International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

Meaningful Data Extraction Using Data Correlation Technique

B. Uma*, S. Shanawaz Basha**, G. Sesha Phaneendra Babu**
*M.Tech(C.S.E) ,IPCET(W), Kurnool,
**Assistant Professor, AVR & SVR Engg College, Kurnool,
** Assistant Professor, INTELL ENGINEERING COLLEGE, Anantapur .

Abstract

Time marked texts, or text sequences, are everywhere in real-world applications. several text sequences are often related to each other by sharing common topics. The correlation among these sequences provides more meaningful and comprehensive clues for topic mining than those from each individual sequence. It is nontrivial to explore the correlation with the existence of asynchronism among multiple sequences, i.e., documents from different sequences about the same topic may have different timestamps. In this paper, we address this problem and provide a novel algorithm based on the generative topic model. Our algorithm consists of two steps: In the first step we take out common topics from various sequences depending on the adjusted time stamps from the second step; the second step adjusts the time stamps of the documents according to the time distribution of the topics discovered by the first step. We perform these two steps repetitively our objective function can be guaranteed.

Index Terms— Asynchronous sequences, temporal correlation, text mining, time synchronization.

1. INTRODUCTION

Text mining, also referred to as text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

In many real-world applications, we are facing multiple text sequences that are correlated with each other by sharing common topics. Intuitively, the interactions among these sequences could provide clues to derive more meaningful and comprehensive topics than those found by using information from each individual stream solely[4].

Asynchronism among multiple sequences, on the same topic have different time stamps, is actually very common in practice. For instance, in news feeds, there is no guarantee that news articles covering the same topic are indexed by the same time stamps. An example is research paper archives, where the latest research topics are closely followed by newsletters and communications within weeks or months, then the full versions may appear in conference proceedings, which are usually published annually and at last in journals, which may sometimes take more than a year to appear after submission. To visualize it, we have the relative frequency of the occurrences of two terms warehouse and mining, respectively, in the titles of all research papers here we deal with text sequences that share common topics yet are temporally asynchronous.

2. Background and Related Work

Topic mining has been extensively studied in the literature, starting with the Topic Detection and Tracking (TDT) In many real applications, text collections carry generic temporal information and, thus, can be considered as text sequences. To capture the temporal dynamics of topics, various methods have been proposed to discover topics over time in text sequences In [1], the authors introduced hyper-parameters that evolve over time in state transfer models in the sequence. For each time slice, a hyperparameter is assigned with a state by a probability distribution, given the state on the former time slice. The time dimension of the sequence was cut into time slices and topics were discovered from documents in each slice independently. As a result, in multiple-sequence cases, topics in each sequence can only be estimated separately and potential correlation between topics in different sequences, both semantically and temporally, could not be fully explored. We also note that there is a whole literature on similarity measure between time series (sequences). Various similarity functions have been proposed, many of which addressed the asynchronous nature between time series[2] [3]. However, defining an asynchronism-robust similarity measure alone does not necessarily solve our problem. In fact, most of the similarity measures deal with synchronism implicitly, rather than fix the asynchronism explicitly, like what we do in this work.

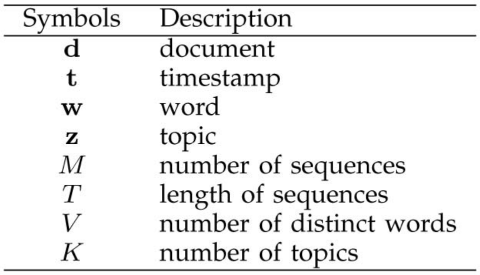
3. Algorithm

We implement this method in two deferent steps. The outline of our algorithm is:

. **Step 1**. We assume that the current time stamps of the sequences are synchronous and extract common topics from them.

. **Step 2**. We synchronize the time stamps of all documents by matching them to most related topics, respectively. Then, we go back to Step 1 and iterate until convergence.

The main symbols used throughout the paper are listed in Table 1.

