Online Cross-Lingual PLSI for Evolutionary Theme Patterns Analysis

Xin Xin, Kun Zhuang, Ying Fang, and Heyan Huang

School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
{xxin,kzhuang,yfang,hhy63}@bit.edu.cn

Abstract. In this paper, we focus on the problem of evolutionary theme patterns (ETP) analysis in cross-lingual scenarios. Previously, cross-lingual topic models in batch mode have been explored. By directly applying such techniques in ETP analysis, however, two limitations would arise. (1) It is time-consuming to re-train all the latent themes for each time interval in the time sequence. (2) The latent themes between two adjacent time intervals might lose continuity. This motivates us to utilize online algorithms to solve these limitations. The research of online topic models is not novel, but previous work cannot be directly employed, because they mainly target at monolingual texts. Consequently, we propose an online cross-lingual topic model. By experimental verification in a real world dataset, we demonstrate that our algorithm performs well in the ETP analysis task. It can efficiently reduce the updating time complexity; and it is effective in solving the continuity limitation.

Keywords: Temporal Text Mining, Evolutionary Topic Patterns, Incremental/Online PLSI, Cross-lingual PLSI.

1 Introduction

Tracking the temporal evolutionary theme patterns (ETP) of documents is an important research issue nowadays. The original documents on the Web, such as news articles, research papers, etc., are generated in an unstructured stream, which would bring information overload problem to users. With the help of ETP analysis, it is more user-friendly to automatically group the documents in a temporal hierarchical structure. Figure 1(left) shows the ETP analysis result from the news articles (from NetEase, VOC, etc.) of the event Euro 2012. Before game starts, themes are mainly about the basic information of different teams, such as the stars, the strategies, etc.; after the game starts, the themes are evolved to the performances and reviews; before the final, the themes of the two participants in the final, Spain and Italy, merge into one theme, whose content includes the predictions from the fans, the comparison analysis, etc.; finally after the final, the theme is involved into the celebrations, the dominations, the reviews, and etc. We can see from this example that such a temporal hierarchical structure is very informative in tracking the development of an event. In addition,

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 74–85, 2013.

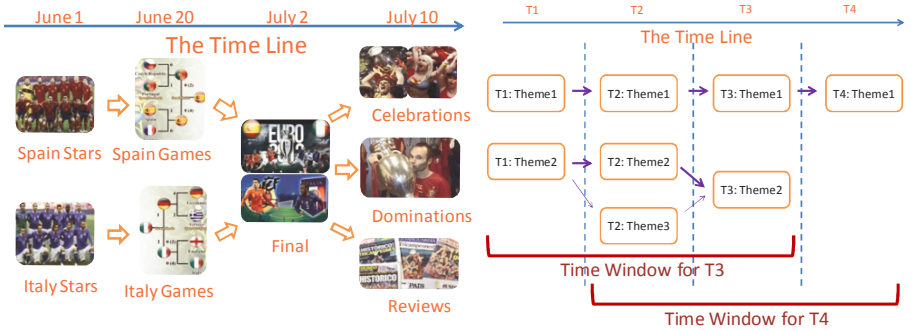


Fig. 1. An example of ETP analysis

it would be very helpful for information retrieval tasks from the semantic level. As a result, the ETP analysis has attracted more and more attention in both academia and industry [10–12].

In this paper, we focus on the issue of ETP analysis in the cross-lingual scenarios, which is different from previous ETP analysis [12] in monolingual texts. In this case, a topic is presented by documents of multiple languages rather than a monolingual one. Cross-lingual data would make the information more complete and the ETP analysis more accurate. For example, in Euro 2012 news, the discussion attitudes between a participant (e.g., Spain), and a third party member (e.g., China) might be quite different. The former would have bias to Spain; but the latter might be more neutral. Both of them should be considered in the ETP analysis.

Cross-lingual latent theme modeling is the key technique in cross-lingual ETP analysis. Competitive approaches include the cross-lingual probabilistic latent semantic indexing (PLSI) [15] model and the cross-lingual latent Dirichlet allocation (LDA) [9] model. The limitation of these competitive methods is that both of them are in batch mode. This would make the ETP analysis face the following two limitations.

