News Topic Evolution Tracking by Incorporating Temporal Information

Jian Wang, Xianhui Liu*, Junli Wang, and Weidong Zhao

Engineering Research Centre of Ministry of Education
on Enterprise Digitalization Technology, Tongji University, China
lxh@tongji.edu.cn

Abstract. Time stamped texts or text sequences are ubiquitous in real life, such as news reports. Tracking the topic evolution of these texts has been an issue of considerable interest. Recent work has developed methods of tracking topic shifting over long time scales. However, most of these researches focus on a large corpus. Also, they only focus on the text itself and no attempt have been made to explore the temporal distribution of the corpus, which could provide meaningful and comprehensive clues for topic tracking. In this paper, we formally address this problem and put forward a novel method based on the topic model. We investigate the temporal distribution of news reports of a specific event and try to integrate this information with a topic model to enhance the performance of topic model. By focusing on a specific news event, we try to reveal more details about the event, such as, how many stages are there in the event, what aspect does each stage focus on, etc.

Keywords: Temporal Distribution, LDA, News Topic Evolution.

1 Introduction

Wth the dramatic increase of these digital document collections, the amount of information is far more beyond that person can efficiently and effectively process. There is a great demand for developing automatic text analysis models for analyzing these collections and organizing its contents. Probabilistic topic models, such as Latent Dirichlet Allocation (LDA) [1], Author-Topic Model [2] were proven to be very useful tools to address these issues.

With the need to model the time evolution of topics in large document collections, a family of probabilistic time series models were developed. Dynamic topic model (DTM) [3] captures the evolution of topics in a sequentially organized corpus of documents. Topic over Time (TOT) [4] model treats time as a continuous variable. Continuous Dynamic Topic Model (cDTM) [5] uses Brownian motion to model continuous time topic evolution. iDTM [6] is an infinite dynamic topic model which allows for an unbounded number of topics and captures the appearance and vanishing of topics. These models are quite useful when dealing with corpus with many different topics mixed up, but when it comes to a specific

* Corresponding author.

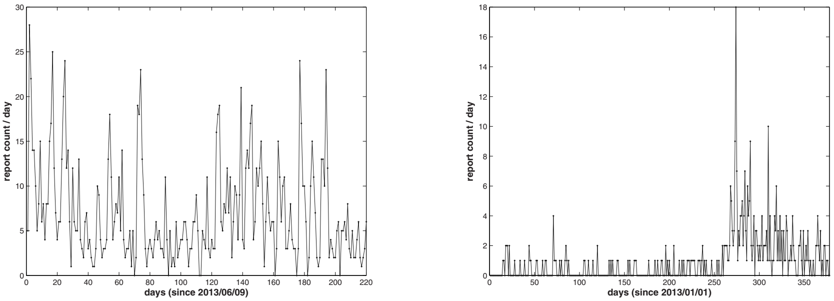
C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 465–472, 2014.

news event, the result seems to be not such remarkable. A couple of methods for generating timeline were proposed to deal with these issues in recent year [7], [8], [9].

However, most of these methods are trying to get a summarization of the event from the text but none of these methods mentioned above have taken the advantage of the temporal information as a prior knowledge. In this paper, we first explore the temporal distribution of news event, then propose an algorithm to automatically divide the corpus into different stages, in which the documents may have more coherence. By incorporating temporal information to topic model, we introduce a framework for tracking a specific news event evolution. The rest of this paper is organized as follows. In section 2, we illustrate the temporal distribution of the news event and describe the division algorithm in detail. In section 3, we propose our analysis framework and explain how temporal information can enhance the topic model. In section 4, we present the case study experiment in detail. In section 5, we conclude the paper with some analysis and outlook for future work.

2 Temporal Distribution of News Events

When a sensational event burst out, related reports will overflow in media very soon. Later the quantity of related reports would gradually decline. But once new details are disclosed or someone else is involved, the event gains its popularity again and likewise the amount of related reports would move up sharply. Suppose we label each of the popular period as a stage. Generally, most news events may have several stages, and each stage has its own focus. The results showed in the Fig.1 are exactly in conformity with what we've assumed above.



