MATEC Web of Conferences **22**, 01062 (2015)
DOI: 10.1051/matecconf/20152201062

Microblog Hot Spot Mining Based on PAM Probabilistic Topic Model

Yaxin Zheng & Liu Ling
Chongqing University, Chongqing, China

ABSTRACT: Microblogs are short texts carried with limited information, which will increase the difficulty of topic mining. This paper proposes the use of PAM (Pachinko Allocation Model) probabilistic topic model to extract the generative model of text's implicit theme for microblog hot spot mining. First, three categories of microblog and the main contribution of this paper are illustrated. Second, for there are four topic models which are respectively explained, the PAM model is introduced in detail in terms of how to generate a document, the accuracy of document classification and the topic correlation in PAM. Finally, MapReduce is described. For the number of microblogs is huge as well as the number of contactors, the totally number of words is relatively small With MapReduce, microblogs data are split by contactor, document-topic count matrix and contactor-topic count matrix can be locally stored while the word-topic count matrix must be globally stored. Thus, the hot spot mining can be achieved on the basis of PAM probabilistic topic model.

Keywords: microblog; hot spot; PAM probabilistic topic model; MapReduce

1 INTRODUCTION

Microblog is becoming a major source for producing and spreading hot spot on the internet, it provides a short and convenient way for users to express and share their attitudes instantly. With the development of Web2.0, there is a continuous improvement in people's participation and the method of using the Internet has been changed that people are no longer passive in acquisition of knowledge from the network, but to express their own views initiatively or attitude towards others or events through a network. Microblogging updates message via short 140 characters, and achieve instant share of multi-tool, making it an important new media on the Internet. The emergence of microblogging makes the information present in microblogging with characteristics of fragmentation, instant and mobility, rather than complete contextual information. Through the microblogging freedom, convenience and instant way to express their feelings, it has become fashion on the Internet, but also makes it an important place to generate and talk about hot spot, which means the event, the topic or the information has been widely concerned, debated and discussed within a certain time, so the study on hot spot finding, monitoring and management in microblogging platform will become increasingly important.

MapReduce is a framework for processing huge datasets with distributed Map and Reduction operations (Figure 1). In Map step, the master node splits input data into partitions, which can be processed by user-defined Map functions, and produces (key, value) tuples as the intermediate output. In the reduce step, the reduce functions are used to merge all intermediate

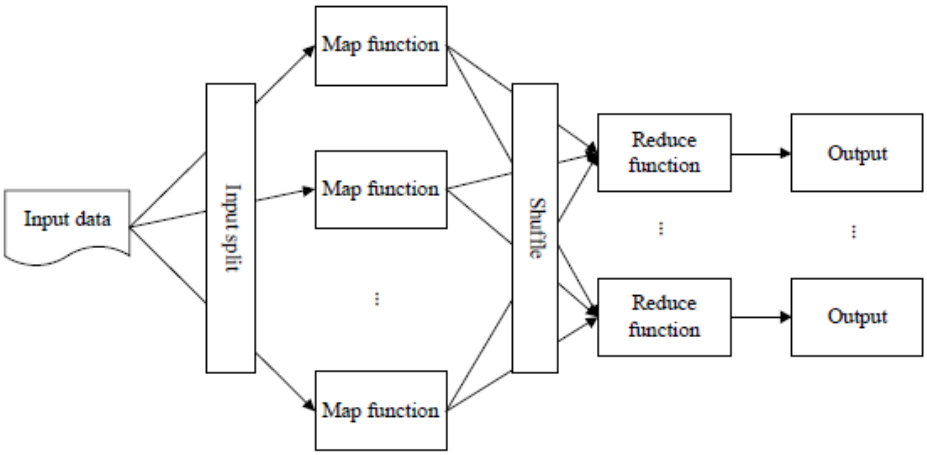


Figure 1. MapReduce framework

tuples with the same key, sort them and output the final (key, value) tuples. Both the map function and reduce function can be executed in parallel on no overlapping input and intermediate data.