Time-consuming Limitation. In the process of ETP analysis, the themes of documents in a new time interval should be integrated into the previous themes. In the batch mode, all the latent themes should be re-trained among the new documents and the old documents. In practice, the re-train process needs the enumeration of all the words in all the documents. Since the initial points are randomly chosen, it would also take many iterations of the enumeration before the model converges. Therefore, it is time-consuming to re-train the whole model for each time interval in the ETP analysis.

Continuity Limitation. If the re-train process is conducted for each time interval, the latent themes between adjacent time intervals might lose the continuity. The "continuity" has two meanings. 1) It is difficult to match the latent theme *ids* between the current time interval and the previous time interval. Because such *ids* are randomly assigned in the initial step of the algorithm. Thus the evolution process of themes might be not clear. 2) The latent themes between

adjacent time intervals might have different category angles and might not match each other. For example, in Fig. 1 (left), the latent themes are divided by teams. However, the latent themes could also be divided by the activities from fans, the activities from the match, etc. In fact, these themes do exist in the data. But it is not proper to have one category in one time interval, and have another in the next. An ideal ETP tracking should keep the continuity in all the time intervals.

These two limitations motivate us to investigate online algorithms for ETP analysis. Intuitively, when new documents arrive, the online algorithms only make local update to the model instead of global one, thus the time complexity would be much less than re-training the whole model. In addition, while the online algorithm retains the main body of the previous model, the continuity would be kept. Thus both the above limitations could be solved.

Online algorithms for topic models are not novel research tasks either. Both online PLSI [4] and online LDA [6] have been explored. But these models are mainly for monolingual scenarios. Thus they cannot be directly employed. Therefore, in this paper, we combine the ideas of previous work and propose an online cross-lingual topic model for cross-lingual ETP analysis. Comparing between the PLSI model and the LDA model, the LDA model utilizes the regularization technique to smooth the data fitting, which indeed performs well in previous tasks, but it also increases the training cost and the model complexity. As a preliminary study, in this paper, we first choose to combine the cross-lingual PLSI [15] and the online PLSI [4] for simplicity. We leave the combination of the cross-lingual LDA[9] and the online LDA [6] as our future work.

The main contributions of this paper lie in that we combine the ideas of previous cross-lingual PLSI and online PLSI together and propose an online cross-lingual topic model. In addition we utilize it in ETP analysis and demonstrate its efficiency for reducing the time complexity and its effectiveness for keeping the continuity property.

In the following of this paper, to make it consistent with previous work and easy for comparisons, *many previous definitions and variables in [4, 12, 15] are retained if being not necessarily changed, and our contribution is to combine these ideas in the cross-lingual ETP analysis task.*

2 Problem Definition

2.1 Preliminary Definitions

Definition 1 (Cross-lingual Corpus). *Following [15], the cross-lingual corpus is a dataset with multiple languages. We utilize* $C = \{C_1, C_2, ..., C_s\}$ *to denote the set of data collections with s languages. Each* C_i *denotes the data collection of a single language, whose vocabulary is denoted by* $W_i = \{w_1^i, w_2^i, ...w_{N_i}^i\}$. N_i *is the total word number in the* i_{th} *language. Each data collection* C_i *contains a set of documents,* $C_i = \{d_1^i, d_2^i, ..., d_{M_i}^i\}$. *Each document is presented by a bag of words, and a function* $c(w_k^i, d_j^i)$ *is utilized to denote the occurrence count of word* w_k^i *in document* d_j^i.