(a) Edward Snowden

(b) Obamacare

Fig. 1. Temporal distribution of news reports quantity about "Edward Snowden" and "Obamacare". X-axis represents day's interval from the beginning date and the Y-axis represents article count. Articles were crawled from The Guardian with the key words "Edward Snowden" from 9 June 2013 to 10 Jan 2014, and "Obamacare" from 1 Jan 2013 to 10 Jan 2014.

2.1 Documents Division and the Adaptive K-Means Algorithm

In Blei's DTM [3], the corpus was evenly divided by time, thus every episode has the same time scale. Let's take a look at the Fig.1, if the documents are divided by time evenly, the dividing points may just locate at the peak point. Intuitively, this is not a good choice. Because the reports around the peak point mainly focus on the same aspect and they have a strong coherence.

So, we propose a simple but efficient method which is called the Adaptive K-Means algorithm. This algorithm is based on the K-Means algorithm[10] and could automatically divide the documents without setting the cluster number K in advance. In this paper the cluster number K means the episode number. The algorithm is described as follow:

Algorithm 1: Adaptive K-Means algorithm

```
Data: $X$: news count of each day; $max\_k$: the maximum k; $t$: threshold value
Result: $count$: article count of each episode; $dists$: weighted mean distance
        array; $K$: the best count of cluster

$Y \longleftarrow$ remove $zero$ points from $X$
for $i \leftarrow 1$ to $max\_k$ do
    $[count, sumd] = kmeans(Y, i)$;
    // $count$: point count of each cluster
    // $sumd$: sum distance of each cluster
    $means \leftarrow calc\_mean\_distance(count, sumd)$;
    // $means$: mean distances of all clusters
    $dists[i] \leftarrow calc\_weighted\_mean\_distance(means)$;
    if $i > 1$ then
        if $dists[i] - dists[i-1] < t$ then
            $K \leftarrow (i-1)$; break;
        end
    end
end
if $K = 0$ then
    $K \leftarrow max\_k$;
end
```

The Adaptive K-Means algorithm starts with a small number of clusters, and adds the number one by one. At each iteration of the algorithm, we calculate the *weighted mean distance* of all clusters. The *weighted mean distance* is defined as follows:

$$Weighted\ Mean\ Distance = \frac{\sum_{i=1}^{n} mean\ distance\ of\ cluster\ i}{n} \qquad (1)$$

The distance calculated in the Equation.1 refers to Euclidean distance. In the beginning, the number of centers is much smaller than the best K, so the *Weighted Mean Distance* would decline rapidly. With the number getting closer to the best

K, the decrease value becomes smaller and smaller. Once the decrease value is smaller than a specific threshold value, then the current number of centers is regarded as the best K.

3 Incorporating Temporal Information into Topic Model

In this section, we illustrate how to use topic model to track news topic evolution and why temporal information can improve the analysis result.

3.1 Basic Concepts

First of all, we would like to give the definitions of some basic concepts which would be frequently mentioned.

1. A ***stage*** is a time episode in which documents have a strong coherence, and documents are likely related to a same aspect of the event
2. The ***main topic*** could run throughout all stages and is the line connecting all the episodes;
3. The ***auxiliary topics*** are all the other topics besides the main topic, which present the new aspects of the main topic in different stages; The auxiliary topics could be regarded as the progresses of the event because they very a lot along the time

In order to discover the main topic and track the evolution, we need to calculate the similarity between adjacent episodes. As the topic is characterized by a distribution over words. A simple measure method of similarity between topics is the *Kullback-Leibler divergence* (also called *relative entropy*). However, the *KL divergence* is not a proper distance measure method because it is not symmetric. An alternative option is the *Jensen-Shannon distance*, which is a smoothed and symmetric extension of the *KL divergence*. For discrete probability distribution P and Q, the *JS Distance* is defined to be

$$D_{JS}\left(P||Q\right) = \frac{1}{2}D_{KL}\left(P||M\right) + \frac{1}{2}D_{KL}\left(Q||M\right) \tag{2}$$

With the averaged variable $M = \frac{1}{2}(P + Q)$.