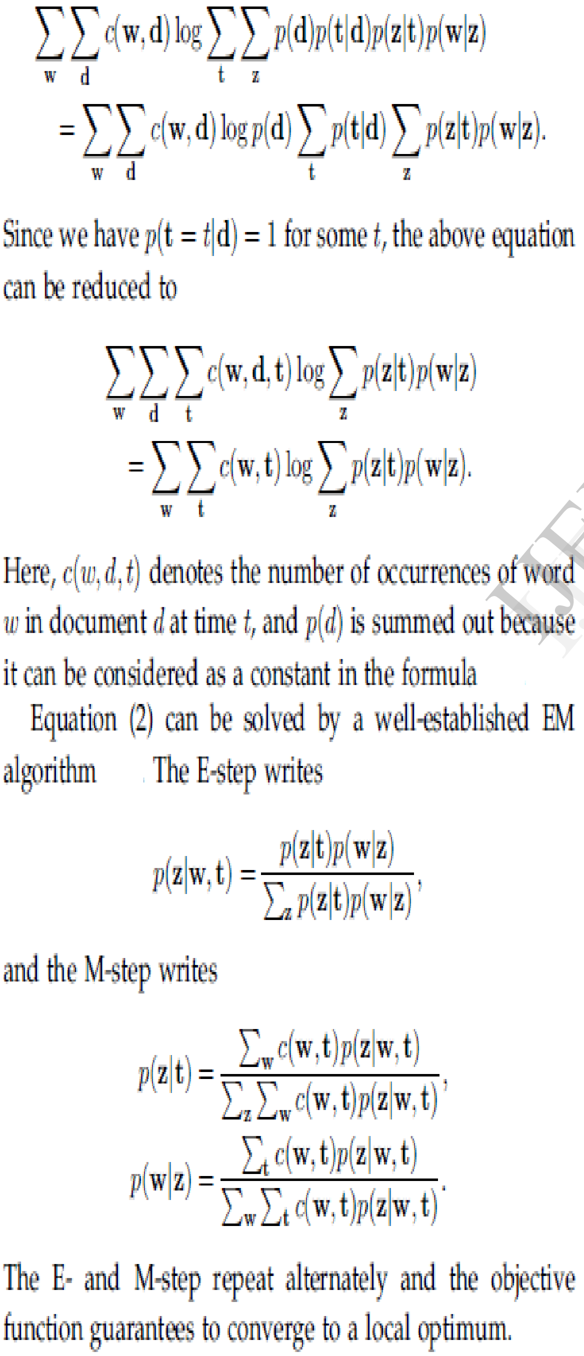


Symbols	Description
$\mathbf{d}$	document
$\mathbf{t}$	timestamp
$\mathbf{w}$	word
$\mathbf{z}$	topic
M	number of sequences
T	length of sequences
V	number of distinct words
K	number of topics

TABLE 1
Symbols and Their Meanings

Topic Extraction

we assume the current time stamps of all sequences are already synchronous and extract common topics from them. In other words, now p(t|d) is fixed and we try to maximize the likelihood function by adjusting p(t|z) and p(w|z). Thus, we can rewrite the likelihood function as follows:



$$\sum_{w}\sum_{d} c(w,d)\log\sum_{t}\sum_{z} p(d)p(t|d)p(z|t)p(w|z)$$

$$=\sum_{w}\sum_{d} c(w,d)\log p(d)\sum_{t} p(t|d)\sum_{z} p(z|t)p(w|z).$$

Since we have $p(t = t|d) = 1$ for some t, the above equation can be reduced to

$$\sum_{w}\sum_{d}\sum_{t} c(w,d,t)\log\sum_{z} p(z|t)p(w|z)$$

$$=\sum_{w}\sum_{t} c(w,t)\log\sum_{z} p(z|t)p(w|z).$$

Here, $c(w,d,t)$ denotes the number of occurrences of word w in document d at time t, and $p(d)$ is summed out because it can be considered as a constant in the formula

