Emergency Event Matching using Hierarchical Blocking Method

To cite this article: Chang Wen and Yu Liu 2019 J. Phys.: Conf. Ser. 1187 052095

View the article online for updates and enhancements.
Emergency Event Matching using Hierarchical Blocking Method

Chang Wen¹²³⁴*, Yu Liu¹²³⁴

¹College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
²Key Laboratory of Intelligent Information Processing and Real-time Industrial System in Hubei Province, Wuhan 430065, China
³Institute of Big Data Science and Engineering, Wuhan University of Science and Technology, Wuhan 430065, China
⁴Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, National Press and Publication Administration, Beijing 100038, China

597073407@qq.com

Abstract. With the extensive application of the knowledge base (KB), how to complete it is a hot topic on Semantic Web. However, many problems go with the big data, and the event matching is one of these problems, which is finding out the entities referring to the same things in the real world and also the key point in the extending process. To enrich the emergency knowledge base (E-SKB) we constructed before, we need to filter out the news from several web pages and find the same news to avoid data redundancy. In this paper, we proposed a hierarchy blocking method to reduce the times of comparisons and narrow down the scope by extracting the news properties as the blocking keys. The method transforms the event matching problem into a clustering problem. Experimental results show that the proposed method is superior to the existing text clustering algorithm with high precision and less comparison times.

1. Introduction

Events based on the text description contain a large number of potentially useful information, one of the most common type is the web news. Extracting the news messages in the specific fields such as finance, economy, can help us build knowledge base to develop some applications in the daily life. Thereinto, the emergency is very important because of its burstiness and unpredictability, it always produce disasters and lead to the economic losses. Therefore, we have built a semantic knowledge base of emergency based on the case database from Jinan University in previous work [1]. However, this database is small and limited, we have to extend our knowledge base by collecting the news from the web, and then transform the news into structured triples.

However, thousands of news appearing on webs may have the repetitive ones certainly. They are in different forms and written by different authors. Adding impure news into the knowledge base will lead to data redundancy. Hence, we need to filter out the news to grant the pureness of the database by Entity Matching. Entity Matching means recognizing the entities which refer to the same things in the real world. In the same way, the news entity matching is identifying the news reporting the same events. On
account of the flexibility we have discussed, the simple String Matching will cost large quantities of operation space. To reduce the times of comparing, we have to narrow down the comparative scope by extracting the data properties and ignoring the irrelevant matching-pairs. At the same time, extracting the critical properties can help lessen the workloads of transforming data into the knowledge base.

2. Related Work
Event Matching is to find out semantic projection or similarity between entities. Up to now, there are four major types of Entity Matching – based on the linguistics, the structure, the instances and the multi-strategy. The linguistics method uses the natural language processing technology to calculate the similarity of the name, label or description between entities. Event [2] is an activity in which a specific entity participates at a specific time and place. Information such as body, time, and activity can be presented as an element of the event [3]. The participants of the event are usually named entities. The integrated events contain a large number of entity schema information, which can be regarded as a natural dictionary as the basis for the identification of named entities. Compared with the identification of named entities in the literature [4] and link matching, information. The named entity form is free, and there are variants such as abbreviations, abbreviations, and nicknames. For the product names that appear in a large number of comments, it is necessary to add mode information [5] and symbol information. The entity in structure method is abstracted into a node and the relation becomes the edge. The application based on the instances and multi-strategy are less used because they have problems of sample size, algorithm performance and so on.

3. Challenges in Blocking News
Compared with the data in knowledge bases like DBpedia or WordNet, the form of the news is optional and informal. In addition, to grant the sufficient coverage of the data set, only collecting one website is not enough. We usually have to crawl the news data from two or more web pages. With the enlargement of the data set, there are two problems in the news collection, we call them time difference and classification difference.

- Time difference: due to the burstiness of the emergency, there is not enough time for some authors to deliver the news timely. It will result in the case that the same report will be published at different dates. Therefore, to regard the date of publication as the time of the event is not rigorous, we have to extract the correct date form the news to avoid the time difference though it is infrequent.

- Classification difference: the blocking method is based on the property of the event and one of the important properties is classification. Whereas each news web has its own classification framework, the Figure 1 shows that the same news reported in the two web pages have different classifications which have respective definitions. So, comparing the news events according to the classification directly is infeasible. We have to normalize the news classifications by using a unified method, the THUCTC (THU Chinese Text Classification) [7] for classification.