3.2 Framework of Our Method

Our analytical framework is based on the LDA [1], a generative latent variable model that treats documents as bags of words generated by one or more topics. We perform parameter estimation using collapsed Gibbs sampling [11] [12]. We could firstly divide the corpus into several subsets by time, and apply LDA within each subset, respectively. As for the division, we've described the *Adaptive K-Means algorithm* above which makes more sense than the method of simply dividing the corpus by time evenly

By incorporating temporal information, the overall framework of analysis process is as follows:

1. Prepare documents for each episode with *Adaptive K-Means algorithm*;
2. Preprocess of the documents in each episode;
3. Draw topic distribution of each episode from topic model (LDA);
4. Discover the main topic and draw the evolution map of the event.

3.3 Document Coherence and Evaluation

Document Coherence measures the topic similarity among documents. Intuitively, if articles within a corpus are more coherent, more detail could be revealed by topic models. In order to better present document coherence we propose a new evaluation method which is called *n Topic Coverage Rate(TCR_n)*.

$$TCR_n = \frac{\|articles\ \ belong\ to\ these\ n\ topics\|}{\|all\ articles\|} * 100\% \tag{3}$$

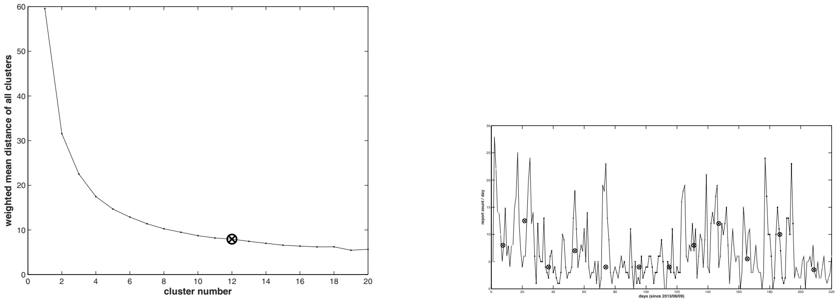
In this equation, $\|\cdot\|$ denotes the element count of a collection. TCR_n measures the documents cover rate of the top n topics within the whole corpus. From the Equation.3, we can see that with topic number n fixed, the bigger the TCR_n is, the more coherent the articles are. In other words, with the TCR_n fixed, the smaller the n is, the more coherent the articles are.

4 Experiment Result and Analysis

In this section we illustrate the result of the topic model with temporal information incorporated. First of all, we demonstrate the division result of the *Adaptive K-Means algorithm* with two corpora, and later focus on one of them to give a deep illustration. We analyze 1550 documents crawled from the Guardian with the key words "Edward Snowden" which is one of the top events of 2013. The time of these documents varies from June 9 of year 2013 to the end of year 2013. Our corpus is made up of approximately 1.5 million words. First of all, we use Stanford Parser[1] to parse the full text, and only keep words which are noun, proper noun, verb and adjective. Next, we lemmatize the remaining words. At last, we prune the vocabulary by removing stop words and removing terms that occurred less than 5 times. The remaining vocabulary size is 7732.

Fig.2 shows the result after applying the *Adaptive K-Means algorithm*. Our corpus of "Edward Snowden" is divided into 12 subsets. For the sake of document coherence comparison, we also divide the corpus into another 12 subsets by time evenly. We set the initialization parameters as follow: LDA parameters, α=2 and β=0.5; Gibbs sampling parameters, total iterations=1000, burn-in iterations=200, sample interval=3. After running topic model in each episode, respectively, we calculate TCR_n of all episodes. Fig.3(a) is an example of top 5 topics' coverage. Obviously, our division algorithm has a higher coverage. To be more convincible, we calculate the average coverage rates of all episodes with different topic numbers, and the Fig.3(b) shows that our method has a general advantage.