In MapReduce framework, microblogs must be split into partitions. Under the situation of real world application, the number of microblogs is huge (over millions) as well as the number of contactors, while the total number of words is relatively small. Through this analysis on the inference process of PAM, we find out if microblogs data are split by contactor, the document-topic count matrix and the contactor-topic count matrix can be locally stored while the word-topic count matrix must be globally stored because only the word-topic count matrix needs global updates.

At present, the domestic microblogging research is in its infancy, and many research questions are to be solved in the field. Today, the monitoring and management of hot spot mining is becoming important area of research in the huge information flow of microblogging. When a certain event occurs in hot microblogging platform, people can use microblog to express their view or attitude towards the hot spot. As to the microblog hot spot mining, the PAM probabilistic topic model can be used in microblogging platform. The probabilistic topic model is used to regard the theme as a lexical item of probability distribution, and the text is seen as a random mix theme. Compared with the clustering method, using the distribution of the word, a text message will be easily transformed into digital information modeling, with the ability to recognize the potential of large-scale centralized theme text message. Themes can be more intuitive expressions, which greatly simplify the complexity of the problem, and the application of topic model is increasingly widespread. As microblog is becoming more and more popular, microblog services have become information provider on a web scale, so researches on microblog begin to focus more on its content mining than solely user's relationship analysis before. Although traditional text mining methods have been studied well, no algorithm is especially designed for microblog data, which contain structured information on social network besides plain text.

With the continuous development of the theme model, it has a very wide range of applications in terms of text classification, information retrieval and natural language processing. The PAM probabilistic topic model has become one of the popular models because of its structure flexibility, it can access to a wealth of semantic association and it's difficult to produce the phenomenon of over-fitting and so on, and it currently has a certain application in the image retrieval, the document classification and the object recognition and so on. The goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure, the topics, per-document topic distributions, and the per-document per-word topic assignments are hidden structure.

2 CATEGORIES OF MICROBLOG AND PAPER CONTRIBUTION

Microblogs can be divided into three categories: broadcast, conversation, and retweet messages. Broadcast messages are most common and can be seen by any users; conversation messages starting with a special symbol "@" have specific contactors to talk to; retweet messages which is identified by a special symbol "RT" allows users to repost other's messages with their personal comments.

In the information explosion era, how to effectively dig out latent topics and internal semantic structures from large scale data is an important issue. Microblogs contain the structured information on social network except the plain text, and the relationships on social network can play a supporting role in topic mining. On the other hand, microblogs are short texts carried with limited information (which is restricted to 140 characters), which will increase the difficulty of the topic mining. These natural features of microblogs mentioned above can prevent the traditional text mining algorithms to be directly employed with their full potentials.

In this paper, we make the following contributions:

- A novel model PAM is proposed, it is suitable for microblog data by taking both structured information and unstructured information into consideration.
- Distributed PAM in MapReduce framework is proposed in order to meet the requirement of processing large scale microblogs with high scalability.

The rest of this paper is organized as follows. Section 3 introduces the novel model PAM for microblog mining. Section 4 proposes the PAM probabilistic topic model in MapReduce framework for large scale situations and it applies performance model to further optimize our approach.

3 PAM MODEL

There are four topic models, the model structures are respectively shown in Figure 2.

(a) Dirichlet Multinomial: For each document, a multinomial distribution on words is sampled from a single Dirichlet.

(b) LDA (*latent Dirichlet allocation*): This model samples a multinomial on topics for each document, and then generates words from the topics.

(c) Four-Level PAM: A four-level hierarchy consists of a root, a set of sub-topics, a set of super-topics and a word vocabulary. The super-topics and the root are both associated with Dirichlet distributions, from which we sample multinomial over their children for each document.