Equation (2) can be solved by a well-established EM algorithm The E-step writes

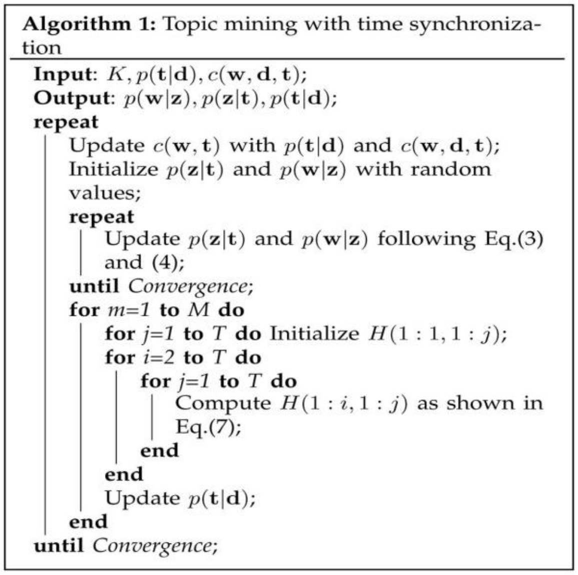
$$p(z|w,t)=\frac{p(z|t)p(w|z)}{\sum_{z} p(z|t)p(w|z)},$$

and the M-step writes

$$p(z|t)=\frac{\sum_{w} c(w,t)p(z|w,t)}{\sum_{z}\sum_{w} c(w,t)p(z|w,t)},$$

$$p(w|z)=\frac{\sum_{t} c(w,t)p(z|w,t)}{\sum_{w}\sum_{t} c(w,t)p(z|w,t)}.$$

The E- and M-step repeat alternately and the objective function guarantees to converge to a local optimum.



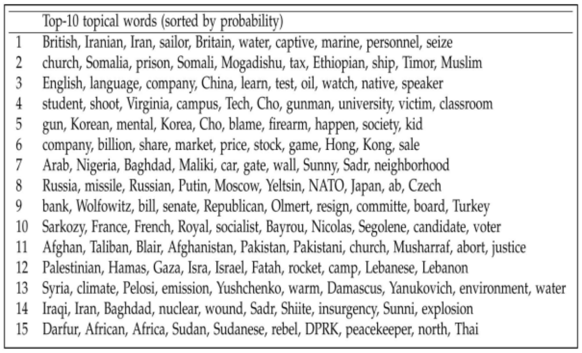
```
Algorithm 1: Topic mining with time synchroniza-
tion
  Input: $K, p(\mathbf{t}|\mathbf{d}), c(\mathbf{w},\mathbf{d},\mathbf{t})$;
  Output: $p(\mathbf{w}|\mathbf{z}), p(\mathbf{z}|\mathbf{t}), p(\mathbf{t}|\mathbf{d})$;
  repeat
      Update $c(\mathbf{w},\mathbf{t})$ with $p(\mathbf{t}|\mathbf{d})$ and $c(\mathbf{w},\mathbf{d},\mathbf{t})$;
      Initialize $p(\mathbf{z}|\mathbf{t})$ and $p(\mathbf{w}|\mathbf{z})$ with random
      values;
      repeat
          Update $p(\mathbf{z}|\mathbf{t})$ and $p(\mathbf{w}|\mathbf{z})$ following Eq.(3)
          and (4);
      until Convergence;
      for $m$=1 to $M$ do
          for $j$=1 to $T$ do Initialize $H(1:1,1:j)$;
          for $i$=2 to $T$ do
              for $j$=1 to $T$ do
                  Compute $H(1:i,1:j)$ as shown in
                  Eq.(7);
              end
          end
          Update $p(\mathbf{t}|\mathbf{d})$;
      end
  until Convergence;
```

Fig. 1 Topic Mining with time Synchronization.