[1] http://nlp.stanford.edu/software/lex-parser.shtml



(a) Edward Snowden

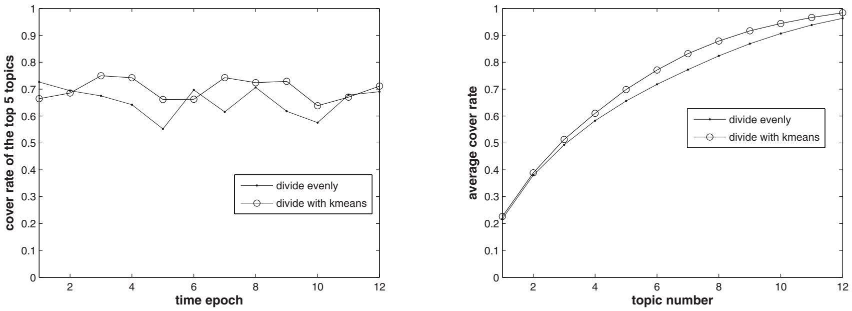
(b) Edward Snowden

Fig. 2. Cluster number determination process and documents division results by applying adaptive K-Means algorithm. The black ⊗ indicates the best number of clusters in (a), and the centre point in (b). In (b), data points in different clusters are labelled with different colours. (the maximum value of episode is 20 and the threshold value is 0.5)

4.1 Evolution Map of "Edward Snowden"

Considering that the corpus is about one specific event and there won't be too many aspects in each episode, so we only pick the top 3 topics of each episode to draw the **evolution map**.

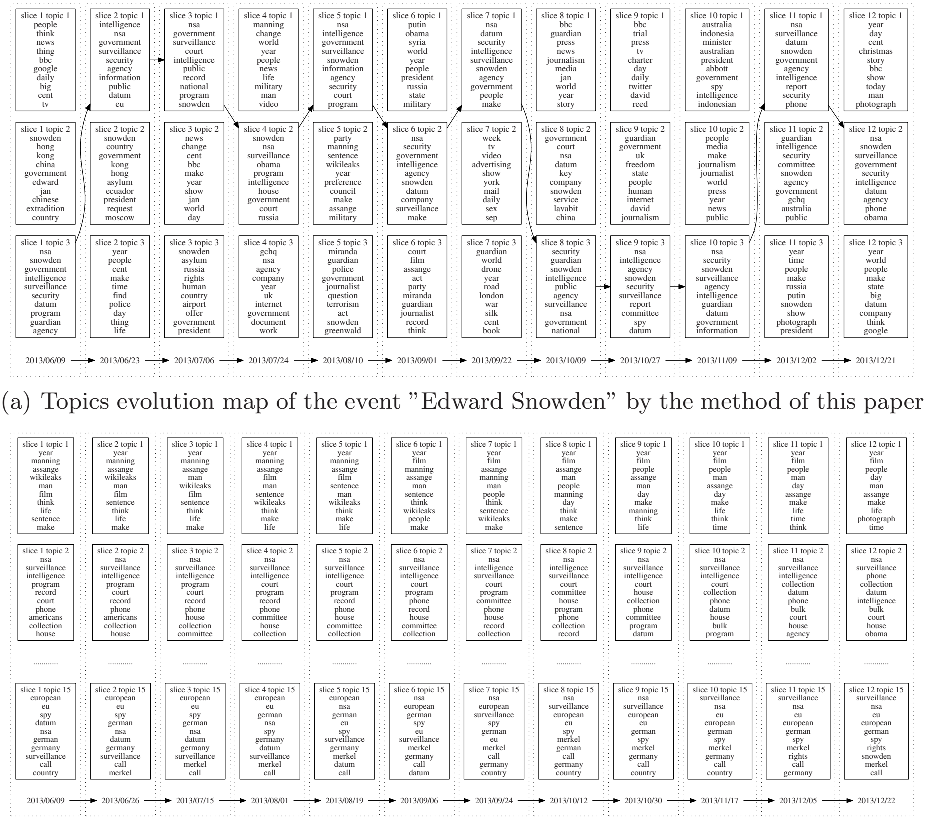
Fig.4(a) is the evolution map of the event "Edward Snowden" drawn from the method introduced in this paper. On this map, the main topic which runs throughout all stages is the one chained with arrow lines. To make a contrast, we also apply the DTM to generate evolution map Fig.4(b). From the Fig.4(b), we can see that topics in all the episodes seem to be almost the same. That's because DTM assumes that topic number is fixed during all episodes and no new topic would emerge and all the topics are evolved from the first episode. Thus, it



