2011 International Conference on Asia Agriculture and Animal
IPCBEE vol.13(2011) © (2011)IACSIT Press, Singapore

Mining and Visualizing Historical Events Using Wikipedia and Google Map

Ping Dong, An-Chi Shi, Yuan-Cheng Yu and Liang-Chih Yu [+]

Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C.

Abstract. Historical events such as natural disasters, financial crisis, and new invention are common knowledge sources in social media such as Wikipedia. Internet users can search for a particular historical event through the support of information extraction and retrieval techniques. However, the users may have another need — query the events occurred in the same year worldwide. To acquire such knowledge, one possible way is to browse the events year by year, which is a labor intensive and time consuming process. Therefore, this work proposes a framework that can mine the historical events occurred in the same year. For visualization purpose, the mined historical events are displayed on the Google Map according to their locations. Such visualization information can provide an integrated view of temporal and spatial information of historical events.

Keywords: Historical events, data mining, data visualization, Wikipedia, Google map

1. Introduction

Historical events are common knowledge sources in social media. For example, Wikipedia has a large number and variety of historical events such as natural disasters, medical crisis, financial crisis, wars, political issues, as well as new invention. Internet users can search for historical events by the following two schemes: specifying a set of keywords relating to the events and browsing the events year by year. For the former one, the aim is usually to find the information relevant to a particular event. In this circumstance, information extraction and retrieval techniques can be used to facilitate the search process; that is, to help users obtain more precise event information more efficiently. For the later one, the users may have the need to query the events occurred in the same year worldwide. For instance, the Iraq War began in 2003. In the same year, the outbreak of Severe Acute Respiratory Syndrome (SARS) occurred, China launched its first human spacecraft, the ShenZhou-5, and a magnitude 8 earthquake occurred in Hokkaido. To acquire such knowledge, the users can browse the events year by year to find the events occurred in the same year. However, browsing and searching all events year by year is time consuming and tends to become overwhelming due to the large number of events in the web. Therefore, this work proposes a framework that can mine the historical events occurred in the same year. For visualization purpose, the mined historical events are displayed on the Google Map according to their locations. The users can benefit from such visualization information to simultaneously capture the temporal and spatial information of historical events. The aim of this work is summarized below.

- *Mine the historical events occurred in the same year*: The process begins with a user input of a particular historical event. The year of the input event is then extracted to find more events occurred in the same year.
- *Display the mined events on the Google Map*: Extract the locations for each mined event and the input event, and based on which to show all events on the Google Map.

[+] Corresponding author. Tel.: + 886-3-463-8800; fax: + 886-3-435-2077.
E-mail address: *lcyu@saturn.yzu.edu.tw*.

The rest of this work is organized as follows. Section 2 presents some related work. Section 3 describes the framework of historical event mining and visualization. Section 4 summarizes the experimental results. Conclusions are finally drawn in Section 5.

2. Related Work

The use of information extraction techniques can extract the specified events more precisely. Recent studies have demonstrated that different types of events can be extracted using different methods. For business events, Stevenson and Greenwood proposed a semantic approach to identify the relations between person name and company [1]. Yu et al. proposed a cascaded hybrid model to extract the detail information from a corpus of resumes [2]. In the biomedical domain, Wu et al. used a dependency-based method to identify depressive symptoms [3]. Their method considered the dependencies between each word token and its head in a sentence, thus yielding higher performance than the bag-of-words method. Yu et al. devised an evolutionary inference algorithm based on the Hyperspace Analog to Language (HAL) model to extract negative life events from psychiatry web corpora [4]. The HAL model constructs a high-dimensional context space to represent words, which provides an informative infrastructure for the evolutionary inference algorithm. Wu et al. used a data mining algorithm called association rule mining [5][6] to mine emotion events [7]. They manually constructed a set of emotion generation rules from psychology textbooks. These rules were then considered as the basis to generate more frequent emotion association patterns for identifying emotion events. Lan et al. and Bj¨orne et al. investigated the use of graph-based methods to extract biomedical events such as protein-protein interactions [8][9]. In the emerging field of social network, Agarwal and Rambow devised a new sequence kernel to mine social events such as human-human interaction [10]. The new sequence kernel considered both syntactic and semantic insights from the dependency trees of sentences, thus achieving better performance than previously proposed sequence kernels. Additionally, phrase, sentence, and document-level information can also be incorporated to mine terrorism events [11].