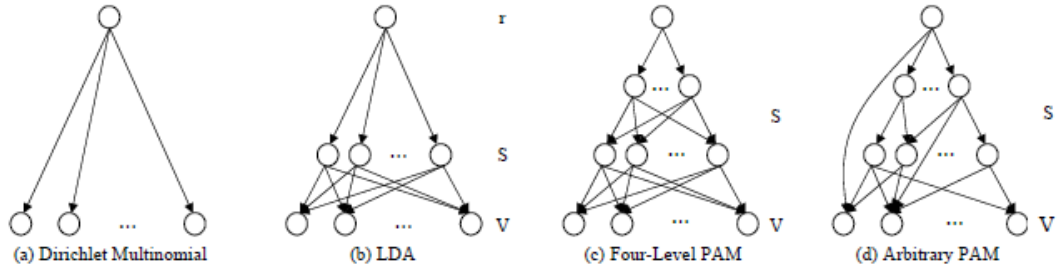


Figure 2. Model structures for four topic models

(d) PAM: An arbitrary DAG structure is used to encode the topic correlations. Each interior node is considering a topic and it is associated with a Dirichlet distribution.

In this section, we introduce the pachinko allocation model (PAM), which uses a directed acyclic graph (DAG) structure to represent and learn arbitrary-arity, nested, and possibly sparse topic correlations. In PAM, the concept of topics is extended to be distributions not only over words, but also over other topics. The model structure consists of an arbitrary DAG, in which each leaf node is associated with a word in the vocabulary, and each non-leaf\interior node that corresponds to a topic has a distribution over its children. An interior node whose children are all leaves would correspond to a traditional LDA topic. But some interior nodes that may also have children are other topics, thus they are represented as a mixture over topics. With many nodes like these, PAM therefore captures not only correlations among words (as in LDA), but also correlations among topics themselves. The LDA and four-level PAM graphical models are shown in Figure 3. As we can see, the major difference is that PAM has one additional layer of super-topics modeled with Dirichlet distributions, which are the key components of capturing topic correlations here.

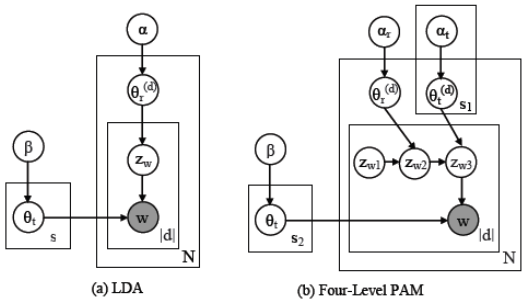


Figure 3. LDA and Four-Level PAM

To generate a document, LDA samples a multinomial distribution over topics from g (α), then repeatedly samples a topic from this multinomial, and a word from the topic. Now we introduce notation for the pachinko allocation model. PAM is used to connect words in V and topics in T with an arbitrary DAG, where topic nodes occupy the interior levels and the leaves are words.

To generate a document d, we follow a two-step process:

(1) Sample $\theta_{t1}, \theta_{t2}, \ldots, \theta_{ts}$ from $g_1(\alpha_1), g_2(\alpha_2), \ldots, g_s(\alpha_s)$, where θ_{ti} is a multinomial distribution of topic t_i over its children.

(2) For each word ω in the document,

– Sample a topic path Z_ω of length L_ω : $< Z_{\omega 1}, Z_{\omega 2}, \ldots, Z_{\omega L_\omega}, >$. $Z_{\omega 1}$ is always the root and $Z_{\omega 2}$ to $Z_{\omega L_\omega}$ are topic nodes in T. $Z_{\omega i}$ is a subset of $Z_{\omega(i-1)}$ and it is sampled according to the multinomial distribution $\theta_{Z_{\omega(i-1)}}$.

– Sample word ω from $Z_{\omega L_\omega}$.

Following this process, the joint probability of generating a document d, the topic assignments z and the multinomial distributions θ are shown as follows:

$$P(\mathrm{d}, \mathrm{z}, \theta | \alpha) = \prod_{i=1}^{s} P(\theta_{ti} | \alpha_i) \times \prod_{\omega} \prod_{i=2}^{L\omega} P(\mathrm{z}_{\omega i} | \theta_{\mathrm{z}_{\omega(i-1)}}) P(\omega | \theta_{\mathrm{z}_{\omega L\omega}})$$

Through integrating out θ and summing over z, we calculate the marginal probability of a document which is shown as follows:

$$P(\mathrm{d} | \alpha) = \int \prod_{i=1}^{s} P(\theta_{ti} | \alpha_i) \times \prod_{\omega} \sum_{z_\omega} (\prod_{i=2}^{L\omega} P(\mathrm{z}_{\omega i} | \theta_{\mathrm{z}_{\omega(i-1)}}) P(\omega | \theta_{\mathrm{z}_{\omega L\omega}})) \mathrm{d}\theta$$