Definition 2 (Cross-lingual Theme). *Following [15], a cross-lingual theme* θ *is presented as a multinomial distribution of words, denoted by* $p(w|\theta)$, $w \in W_1 \bigcup W_2 \bigcup ... \bigcup W_s$. *For each* θ, *we have* $\sum_{i=1}^{s} \sum_{w \in Wi} p(w|\theta) = 1$. *Under this definition, a cross-lingual theme would gather the words of all the languages with related semantic meanings together. This is a super set of monolingual theme. A monolingual theme could be extracted from the cross-lingual theme by normalization as* $p_i(w^i|\theta) = \frac{p(w^i|\theta)}{\sum_{w \in W_i} p(w|\theta)}$.

Definition 3 (Time Interval and Time Window). *As shown in Fig. 1 (right), the time line is divided into discrete time intervals. Each time interval is a fixed number of days, e.g., a week of seven days. For each time interval, the themes are automatically learned from the documents in both the current interval and the recent intervals to retain the overlap between new documents and old documents [12]. By doing this, the evolution analysis would be more smooth and robust. Thus a time window is defined for each time interval. It contains the current interval and a fixed number* $l - 1$ *of the recent intervals. The themes in a time interval is learned from the documents in its corresponding time window. Take* $l = 3$ *as an example, in Fig. 1 (right), the themes in interval* T_3 *is trained from the documents in the time window of* T_1,T_2 *and* T_3*; and the themes in* T_4 *is trained from the documents in a time window of* T_2,T_3 *and* T_4.

Definition 4 (Cross-lingual Evolutionary Transition). *Following [12], we utilize* θ^{T_i} *to denote a theme with a time stamp in the cross-lingual data. Suppose a theme* $\theta_a^{T_{i-1}}$ *and a theme* $\theta_b^{T_i}$, *if the similarity between them is larger than a threshold, it is defined that there is a cross-lingual evolutionary transition from* $\theta_a^{T_{i-1}}$ *to* $\theta_b^{T_i}$.

Definition 5 (Cross-lingual Theme Evolutionary Graph). *Following [12], the cross-lingual theme evolutionary graph* $G = (N, E)$ *is defined as a weighted graph as shown in Fig. 1 (right). In the graph, each node* $v \in N$ *denotes a theme with a time stamp; and each edge* $e \in E$ *denotes whether the two nodes in adjacent time intervals have an evolutionary transition. The weight of an edge (denoted by the thickness of the edge), stands for the similarity value between the current theme and the evolved theme. The larger the similarity is, the thicker the edge would be.*

2.2 The Task of Cross-Lingual ETP Analysis

Following the idea in [12], we define the task of cross-lingual ETP analysis as to draw a cross-lingual theme evolutionary graph from the unstructured data stream. The result of ETP analysis would give a clear temporal hierarchical overview of the themes and their evolutions for different events, which would serve for both users and other intelligent services.

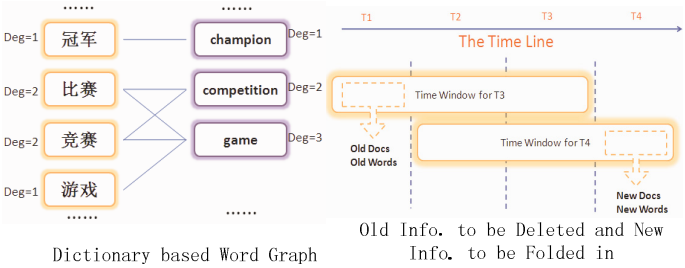


Fig. 2. Illustration for online cross-lingual PLSI

3 General Framework

The general framework for cross-lingual ETP analysis is a two-step process. The first step is to extract the themes in each time interval from their time windows; and the second step is to construct the evolutionary transitions among the themes in adjacent time intervals.

The second step is more intuitive by calculating the KL-divergence $D(\theta_2||\theta_1)$ between two themes [12]. In the cross-lingual case, it is defined as

$$D(\theta_2||\theta_1) = \sum_{w_i \in W_1 \bigcup W_2 \bigcup \cdots \bigcup W_s} p(w_i|\theta_2) \log \frac{p(w_i|\theta_2)}{p(w_i|\theta_1)}. \quad (1)$$

The main difference compared with [12] is that the cross-lingual theme contains words from different languages.