Once the events are identified, information retrieval can be applied in the later stage to help users find desired event information more efficiently. Information retrieval systems generally adopt a keyword-based approach to accomplish retrieval tasks [12]. That is, users formulate their information needs by using a set of keywords, based on which, the retrieval system returns a set of documents. In this scenario, both queries and documents are usually represented using a bag-of-words approach. Retrieval models, such as the vector space model (VSM) [13] and Okapi model [14][15] are then adopted to estimate the relevance between queries and documents. The VSM represents each query and document as a vector of words, and adopts the cosine measure to estimate their relevance. The Okapi model, which has been used on the Text REtrieval Conference (TREC) collections, developed a family of word-weighting functions for relevance estimation. These functions consider word frequencies and document lengths for word weighting. Both the VSM and Okapi models estimate the relevance by matching the words in a query with the words in a document. The above word-based methods are conceptually simple, easy to implement, and can achieve satisfactory results in many application domains. Other retrieval methods that incorporate high-level information can also improve retrieval results [16][17][18].

3. Framework of Historical Event Mining and Visualization

Figure 1 shows the overall framework of historical event mining and visualization. The framework begins with receiving a historical event specified by the user. The input event is represented as a set of keywords, which is used to find the event in the Wikipedia History database. Once the input event is found, the event year is extracted by the year extraction module. The extracted event year is then taken as input to find the other events occurred in the same year. For each event retrieved from the Wikipedia History database, the event locations are extracted by the location extraction module. Finally, all the events occurred in the same year are displayed on the Google Map according to their locations, so that the user can simultaneously capture the temporal and spatial information of historical events. In the following, we describe the detailed implementation of each module.

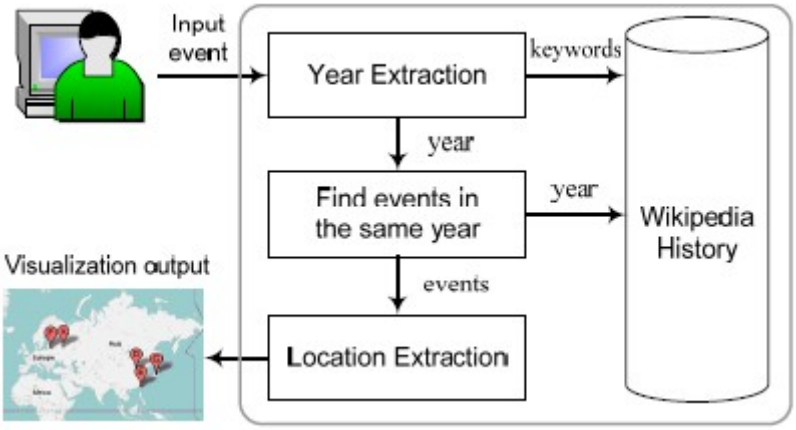

Fig. 1: Framework of historical event mining and visualization.