The computational complexity of the topic extraction step (with the EM algorithm) is O(KVT) while the complexity of time synchronization step is approximately O(VMT^{3}). Thus, the overall complexity of our algorithm is O(VT(K + MT^{2})), where V is the size of vocabulary, T the number of different time stamps, K the number of topics, and M the number of sequences. If we take V , K and M as constants and only consider the length of sequence, which is T, the complexity of Algorithm 1 becomes O(T^{3}). We will show in the next section how to reduce it to O(T^{2}) with a local search strategy.

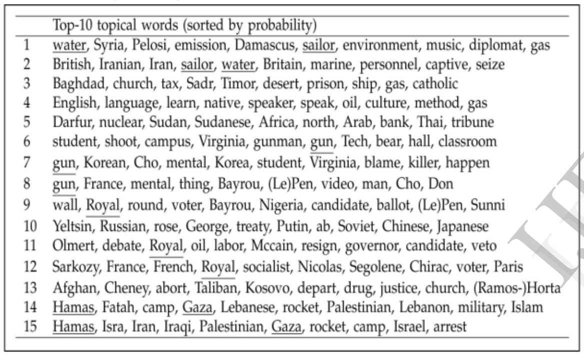
4. Performance Analysis

Here the performance of our method on the news feeds data set. We extracted 15 common topics (K = 15) from two news feeds consisting of 61 days' news reports with full texts. The local search radius was set to be 3, as we assumed that time difference between (online) news articles belonging to the same topic normally will not exceed 3 days. The topic extraction step remained the same. We list in Fig. 2 the topical words of all 15 common topics extracted by our method (sync) and those by the baseline method (no_sync) in Fig. 3 Comparing these two sets of results, we can see that both methods discovered some common topics in the sequences, e.g., British sailors captured in Iran, Campus shooting at VT, France presidential election, Darfur problem,etc.



	Top-10 topical words (sorted by probability)
1	British, Iranian, Iran, sailor, Britain, water, captive, marine, personnel, seize
2	church, Somalia, prison, Somali, Mogadishu, tax, Ethiopian, ship, Timor, Muslim
3	English, language, company, China, learn, test, oil, watch, native, speaker
4	student, shoot, Virginia, campus, Tech, Cho, gunman, university, victim, classroom
5	gun, Korean, mental, Korea, Cho, blame, firearm, happen, society, kid
6	company, billion, share, market, price, stock, game, Hong, Kong, sale
7	Arab, Nigeria, Baghdad, Maliki, car, gate, wall, Sunny, Sadr, neighborhood
8	Russia, missile, Russian, Putin, Moscow, Yeltsin, NATO, Japan, ab, Czech
9	bank, Wolfowitz, bill, senate, Republican, Olmert, resign, committe, board, Turkey
10	Sarkozy, France, French, Royal, socialist, Bayrou, Nicolas, Segolene, candidate, voter
11	Afghan, Taliban, Blair, Afghanistan, Pakistan, Pakistani, church, Musharraf, abort, justice
12	Palestinian, Hamas, Gaza, Isra, Israel, Fatah, rocket, camp, Lebanese, Lebanon
13	Syria, climate, Pelosi, emission, Yushchenko, warm, Damascus, Yanukovich, environment, water
14	Iraqi, Iran, Baghdad, nuclear, wound, Sadr, Shiite, insurgency, Sunni, explosion
15	Darfur, African, Africa, Sudan, Sudanese, rebel, DPRK, peacekeeper, north, Thai

Fig. 2. Common topics extracted by our method (sync) from news feeds (K =15).