The first step is the key point that would be discussed in this paper. A naive way is to re-train all the documents in the time window in a batch mode. For example, the cross-lingual PLSI [15] model could be directly utilized here for theme extraction. However, as discussed in previous sections, it is time-consuming to re-train all the latent themes for each time interval; and the latent themes between two adjacent time intervals might lose continuity. Therefore, in this paper, we combine the idea in [15] and [4] to propose an online cross-lingual PLSI model for cross-lingual ETP analysis, which would be illustrated in detail in the following section.

4 Online Cross-Lingual PLSI

4.1 Cross-Lingual PLSI in Batch Mode

The cross-lingual PLSI [15] model is described by two kinds of parameters, $p(\theta|d)$ and $p(w|\theta)$. θ denotes the theme; d denotes the document; and w denotes the word.

The learning process of $p(\theta|d)$ and $p(w|\theta)$ is to optimize the linear combination of 1) the log-likelihood of the data and 2) the constraint function of cross-lingual connection, denoted as

$$F = (1 - \lambda)L(C) + \lambda R(C), \tag{2}$$

where

$$L(C) = \sum_{i=1}^{s} \sum_{d \in C_i} \sum_{w} c(w, d) log \sum_{j=1}^{k} p(\theta_j|d)p(w|\theta_j), \tag{3}$$

$$R(C) = -\frac{1}{2} \sum_{<w_u, w_v> \in E} \sum_{j=1}^{k} (\frac{p(w_u|\theta_j)}{Deg(w_u)} - \frac{p(w_v|\theta_j)}{Deg(w_v)})^2. \tag{4}$$

$L(C)$ is the log-likelihood of the data in multiple languages. In constructing the constraint function $R(C)$, a multi-partite undirected graph $G =< W, E >$ is built from each bilingual dictionaries as shown in Fig. 2 (left). Each node in W denotes a word. If two words from different languages, e.g., w_u and w_v, could interpret each other, there would be an edge between them. $Dev(w_u)$ denotes the degree of the word w_u. From the definition, by optimizing the function $R(C)$, words from multiple languages that have similar meanings, would have similar probabilities within a theme.

The optimization is based on the EM algorithm as follows.

1. E-Step

$$P(\theta|w, d) = \frac{P(w|\theta)P(\theta|d)}{\sum_{\theta'} P(w|\theta')P(\theta'|d)} \tag{5}$$

2. M-Step

$$P(w|\theta) = \frac{\sum_{i=1}^{s} \sum_{d \in C_i} c(w, d)P(\theta|w, d)}{\sum_{i=1}^{s} \sum_{d \in C_i} \sum_{w' \in d} c(w', d)P(\theta|w', d)} \tag{6}$$

$$p(\theta|d) = \frac{\sum_{w \in d} c(w, d)P(\theta|w, d)}{\sum_{w \in d} c(w, d)} \tag{7}$$

$$p^{(t+1)}(w_u|\theta_j) = (1 - \alpha)p^{(t)}(w_u|\theta_j) + \alpha \sum_{<w_u, w_v> \in E} \frac{1}{Deg(w_v)} p^{(t)}(w_v|\theta_j) \tag{8}$$

4.2 Proposed Cross-Lingual PLSI in Online Mode

Following the idea in [4], the process of the online cross-lingual PLSI is shown in Fig. 2 (right), including 1) discarding old documents and old terms; 2) folding in new documents and new terms and 3) updating the PLSI parameters. Compared with the online monolingual PLSI in [4], in this paper, step 2 and 3 are extended in order to fit the cross-lingual data. Therefore, in this section, we would focus these two steps. The first step could be referred in [4].