Finally, the probability of generating a whole corpus is the product of the probability for every document:

$$P(\mathrm{D} | \alpha) = \prod_{d} P(d | \alpha)$$

As to the accuracy of document classification, we can refer to [11], it is shown as follows:

class	# docs	LDA	PAM
Graphics	243	83.95	86.83
Os	239	81.59	84.10
Pc	245	83.67	88.16
Mac	239	86.61	89.54
Windows. x	243	88.07	92.20
Total	1209	84.70	87.34

Figure 4. Accuracy of document classification

When we randomly choose a subset of abstracts

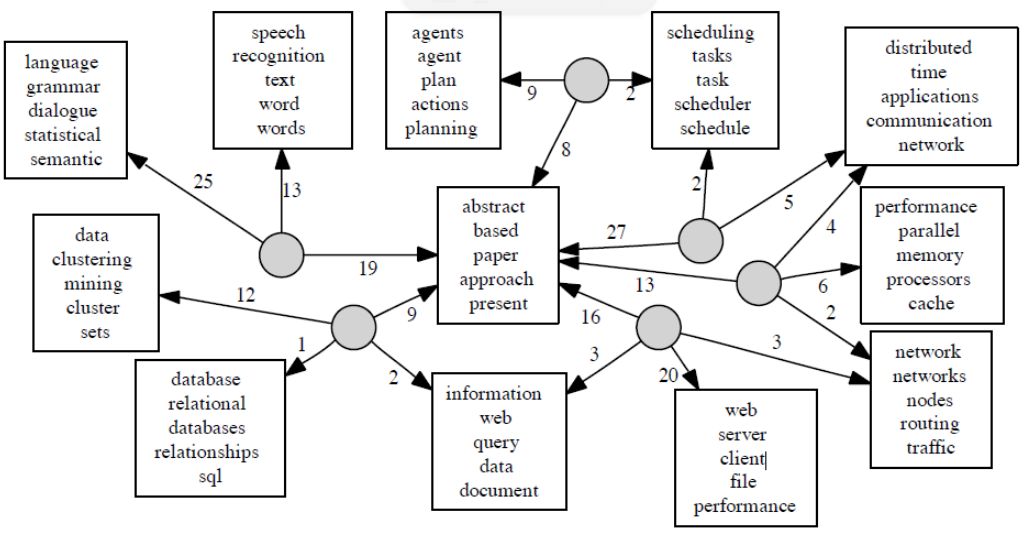


Figure 5. Topic correlation in PAM

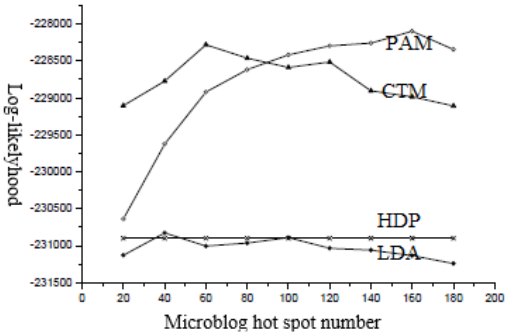
from its large collection, where there are 4000 documents, 278438 word tokens and 25597 unique words. Based on PAM, the topic correlation can be seen in Figure 5.

Each circle corresponds to a super-topic each box corresponds to a sub-topic. One super-topic can be used to connect to several sub-topics and capture their correlation. The numbers on the edges are the corresponding values for the (super-topic, sub-topic) pair. Figure 4 shows a subset of super-topics in the data, and how they capture correlations among sub-topics.