- **Year extraction**: A historical event may occur in a single year or last for many years. That is, the year extraction module may find more than one distinct year for an event in the Wikipedia History database. For instance, the event "Iraq War" (2003-2010) began in 2003 and lasted for seven years. To address this problem, we restrict the event year as the beginning year of an event. The year extraction procedure is formulated as follows. Let x be an input event and $Y = \{y_1, ..., y_n\}$ be the set of the event years of x found in the Wikipedia History database. The event year of x output by the year extraction module is defined as

$$Y_x = \begin{cases} y_1, & \text{if } n=1 \\ \min\{y_n\}, & \text{if } n>1 \end{cases} \quad (1)$$

  The above equation shows that in case of an event lasted for many years, the event year is the minimum value, i.e., beginning year, of the event years in the set.
- **Find events in the same year**: Once the event year of the input event is extracted, it is taken as the input to find the other events occurred in the same year from the Wikipedia History database. Table 1 lists some example events occurred in 2003, i.e., the beginning year of the input event "Iraq War". In these examples, we only show the event titles and ignore the detailed description.
- **Location extraction**: The locations of events in event titles are usually expressed in terms of nouns. Therefore, the first step of location extraction is the word segmentation and part-of-speech (POS) tagging. To accomplish this, we use the CKIP, which is developed by Academia Sinica, Taiwan (http://ckipsvr.iis.sinica.edu.tw). Additionally, only the words with the linguistic category Nc are considered as locations. The third column of Table 1 shows some example locations extracted from the event titles through POS filtering. The POS filtering method may suffer from some problems in location extraction. First, it may fail to identify the locations with the linguistic category other than Nc, which may decrease the recall of extraction. Second, not all Nc nouns are locations, which may decrease the precision of extraction. For instance, the noun "Columbia" in event B is the name of a space shuttle. Similarly, the noun "Liverpool" in event E is the name of a football team. These two issues will be investigated in experiments.

Table 1. Example events occurred in the same year with the input event "Iraq War".

No.	Event Title	Event Location
A	Iraq war	Iraq
B	The space shuttle Columbia disintegrated over Texas	Columbia, Texas
C	China's first human spacecraft, the ShenZhou-5, successfully launched.	China
D	World's first tongue transplant performed at Vienna's General Hospital	Vienna
E	English Premier League Liverpool visited Hong Kong	Liverpool, Hong Kong

F	2003 Hokkaido earthquake	Hokkaido

Fig. 2: Visualization of events on the Google Map.

4. Experimental Results

In experiments, we first show the visualization results of events on the Google Map, and then evaluate the performance of historical event mining. Figure 2 shows the visualization results of the events presented in Table 1. All events are displayed according to their locations. Additionally, in case of events with multiples locations (e.g., event B and E), each location will be displayed. In case of locations with the country name only (e.g., event A and C), the capital of the country will be displayed. In evaluation of the performance of event mining, the perfect case is that all events occurred in the same year can be exactly extracted and displayed using the POS filtering method. The evaluation metrics used herein include *recall*, *precision*, and *F-measure*, defined as

$$Recall = \frac{\text{number of correct events found by the method}}{\text{total number of events occurred in the same year}}, \tag{2}$$

$$Precision = \frac{\text{number of correct events found by the method}}{\text{total number of events found by the method}}, \tag{3}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}. \tag{4}$$

Table 2 shows the performance of historical event mining using the POS filtering method. The test events are “Iraq War” (Q1), “Three Gorges Dam” (Q2), and “Kobe earthquake” (Q3). The results show that among the three test queries, Q2 had the highest recall and F-score, and Q3 had the highest precision. Additionally, the recall of the POS filtering method was much higher than the precision, and the average F-measure was 62.31%. Although the use of the linguistic category Nc can find 75.61% historical events in average, some events were still misidentified because there is no explicit description of locations in the event titles. The precision is low, 52.99% in average, indicating that about half of the Nc nouns in the event titles were not actual locations.