1. Fold in new documents and terms. The target of folding in new documents and terms is to estimate $P(\theta|d_{new})$ and $P(\theta|w_{new})$. The difference of our work from previous work [4] is that the process should be extended for incorporating the cross-lingual word relations. Because words in different languages that have similar meanings should have more probability to be assigned to the same theme. In the batch learning mode, this problem has been solved by a global update in Eq. 8; in the online learning mode, this would be solved by a local update. Consequently, we add an local update step as an extension of previous updating process. The whole process is conducted as follows by EM algorithms. By fixing the parameter of $P(w|\theta)$, $P(\theta|d_{new})$ could be estimated by iteration calculation of

$$P(\theta|w, d_{new}) = \frac{P(w|\theta)P(\theta|d_{new})}{\sum_{\theta' \in \Theta} P(w|\theta')P(\theta'|d_{new})}, . \tag{9}$$

$$P(\theta|d_{new}) = \frac{\sum_{w \in d_{new}} c(w, d_{new})P(\theta|w, d_{new})}{\sum_{\theta \in \Theta} \sum_{w \in d_{new}} c(w, d_{new})P(\theta'|w, d_{new})}. \tag{10}$$

After $P(\theta|d_{new})$ is estimated, $P(d_{new}|\theta)$ could be estimated by

$$P(d_{new}|\theta) = \frac{\sum_{w \in d_{new}} c(w, d_{new})P(\theta|w, d_{new})}{\sum_{d \in D_{new}} \sum_{w \in d} c(w, d)P(\theta|w, d)}, \tag{11}$$

where D_{new} is all the documents in the new time window, including documents in the remaining intervals and the new interval. By fixing the parameter of $P(d_{new}|\theta)$, $P(\theta|w_{new})$ could be estimated by iteration calculation of

$$P(\theta|w_{new,d_{new}}) = \frac{P(d_{new}|\theta)P(\theta|w_{new})}{\sum_{\theta' \in \Theta} P(d_{new}|\theta')P(\theta'|w_{new})}, \tag{12}$$

$$P(\theta|w_{new}) = \frac{\sum_{d \in D_{new}} c(w_{new}, d)P(\theta|w_{new}, d)}{\sum_{d' \in D_{new}} c(w_{new}, d')}. \tag{13}$$

$$P'(\theta|w_{new}) = (1 - \alpha)P(\theta|w_{new}) + \alpha \sum_{<w_{new}, w_v> \in E} \frac{1}{Deg(w_{new})} P(\theta|w_v) \tag{14}$$

In the last equation, the cross-lingual smoothing is incorporated following the idea of [15]. Consequently, the similar words in different languages would have similar latent theme distributions.

2. Update the PLSI parameters. In the folding in process, $p(w_{old}|\theta)$ has not been changed and $p(w_{new}|\theta)$ has not been incorporated. Thus a normalization is utilized following [4] as

$$P(w|\theta) = \frac{\sum_{d \in D_{new}} c(w, d)P(\theta|w, d)}{\sum_{d' \in D_{new}} \sum_{w' \in d'} c(w', d')P(\theta|w', d')}. \tag{15}$$

If calculation complexity is limited in a small range, the updating has been finished. But if more calculation is allowed, Eq. 5, Eq. 6, Eq. 7 and Eq. 8 could be executed iteratively to optimize the parameters further.

5 Experiments

The experiments are targeted at justifying the following issues for the proposed online cross-lingual PLSI model.

1. What is its overall performance in the ETP analysis?
2. Whether it is efficient in reducing the time complexity?
3. Whether it is effective in solving the continuity limitation?

To issue 1, we show an intuitive example of the ETP analysis result; and we also compare it with a version of monolingual topic model by translating multiple languages into a single language, in order to show the advantages of cross-lingual. To issue 2, we compare the speed of convergence of the proposed model with the original cross-lingual model in batch mode. To issue 3, we give both qualitative and quantitative justification compared with the original cross-lingual model in batch mode.

5.1 Dataset

Our dataset is a set of news articles, which is collected between June 1st 2012 to June 23th 2012. The articles are either in English or in Chinese. Totally, 269,144 Chinese articles and 64,897 English articles are collected. A pre-processing is conducted before the experiments, including splitting words, removing the stop words and stemming. We choose mandarintools[1] to build the bilingual word graph as shown in Fig. 2 (left).