4 SIMULATIONS AND RESULTS

After enough iterations, the model reaches the convergence and calculates out the final count matrices. With these count matrices, we can obtain the related distributions, which are useful in topic mining for microblogs. In MapReduce environment with a cluster's resource, the mappers are mainly used to conduct the sampling process in parallel and the reducers mainly update the count matrices for next iteration, which makes PAM effective and scalable in topic mining for large scale microblogs. Compared with the same PAM-based microblog hot spot mining method in [13], this paper proposes the MapReduce description, and splits microblogs data into contactor so that document-topic count matrix and contactor-topic count matrix can be locally stored. PAM and LDA have been introduced in detail as mentioned above. As to CTM and HDP, we will introduce them in brief. Although CTM and PAM are both trying to model topic correlations directly, PAM takes a more flexible approach that can capture the nested correlations. In fact, CTM is very similar to a special-case structure of PAM, while HDP can capture different topic correlations within these groups by using a nested hierarchy of Dirichlet processes.

To validate the effectiveness and feasibility of the proposed PAM probabilistic topic model method in microblog hot spot mining, we can make simulations of PAM, LDA, CTM and HDP that have been proposed in some references. Choose the index of Log-likelihood as an evaluation index, the simulation results can be seen in Figure 6. Then, we can see that for given microblog hot spot number and training data, the Log-likelihood of PAM is much bigger, which implies that the PAM probabilistic topic model is effective and feasible in the microblog hot spot mining.



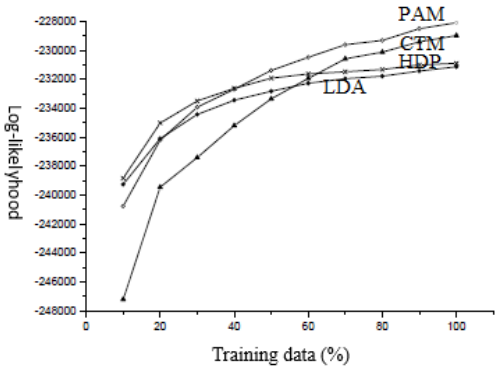


Figure 6. Simulation results

5 CONCLUSIONS

In this paper, we have presented pachinko allocation, a mixture model which uses a DAG structure to capture arbitrary topic correlations. The DAG structure is completely general, and some topic models like LDA can be represented as special cases of PAM. Compared with other approaches for microblog hot spot mining such as hierarchical LDA and correlated topic model, PAM provides more expressive power to support complicated topic structures and adopts more realistic assumptions for generating documents. In addition, this paper introduces the novel model PAM for microblog mining and proposes the PAM probabilistic topic model in MapReduce framework for large scale situations and applies performance model to further optimize our approach.

REFERENCES

[1] Blei D M, Lafferty J D. 2009. Topic models. *Text Mining: Classification, Clustering, and Applications*, 10: 71.

[2] Ramage D, Hall D, Nallapati R, et al. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp: 248-256.

[3] Mimno D, Li W, McCallum A. 2007. Mixtures of hierarchical topics with pachinko allocation. *Proceedings of the 24th International Conference on Machine Learning*. ACM, pp: 633-640.

[4] Li W., McCallum A. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd International Conference on Machine Learning*. ACM, pp: 577-584.

[5] Liu C, Wang F, Shi K, et al. 2014. Robust H∞ control for satellite attitude control system with uncertainties and additive perturbation. *International Journal of Science*, 1(2): 1-9.

[6] Blei D M. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4): 77-84.

[7] Andrzejewski D, Zhu X, Craven M. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, pp: 25-32.

[8] Liu C, Wang F. 2014. In-orbit estimation of inertia parameters of target satellite after capturing the tracking satellite. *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, pp: 3942-3947.

[9] Mimno D, McCallum A. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv preprint arXiv:1206.3278, 2012.

[10] Teh Y W, Jordan M I, Beal M J, et al. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476).

[11] Yu M, Wang J, Zhao X, et al. 2013. Research on PAM Probability topic model. *Computer Science*, 40(5): 1-7.

[12] Liu C, et al. 2014. Mass and mass center identification of target satellite after rendezvous and docking. *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, pp: 5802-5807.

[13] Zhang C, Sun J. 2012. Large scale microblog mining using distributed MB-LDA. *Proceedings of the 21st International Conference Companion on World Wide Web*. ACM, pp: 1035-1042.

[14] Yu M, Zhou Z, et al. 2013. PAM-based microblog hot spot mining. *Technique and Method*, 32(15): 86-89.