Table 2. Performance of historical event mining

	Recall (%)	Precision (%)	F-measure (%)
Q1: Iraq War	73.20	50.68	59.89
Q2: Three Gorges Dam	84.78	51.32	63.94

Q3: Kobe earthquake	74.47	64.81	58.61
Average	**75.61**	**52.99**	**62.31**

5. Conclusions

This work presents a framework for historical event mining and visualization. The proposed framework first uses the POS filtering method to extract the historical events occurred in the same year with the input event. The extracted events are then displayed on the Google Map according to their locations. Such visualization information can provide an integrated view of temporal and spatial information of historical events. Future work will focus on improving the recall and precision. For the improvement of the recall, other useful information extraction methods will be investigated to extract the events with implicit description of locations. For the improvement of the precision, more significant features such as semantic information will be incorporated instead of using the linguistic category Nc alone.

6. References

[1] M. Stevenson and M. A. Greenwood. A Semantic Approach to IE Pattern Induction. In: *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. 2005, pp. 379-386.

[2] K. Yu, G. Guan, M. Zhou. Resume Information Extraction with Cascaded Hybrid Model. In: *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. 2005, pp. 499-506.

[3] C. H. Wu, L. C. Yu, F. L. Jang. Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records. IEEE *Intell. Syst.* 2005, **20**(6): 50-58.

[4] L. C. Yu, C. H. Wu, J. F. Yeh, F. L. Jang. HAL-based Evolutionary Inference for Pattern Induction from Psychiatry Web Resources. *IEEE Trans. Evol. Comput.* 2008, **12**(2): 160-170.

[5] C. R. Tseng, G. J. Hwang, W. F. Tsai. A Minimal Perfect Hashing Scheme to Mining Association Rules from Frequently Updated Data. *J. Chin. Inst. Eng.* 2006, **29**(3): 391-401.

[6] J. Nahar, K. S. Tickle, S. Ali, P. Chen. Significant Cancer Prevention Factor Extraction: An Association Rule Discovery Approach, *J. Med. Syst*. 2009, DOI: 10.1007/s10916-009-9372-8.

[7] C. H. Wu, Z. J. Chuang, Y. C. Lin.. Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models. *ACM Trans. Asian Language Information Processing*. 2006, **5**(2): 165-182.

[8] M. Lan, C. L. Tan, J. Su. Feature Generation and Representations for Protein–Protein Interaction Classification. *J. Biomed. Inform*. 2009, **42**(5): 866-872.

[9] J. Bj¨orne, F. Ginter, S. Pyysalo, J. Tsujii, T. Salakoski. Scaling up Biomedical Event Extraction to the Entire PubMed. In: *Proc. of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP-10)*. 2010, pp. 28-36.

[10] A. Agarwal and O. Rambow. Automatic Detection and Classification of Social Events. In: *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*. 2010, pp. 1024-1034.

[11] S. Liao and R. Grishman. Using Document Level Cross-Event Inference to Improve Event Extraction. In: *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*. 2010, pp. 789-797.

[12] Y. H. Chang, C. C. Luo, C. C. Huang. Efficient Evaluation of XML Twig Queries with Keyword Constraints. *J. Chin. Inst. Eng.* 2009, **32(**4): 469-480.

[13] R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, Reading, MA, 1999.

[14] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu M. Gatford. Okapi at TREC-3. In: *Proc. of the Third Text REtrieval Conference (TREC-3)*, NIST, 1995.

[15] B. He and I. Ounis. Combining Fields for Query Expansion and Adaptive Query Expansion. *Inf. Process. Manage.* 2007, **43**(5): 1294-1307.

[16] L. C. Yu, C. H. Wu, F. L. Jang. Psychiatric Consultation Record Retrieval Using Scenario-based Representation and Multilevel Mixture Model. *IEEE Trans. Inf. Technol. Biomed.* 2007, **11**(4): 415-427.

[17] J. Lu, D. Kang, Y. Zhang, Y. Li. Approximate Information Retrieval based on Multielement Bounds. *Knowledge-Based Syst.* 2008, **21**(2): 123-139.

[18] L. C. Yu, C. H. Wu, F. L. Jang. Psychiatric Document Retrieval Using a Discourse-Aware Model. *Artif. Intell.* 2009, **173**(7-8): 817-829.