The dataset is divided into three time intervals evenly, denoted as T_1, T_2 and T_3. The time window length is $l = 2$. Thus the time window for T_2 include T_1 and T_2; and the time window for T_3 is T_2 and T_3. When documents of T_3 arrive, the topic model should discard the old words and documents in T_1 and fold in the new words and documents in T_3. In the proposed online PLSI, this process is natural. In the original PLSI in batch mode, the model should be re-trained using the documents in T_2 and T_3.

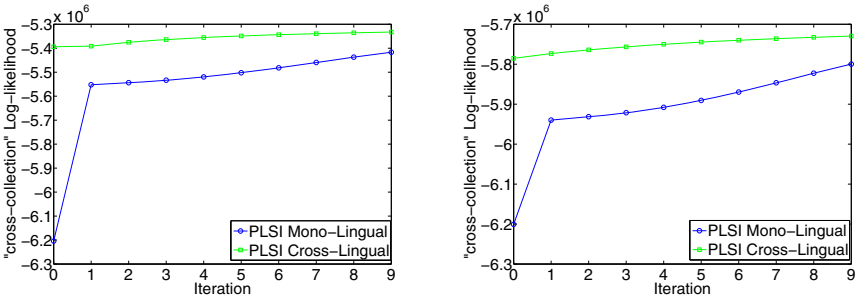
5.2 Overall Performance

To evaluate the overall performance, we utilize the proposed model to extract the themes in T_2 and T_3. Then the cross-lingual evolutionary transitions are generated by calculating the KL-divergence using Eq. 1. A cross-lingual theme evolutionary graph is drawn like Fig. 1 (right). Due to the space limitation, we do not demonstrate the whole graph, but we show part of it instead for analysis.

[1] `mandarintools.com`

Table 1. Overall performance example

Theme extraction in T_2		Theme extraction in T_3 by PLSI in batch mode		Theme extraction in T_3 by the proposed model	
Theme 2.1	Theme 2.2	Theme 3.1'	Theme 3.2'	Theme 3.1	Theme 3.2
England 卡希尔(Cahill) Terry France 保罗(Paul) 揭幕(opening) 热身赛 (warm-up match) Hodgson Curtis 球衣(jersey) 瑞典(Sweden)	Italy Spain 克罗地亚(Croatia) 进球数(goals scored) 球迷(fan) Cassano 皮尔洛(Pirlo) 托雷斯(Torres) player 集训(train) 预测(predict)	警方(police) 酒店(hotel) worker sex 话题(topic) 中国(China) 种族歧视 (racial discrimination) country watch 乌克兰(Ukraine)	coach 乌克兰(Ukraine) 卡希尔(Cahill) 德国(Germany) Spain 意大利(Italian) 英格兰(England) 大战(contest) 出线(advance) knock	England 球迷(fan) 乌克兰(Ukraine) hotel Spain 杰拉德(Gerrard) 鲁尼(Rooney) sex 底线(end line) team 出线(advance)	Italy Ireland 克罗地亚(Croatia) 巴洛特利(Balotelli) 默契(tacit) barbecue 赌球(gambling) 积分(score) 球衣(jersey) 卡萨诺(Cassano) final



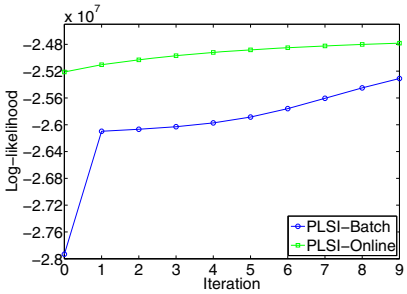
(a) $\alpha = 0.2$ (b) $\alpha = 0.8$

Fig. 3. Comparison with monolingual topic model

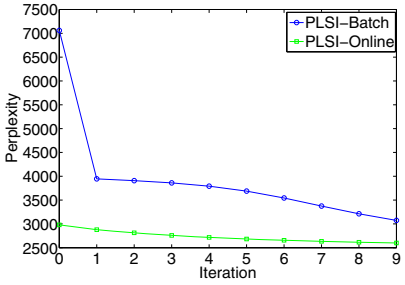
We choose the event of Euro 2012 as an example. Table. 1 shows the ETP analysis result. Column 1 shows the theme extraction result for T_2; Column 2 shows the theme extraction result for T_3 from cross-lingual PLSI in batch mode, which would be utilized to show the improvement of continuity later; and Column 3 shows the theme extraction result for T_3 from the proposed model. *Theme* 3.1 is evolved from *theme* 2.1; and *theme* 3.2 is evolved from *theme* 2.2.