	Top-10 topical words (sorted by probability)
1	water, Syria, Pelosi, emission, Damascus, sailor, environment, music, diplomat, gas
2	British, Iranian, Iran, sailor, water, Britain, marine, personnel, captive, seize
3	Baghdad, church, tax, Sadr, Timor, desert, prison, ship, gas, catholic
4	English, language, learn, native, speaker, speak, oil, culture, method, gas
5	Darfur, nuclear, Sudan, Sudanese, Africa, north, Arab, bank, Thai, tribune
6	student, shoot, campus, Virginia, gunman, gun, Tech, bear, hall, classroom
7	gun, Korean, Cho, mental, Korea, student, Virginia, blame, killer, happen
8	gun, France, mental, thing, Bayrou, (Le)Pen, video, man, Cho, Don
9	wall, Royal, round, voter, Bayrou, Nigeria, candidate, ballot, (Le)Pen, Sunni
10	Yeltsin, Russian, rose, George, treaty, Putin, ab, Soviet, Chinese, Japanese
11	Olmert, debate, Royal, oil, labor, Mccain, resign, governor, candidate, veto
12	Sarkozy, France, French, Royal, socialist, Nicolas, Segolene, Chirac, voter, Paris
13	Afghan, Cheney, abort, Taliban, Kosovo, depart, drug, justice, church, (Ramos-)Horta
14	Hamas, Fatah, camp, Gaza, Lebanese, rocket, Palestinian, Lebanon, military, Islam
15	Hamas, Isra, Iran, Iraqi, Palestinian, Gaza, rocket, camp, Israel, arrest

Fig. 3. Common topics extracted by the baseline method (no_sync) from news feeds (K = 15). Underlined are duplicated topical words.

Fig. 4 proves in quantity that topics extracted by our method (sync) are much more discriminative to each other than those extracted by the baseline method (no_sync).

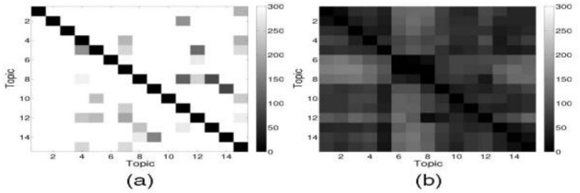


Fig. 4. The pairwise KL-divergence between topics extracted from the news feeds (K = 15). (a) sync. (b) no_sync.

Figs. 5 and 6 show how our method adjusted the time stamps of documents in both news sequences, which is consistent to its behavior on literature repositories: it automatically discovered documents related to the same topic after considering their semantic as well as temporal information and then assigned them to the same time stamp.

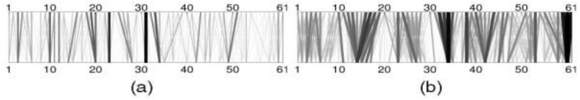


Fig. 5. The mapping from documents' original time stamps (upper axis) to those determined by our method (lower axis) in news streams. (a) IHT. (b) People.

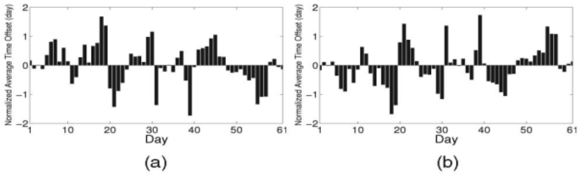


Fig. 6. Normalized average time offset of news articles at each day. (a) IHT. (b) People.

5. Conclusion

In this paper we undertake the problem of mining common topics from multiple asynchronous text sequences. Here we propose a different method which can automatically discover and fix potential synchronism among sequences and consequentially extract better common topics. The key idea of our method is to introduce a self-refinement process by utilizing correlation between the semantic and temporal information in the sequences. Results suggest that the performance of our method is robust and stable against different parameter settings and random initialization.

6. Reference

[1] D.M. Blei and J.D. Lafferty, "Dynamic Topic Models," Proc. Int'l Conf. Machine Learning (ICML), pp. 113-120, 2006.

[2] D.J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," Proc. Knowledge Discovery in Databases (KDD) Workshop, pp. 359-370, 1994.

[3]H. Sakoe, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, no.1, pp. 43-49, Feb. 1978.

[4] X. Wang, C. Zhai, X. Hu, and R. Sproat, "Mining Correlated Bursty Topic Patterns from Coordinated Text Streams," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),pp. 784-793, 2007.