A Bottom-up Method for Constructing Topic Hierarchies

Yuhao Zhang Wenji Mao Xiaochen Li
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{zhangyuhao2012, wenji.mao, xiaochen.li}@ia.ac.cn

***Abstract*—In security-related applications, it is of great need to construct topic hierarchies from the data automatically. It can help us better understand the contents and structure of information and benefit many applications in security informatics. The existing topic hierarchy construction methods either need to specify the structure manually, or are not robust enough for sparse social media data such as microblog. In this paper, we propose an approach to automatically construct topic hierarchies from microblog data in a bottom up manner. We first detect the fine grained topics and then build the topic structure based on a multi-branch hierarchical construction method. We conduct a preliminary empirical study based on Weibo data. The experimental results show that the hierarchical topic structure generated by our method can provide meaningful results.**

Keywords—topic hierarchies; topic detection; social media

I. INTRODUCTION

The analysis of huge social media data is increasingly important in recent years, especially for various security-related applications. Constructing topic hierarchies from online social media can help us better understand the contents and structure of information and facilitate decision making, emergency response and management, and many other applications in security informatics.

Many works have been devoted to constructing topic hierarchies. hLDA [1] can construct topic hierarchies automatically from data based on a non-parametric Bayesian approach, but the results of the method are not very interpretable sometimes for sparse and noisy data. Another generative method hPAM [2] constructs topic hierarchies based on a directed acyclic graph model, but the method itself needs to specify the topic structure manually.

In this paper, we propose an approach that can automatically construct topic hierarchies from microblog data in a bottom up manner. Our approach first extracts fine grained topics and then builds topic hierarchies using a multi-branch hierarchical construction method. We also conduct an empirical study to compare our approach with the related work and show its advantage in topic hierarchy construction.

II. PROPOSED METHOD

We use the nonnegative matrix factorization with L2 regularizer to extract topics from the documents. By specifying a relatively large topic number, we get the fine grained topics. Each one can be seen as a basic topic tree with one node. Then we combine the two trees which are most similar in topic at each round. By iteratively doing this, we can construct the entire topic hierarchies, in which the topics of the non-leaf nodes are the weighted sum of their children.

When merging topic trees, we need to choose a proper combination mode. The modes used here are the same as those used in Bayesian rose tree (BRT) [3]. Unlike BRT handling data in a generative way, we construct the topic hierarchies by comparing three topic similarity measures, the inner topic similarities of two trees I_a, I_b and the pairwise topic similarity between two trees $P_{a,b}$. The inner topic similarity reflects the closeness of the subtopics within a tree. So we define I_a as the average subtopic similarities of Tree(a). If the tree contains no subtopics, I_a is set to infinity. So is I_b. $P_{a,b}$ reflects the topic similarity of two trees and is measured by the similarity between root node topics. By comparing $P_{a,b}$ with γI_a and γI_b respectively (γ is a coefficient between 0-1), we choose the proper combination mode which can best reflect the relationship between two trees.

III. EXPERIMENT

We compare our method with the related work using Weibo dataset on the topic of the 26th APEC. We evaluate the results based on topic intrusion method [4]. The experimental results are summarized as follows: (1) The topic hierarchies our method constructed make sense as the unpopular topic is usually represented by a single node while the hot topic is a subtopic tree. (2) The results of the topic intruder task show the hierarchies our method constructed reflect the parent-child relationship of topics better than the related work.

IV. CONCLUSION

In this paper, we propose an approach to automatically construct topic hierarchies from microblog data in a bottom up manner. We detect topics using NMF and build the topic hierarchies by using a hierarchical construction method. The experiment results show the advantage of our method.

ACKNOWLEDGMENT

This work is supported by NSFC grants #61175040, #71462001, #91024030, #91224008 and #U1435221, Ministry of Health grant #2013ZX10004218, and grant #2013A127.

REFERENCES

[1] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical Topic Models and the Nested Chinese Restaurant Process," in *Neural Information Processing Systems*, 2003.
[2] D. Mimno, W. Li, and A. McCallum, "Mixtures of Hierarchical Topics with Pachinko Allocation," in *Proceedings of the 24th International Conference on Machine Learning*, New York, NY, USA, 2007, pp. 633–640.
[3] C. Blundell, Y. W. Teh, and K. A. Heller, "Bayesian Rose Trees," *arXiv:1203.3468 [cs, stat]*, Mar. 2012.
[4] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Neural Information Processing Systems*, 2009.