It can be observed that in *theme* 2.1, stories are related to the England team, such as the stars, the opponent team, etc. The stories are about warm-up matches, openings, etc. After being evolved to *theme* 3.1, stories are also related to England team. But as time goes by, the content of the stories has been changed to whether the England team can advance, the behavior of its star Rooney, etc. *Theme* 2.2 and *theme* 3.2 have similar theme evolution, but they are related to the Italy team. We can also see that the English media is more open in sensitive topics such as "sex" than the Chinese media, which shows the complementarity of information among the cross-lingual data. This example demonstrates that proposed model is effective in cross-lingual ETP tasks.

To show the advantage of cross-lingual topic model, we compare it with the monolingual PLSI by translating the Chinese documents into English documents. The experimental setup for translation can be referred to [15]. For evaluation,



(a) log-likelihood

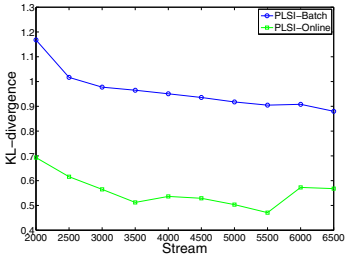


(b) perplexity

Fig. 4. Comparison with cross-lingual PLSI in batch mode (1)

#Theme	PLSI-Batch Convergence time (ms)	PLSI-Online Convergence time (ms)
8	190300	42375 (22.2%)
16	504269	91857 (18.2%)
24	796977	145785 (18.3%)

(a) execution time



(b) continuity

Fig. 5. Comparison with cross-lingual PLSI in batch mode (2)

we utilize the "cross-collection" log-likelihood defined in [15]. Fig. 3 shows the experimental result for different parameters. α is the parameter used in Eq. 8 and Eq. 14. It could be observed that the cross-lingual topic model constantly outperform the monolingual topic model in each iteration. This is because the cross-lingual topic model could well smooth the connection between different languages. The result is consistent with the result in [15].

5.3 Justification for Reducing the Time Complexity

In order to justify the performance of reducing the time complexity, we compare our proposed model with the original cross-lingual PLSI [15] in batch mode. Following the evaluation metrics in [1, 4], we utilize the log-likelihood and perplexity to evaluate the performance of the model in each iteration. For perplexity, the smaller the value, the better the performance. The definition of the metrics could be found in [1, 4]. For the proposed online model, the parameters are updated from the parameters of the previous model; for the original model in batch mode, the parameters are randomly allocated at first and are re-trained thoroughly.

Fig. 4 and Fig. 5(a) shows the experimental results. It could be observed that the proposed online model converges much faster than the model in batch mode,

in both log-likelihood and perplexity metrics. From Fig. 5(a), the convergence time of the proposed model is around 20% of the original model's, which is consistent with the result in [4] (15% to 20% of the batch mode in monolingual case). This demonstrates that the proposed model is efficient for reducing the time complexity in folding in the new documents and words.

5.4 Justification for Solving the Continuity Limitation

In qualitative analysis, we show an example of the ETP analysis from the original PLSI [15] in batch mode in Table. 1. We can see the difference between this model and our proposed model. The themes in T_2 are divided by different teams. One for England and one for Italy. After the evolution in T_3, the themes extracted by our proposed model are also divided by different teams. However, the themes extracted by the original PLSI in batch mode are divided by other dimensions. *theme* $3.1'$ is about the activities of fans; and *theme* $3.2'$ is about the activities of the game. This example demonstrates directly that the proposed model performs better in keeping the continuity.

