided by hashtags can build direct semantic relations between tweets so that the words in tweets have more complex topical relationships than in normal texts.

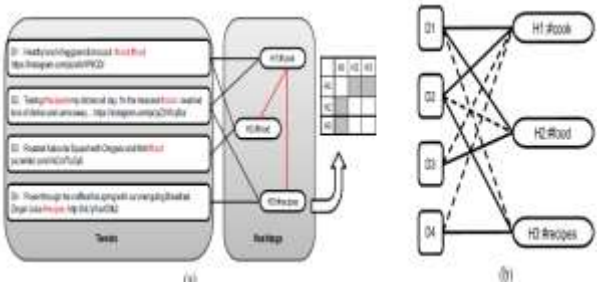

Fig: Using Hashtag Graph-based Topic Model to Connect Semantically-related Words without Co-occurrence in Microblogs

PROPOSED SYSTEM

In the proposed system, treating tweets as semi-structured texts, the system proposes a novel topic model, denoted as Hashtag Graph-based Topic Model (HGTM) to discover topics of tweets. By utilizing hashtag relation information in hashtag graphs, HGTM is able to discover word semantic relations even if words are not co-occurred within a specific tweet. With this method, HGTM successfully alleviates the sparsity problem. Our investigation illustrates that the user-contributed hashtags could serve as weakly-supervised information for topic modeling, and the relation between hashtags could reveal latent semantic relation between words. We evaluate the effectiveness of HGTM on tweet

(hashtag) clustering and hashtag classification problems.

ADVANTAGES

- Improving capability to handle sparseness and noise problem in tweets
- We evaluate the effectiveness of HGTM on tweet (hashtag) clustering and hashtag classification problems.
- We evaluate HGTM on two real-world Twitter data sets to understand different kinds of hashtag graphs and the working of HGTM on extensive tweet mining tasks such as clustering, classification, and topic quality evaluation.
- Compared to the state-of-the-art methods, HGTM shows the ability of handling the sparseness and noise problem in mining tweets by exploiting both explicit and potential relations between hashtags and tweets.

SYSTEM DESIGN

One is explicit relationship that contains inclusion relations between tweets and hashtags and co-occurrence relations between hashtags, as Figure 1(a) shows. Due to the explicit relationship, tweets sharing the same hashtags have highly overlapping correlated topics. The other one is potential relationship shown as dotted lines in Figure 1(b). A tweet should have a possibility to connect or contain those hashtags that have no explicit relationship with, but have a lot of co-occurrences with hashtags the tweet has already contained. Hence, hashtag co-occurrences in tweets indirectly contribute wider semantic relationship between tweets. It is easy to figure out, as shown in Figure 1(a), users anticipate the topic of “Cook” by adding the hashtags “#cook”, “#food”, “#cook” in tweet D1, D2, D3 and D4. The same hashtag bridges tweets with explicit relationship (i.e., hashtag inclusion relation) as an aggregation solution. Furthermore, hashtag co-occurrences in a whole corpus indirectly give a chance to connect tweets with no hashtag sharing. For example, word “Breakfast” in tweet D4 and word “lunch” in tweet D1 are obviously semantically related. Unfortunately, one tweet or the aggregation solution couldn’t handle or find out such a semantic relationship. Whereas, we can connect these two words through the path “D4”-“#recipes”- “#cook”-“D1” based on the hashtag co-occurrences in the whole dataset shown in Figure 1(a). That means D4 should have a potential relationship with “#cook” (in a dotted link as Figure 1(b) shows), and D1 can be connected to ”#recipes” as well. These connections tackle the problem of sparseness in tweets as a weakly-supervised information and build a meaningful semantic relation between words. Inspired by the observations mentioned above, we construct different kinds of hashtag graphs based on statistical information of hashtag occurrence in a crowdsourcing manner that can be acquired without human efforts such as labeling. Based on these hashtag graphs, we propose a novel framework of Hashtag Graphbased Topic Model (HGTM). The basic idea of HGTM is to project tweets into a coherent semantic space by using latent variables via user-contributed hashtags. HGTM provides a robust way for noisy and sparse tweets, which is different from traditional topic models since they normally consider only content information and ignore explicit and

potential semantic connection via noisy hashtags. HGTM is a probability generative model that incorporates such weakly-supervised information\ based on a weighted hashtag graph. The model links tweets via both explicit and potential tweet-hashtag relationship, so that hashtag relationship can connect semantically-related words with\ or without co-occurrences, which alleviates severe sparse and noise problem in short texts. In our previous work, we have verified the effectiveness that HGTM can bridge semanticrelated words when they share no co-occurrences.

CONCLUSION

In this paper , Uncovering topics within tweets has become a vital task for widespread content analysis and social media mining. Different from modeling normal text, tweet mining has suffered a great deal of sparseness and informality problems. In this work, we consider that users have provided hashtags as a powerful and valuable data source in the vast amount of tweets on the web. This paper presents HGTM that first introduces the hashtag relation graphs as weakly-supervised information for tweet semantic modeling. We demonstrate that hashtag graphs contain reliable information to bridge semantically-related words in sparse short texts. HGTM can enhance semantic relations between tweets and reduce noise at the same time. Compared to single documentoriented topic models (e.g., LSA, LDA, ATM, TWTM, TWDA), HGTM has a better ability to capture semantic relations between words with or without co-occurrence by utilizing the wisdom of crowds from user generated hashtags. The model provides a more robust solution for tweet modeling than aggregation strategies with traditional topic models. We also prove that LDA framework inherently can not benefit from hashtag graphs. We achieve significant improvement on the performance of content mining tasks, such as tweet clustering, hashtag clustering and hashtag classification. HGTM discovers more readable and distinguishable topics than the stat-of-the-art models as well. This paper shows one effective alternative of utilizing usercontributed \ hashtags for tweet topic modeling to handle both sparseness and noise in tweets. However, there are still many questions which need to be explored. For example, we would like to explore reasonable and effective ways of combining multimodal hashtag relations for tweet modeling and to model time- sensitive hashtag relations. The resulting model is highly scalable and could be used in a number of real-world applications, such as hashtag recommendation, short text retrieval, and event detection.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” JASIS, vol. 41, no. 6, pp. 391–407, 1990.

[3] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, “Microblogging during two natural hazards events: What Twitter may contribute to situational awareness,” in Proceedings of the

SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1079–1088.
[4] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, "Emerging topic detection for organizations from microblogs," in Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '13. New York, NY, USA: ACM, 2013, pp. 43– 52.
[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: Experiments on recommending content from information streams," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1185– 1194.
[6] K. Tao, F. Abel, Q. Gao, and G.-J. Houben, "TUMS: Twitter-based user modeling service," in The Semantic Web: ESWC 2011 Workshops, ser. Lecture Notes in Computer Science, R. Garca-Castro, D. Fensel, and G. Antoniou, Eds. Springer Berlin Heidelberg, 2012, vol. 7117, pp. 269–283.
[7] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha, "Time is of the essence: Improving recency ranking using Twitter data," in Proceedings of the 19th International Conference on World Wide Web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 331–340.
[8] T. Hofmann, "Probabilistic latent semantic indexing," in Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57.