(a) TCR_5

(b) Average TCR_n

Fig. 3. TCR of corpus about "Edward Snowden"


slice 1 topic 1 people think news thing bbc google daily big cent tv
slice 2 topic 1 intelligence nsa government surveillance security agency information public datum eu
slice 3 topic 1 nsa government surveillance court intelligence public record national program snowden
slice 4 topic 1 manning change world year people news life military man video
slice 5 topic 1 nsa intelligence government surveillance snowden information agency security court program
slice 6 topic 1 putin obama syria world year people president russia state military
slice 7 topic 1 nsa datum security intelligence surveillance snowden agency government people make
slice 8 topic 1 bbc guardian press news journalism media jan world year story
slice 9 topic 1 bbc trial press tv charter day daily twitter david reed
slice 10 topic 1 australia indonesia minister australian president abbott government spy intelligence indonesian
slice 11 topic 1 nsa surveillance datum snowden government agency intelligence report security phone
slice 12 topic 1 year day cent christmas story bbc show today man photograph
slice 1 topic 2 snowden hong kong china government edward jan chinese extradition country
slice 2 topic 2 snowden country government kong hong asylum ecuador president request moscow
slice 3 topic 2 news change cent bbc make year show jan world day
slice 4 topic 2 snowden nsa surveillance obama program intelligence house government court russia
slice 5 topic 2 party manning sentence wikileaks year preference council make assange military
slice 6 topic 2 nsa security government intelligence agency snowden datum company surveillance make
slice 7 topic 2 week tv video advertising show york mail daily sex sep
slice 8 topic 2 government court nsa datum key company service lavabit china
slice 9 topic 2 guardian government uk freedom state people human internet david journalism
slice 10 topic 2 people media make journalism journalist world press year news public
slice 11 topic 2 guardian intelligence security committee snowden agency government gchq australia public
slice 12 topic 2 nsa snowden surveillance government security intelligence datum agency phone obama
slice 1 topic 3 nsa snowden government intelligence surveillance security datum program guardian agency
slice 2 topic 3 year people cent make time find police day thing life
slice 3 topic 3 snowden asylum russia rights human country airport offer government president
slice 4 topic 3 gchq nsa agency company year uk internet government document work
slice 5 topic 3 miranda guardian police government journalist question terrorism act snowden greenwald
slice 6 topic 3 court film assange act party miranda guardian journalist record think
slice 7 topic 3 guardian world drone year road london war silk cent book
slice 8 topic 3 security guardian snowden intelligence public agency surveillance nsa government national
slice 9 topic 3 nsa intelligence agency snowden security surveillance report committee spy datum
slice 10 topic 3 nsa security snowden surveillance agency intelligence guardian datum government information
slice 11 topic 3 year time people make russia putin snowden show photograph president
slice 12 topic 3 year world people make state big datum company think google
2013/06/09
2013/06/23
2013/07/06
2013/07/24
2013/08/10
2013/09/01
2013/09/22
2013/10/09
2013/10/27
2013/11/09
2013/12/02
2013/12/21
slice 1 topic 1 year manning assange wikileaks man film think life sentence make
slice 2 topic 1 year manning assange wikileaks man film sentence think life make
slice 3 topic 1 year manning assange man wikileaks film sentence think life make
slice 4 topic 1 year manning assange film man sentence wikileaks think make life
slice 5 topic 1 year manning assange film sentence man wikileaks think make life
slice 6 topic 1 year film manning assange man sentence think wikileaks people make
slice 7 topic 1 year film assange manning man people think sentence wikileaks make
slice 8 topic 1 year film assange man people manning day think make sentence
slice 9 topic 1 year film people assange man day make manning think life
slice 10 topic 1 year film people man assange day make life think time
slice 11 topic 1 year film people man day assange make life time think
slice 12 topic 1 year film people day man assange make life photograph time
slice 1 topic 2 nsa surveillance intelligence program record court phone americans collection house
slice 2 topic 2 nsa surveillance intelligence program record court phone americans collection house
slice 3 topic 2 nsa surveillance intelligence program court record phone house collection committee
slice 4 topic 2 nsa surveillance intelligence court program record phone committee house collection
slice 5 topic 2 nsa surveillance intelligence court program record phone house committee collection
slice 6 topic 2 nsa surveillance intelligence court program phone record house committee collection
slice 7 topic 2 nsa intelligence surveillance court program committee phone house record collection
slice 8 topic 2 nsa intelligence surveillance court committee house program phone collection record
slice 9 topic 2 nsa surveillance intelligence court house collection phone committee program datum
slice 10 topic 2 nsa surveillance intelligence court collection phone datum house bulk program
slice 11 topic 2 nsa surveillance intelligence collection datum phone bulk court house agency
slice 12 topic 2 nsa surveillance phone collection datum intelligence bulk court house obama
....
slice 1 topic 15 european eu spy datum nsa german germany surveillance call country
slice 2 topic 15 european eu spy german nsa datum germany surveillance call merkel
slice 3 topic 15 european eu spy german nsa datum germany surveillance merkel call
slice 4 topic 15 european eu german nsa spy germany datum surveillance merkel call
slice 5 topic 15 european nsa german eu spy surveillance germany merkel datum call
slice 6 topic 15 nsa european german spy eu surveillance merkel germany call datum
slice 7 topic 15 nsa european surveillance spy german eu merkel call germany country
slice 8 topic 15 nsa surveillance european eu spy merkel german call germany country
slice 9 topic 15 nsa surveillance european eu german spy merkel germany call country
slice 10 topic 15 surveillance nsa eu european german spy merkel germany call country
slice 11 topic 15 surveillance nsa eu european german spy merkel rights call germany
slice 12 topic 15 surveillance nsa eu european german spy rights snowden merkel call
2013/06/09
2013/06/26
2013/07/15
2013/08/01
2013/08/19
2013/09/06
2013/09/24
2013/10/12
2013/10/30
2013/11/17
2013/12/05
2013/12/22