In quantitative analysis, following the idea in [4], we utilize the average KL divergence rate of the two closest latent variables in two adjacent time windows to quantify the continuity performance. The detailed definition of this metric could be found in [4]. Fig. 5(b) shows the performance of our proposed model and the original model in batch mode. It could be observed that the performance of the proposed model outperforms the original model constantly. The average distance of our proposed model is 56% of the original model's.

From above justifications, it can be concluded that the proposed model is effective in solving the continuity limitation.

6 Related Work

The primary work related to ETP analysis in history is the topic detection and tracking (TDT) task [14]. The main difference of the TDT tasks from the ETP analysis is that TDT is in the document level, while ETP is in the word level. Recent work of ETP analysis include [5, 8, 10–12]. The main difference from our work is that our model can model cross-lingual themes. Topic model is the key technique in this paper. Two representatives of topic model include PLSI [7] and LDA [2]. Many variances of topic models are presented over these years, including the online learning [1, 4, 6, 8]; the cross-lingual modeling [3, 9, 13, 15], and etc.

7 Conclusion and Future Work

In this paper, we propose an online cross-lingual PLSI model and utilize this model in the ETP analysis tasks. Experimental verification from a real world dataset demonstrates that the proposed model performs well. It is efficient to reduce the training time in folding in new documents; and the proposed model is effective in keeping the continuity property of themes in different time intervals.

In the future, we will try to combine previous online LDA and cross-lingual LDA model in ETP analysis. We believe through the regularization techniques in LDA, the performance would obtain another improvement.

Acknowledgments. The work described in this paper was fully supported by the National Basic Research Program of China (973 Program, Grant No. 2013CB329605). The authors also would like to thank the reviewers for their helpful comments.

References

1. AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of ICDM 2008, pp. 3–12. IEEE (2008)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022 (2003)
3. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Proceedings of UAI 2009, pp. 75–82. AUAI Press (2009)
4. Chou, T.C., Chen, M.C.: Using incremental plsi for threshold-resilient online event analysis. IEEE Transactions on Knowledge and Data Engineering 20(3), 289–299 (2008)
5. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, L.: Detecting topic evolution in scientific literature: how can citations help? In: Proceeding of CIKM 2009, pp. 957–966. ACM (2009)
6. Hoffman, M.D., Blei, D.M., Bach, F.: Online learning for latent dirichlet allocation. In: Proceedings of NIPS 2010, vol. 23, pp. 856–864 (2010)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of SIGIR 1999, pp. 50–57. ACM (1999)
8. Iwata, T., Yamada, T., Sakurai, Y., Ueda, N.: Online multiscale dynamic topic models. In: Proceedings of KDD 2010, pp. 663–672. ACM (2010)
9. Jagarlamudi, J., Daumé III, H.: Extracting multilingual topics from unaligned comparable corpora. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 444–456. Springer, Heidelberg (2010)
10. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proceedings of the KDD 2003, vol. 7, pp. 373–397 (2003)
11. Lin, C.X., Mei, Q., Han, J., Jiang, Y., Danilevsky, M.: The joint inference of topic diffusion and evolution in social communities. In: Proceedings of ICDM 2011, pp. 378–387. IEEE (2011)
12. Mei, Q., Zhai, C.X.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: Proceedings of the KDD 2005, pp. 198–207. ACM (2005)
13. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Cross lingual text classification by mining multilingual topics from wikipedia. In: Proceedings of WSDM 2011, pp. 375–384. ACM (2011)
14. Wang, C., Zhang, M., Ma, S., Ru, L.: Automatic online news issue construction in web environment. In: Proceedings of WWW 2008, pp. 457–466. ACM (2008)
15. Zhang, D., Mei, Q., Zhai, C.X.: Cross-lingual latent topic extraction. In: Proceedings of ACL 2010. Association for Computational Linguistics, pp. 1128–1137 (2010)