(a) Topics evolution map of the event "Edward Snowden" by the method of this paper

(b) Topics evolution map of the event "Edward Snowden" by the DTM

Fig. 4. Evolution Map

is suitable for the corpus that contains many different topics, such as academic documents. While in this paper, we concentrate on a specific news event, so it is not applicable.

5 Conclusion

In this paper, we address the problem of modeling sequence documents related to a specified news event. We explore the temporal distribution of news reports and treat them as a prior knowledge of the sequence topic model. By incorporating the temporal information we can generate an evolution map of a specific event. In the future, we plan to model the entities involved in the event and explore how they influence the event evolution. We also intend to develop a interactive system to better explore the event detail.

Acknowledgement. This work is supported by the National Science Foundation of China under grant No. 61105047. This work is also partially supported by the Research Program of Science and Technology Commission of Shanghai Municipality of China under grant No 12dz1125400 and 13111103100. Here, we would like to express our deep gratitude.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Michal, R.-Z., Thomas, G., Mark, S., Padhraic, S.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487–494. AUAI Press (2004)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine learning, ICML 2006, pp. 113–120. ACM, New York (2006)
4. Wang., X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 424–433. ACM (2006)
5. Chong, W., David, B., David, H.: Continuous Time Dynamic Topic Models. In: Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI) (2008)
6. Ahmed, A., Xing, E.P.: Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. arXiv preprint arXiv:1203.3463 (2012)
7. Tang, S., Zhang, Y., Wang, H., Chen, M., Wu, F., Zhuang, Y.: The discovery of burst topic and its intermittent evolution in our real world. Communications, China 10(3), 1–12 (2013)
8. Zehnalova, S., Horak, Z., Kudelka, M., Snasel, V.: Evolution of Author's Topic in Authorship Network. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pp. 1207–1210. IEEE Computer Society (2012)
9. Lin, C., Lin, C., Li, J., Wang, D., Chen, Y., Li, T.: Generating event storylines from microblogs. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 175–184. ACM (2012)
10. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 881–892 (2002)
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)
12. Heinrich, G.: Parameter estimation for text analysis (2005)