![Figure 1 Different classification of the same news](image)

4. Hierarchical Blocking
To solve the problems mentioned before and make the events matching, we proposed a blocking method to categorize the news events into several groups and each group contain the events which have the same property values. In this case, we can narrow the comparison range and reduce the number of comparisons.
Firstly, we crawled news from two Chinese news sites, 2,000 news from 163NEWS and 10,000 news from CHINANEWS. The time is from January 1, 2016 to January 6, 2016, the news category covers most of the main categories apart from the life class such as the introduction to life knowledge and makeup introductions.

Figure 2 shows the main steps of news entity matching, we can get four existing properties of the raw text form the news web -- published date, title, text, and URL. Then, we constructed the blocking layers by the blocking keys, which can narrow the scope of comparison layer by layer. Afterwards we extracted the information of blocking keys in each news through the existing properties. Finally, we compared the news in each block to filter out the same news and picked out the emergency news in addition.

4.1 Constructing Blocking Layers

The news from the website is semi-structured because it has some existing properties such as title, published date, text, and its URL, which called EP (existing properties). However, it is impractical to use EP to match the news directly because of the data disunity. Since an event is defined as follow -- something that happens at a given place and time [8], we extracted date, place and the type as the blocking keys according to the EP. These keys finally constitute the hierarchy blocking framework and the news obtaining the same property values will be distributed into the same blocks. Only comparing the news in the block can reduce the number of comparisons effectively, which is significant to the large amounts of data.

Assume we are given a set of news $N_e = \{n_1, n_2, n_3, n_4, \ldots, n_n\}$ where $n_i$ is the news from two news webs. We propose the block formulation $B_k = f(d, p, t) = \{n_{k1}, n_{k2}, \ldots, n_{kn}\}$, which means that a block is determined by the date, place and type. $B_k$ is the final block and $n_{kn}$ is the news event within the block, all the news in one block possess the same property values of date, place and type. The function $f$ is the operation process of the blocking, the news will be handled by the function $d$, $p$ and $t$ sequentially. The input of each function is the output of the previous one, except that the input of the first function is $N_e$. Through the function $d: N_e \rightarrow B_{date}$, the news can be classified into several blocks and these blocks will be the input of the function $p: B_{date} \rightarrow B_{date&place}$. Then the $B_{date&place}$ will generate the final blocks according to the function $t: B_{date&place} \rightarrow B_k$. The comparison works only focus on the $B_k$ sets which can substantially reduce the number of matches.

Figure 3 shows the main process in news blocking, we initially have ten news with redundancy which are in different forms. Firstly, the candidate date of the news will be extracted, so that news will be divided into several groups according to it. Then, the place of each news is used to subdivide the news into fine-grade blocks. Eventually, we unify the type of the news form two different webs and get
the final block sets. Following the definition of the event, the repetitive news will be in the same block therefore we have no need to compare the news across the blocks.

![Figure 3. Generating blocks in each Blocking Layer](image)

**4.2 Extracting Blocking Keys**

In order to extract blocking keys and address the challenges mentioned above, we make full use of the EP (existing properties). The sequence of extraction can be changed optionally, we follow the steps of date, place and type extraction. We now describe the details of each step which can generate the $B_k$ finally.

**4.2.1 Date Extraction**

Time difference is one of the challenges, we can get the published date from the website easily but cannot confirm the date of the news event. Since there will be discrepancies in the time when the authors of different sites publish news.

To solve this problem, we define the candidate date as $CD = \langle PD, ED \rangle$, $CD$ is a date set composed of two types of dates. $PD$ (published date) is the date when the news is published and the $ED = \langle ED_1, ED_2, \ldots, ED_n \rangle$ is another date set we extract form the articles. Based on the characteristics of news narrative, the date of the event is usually mentioned in the first few sentences of the article. Therefore, we use the regular expressions to extract date to join into $ED$.

\[
CD = \begin{cases} 
PD, & \exists date \in ED \rightarrow date = PD \\
PD \cup ED, & \notin date \in ED \rightarrow date = PD 
\end{cases}
\]  

If the $ED$ contains the date which is equal to $PD$, then we can determine the date of the news is $PD$. Otherwise, all the dates in $CD$ will be regarded as the blocking keys to divide the news. Although this situation is rare due to the real-time nature of the news, we have to check it out to prevent data omission.

**4.2.2 Place Extraction**

Another important factor in news events is the place, which is usually in the form of country, province or city. These place words have fixed expressions, using ordinary word segmentation may result in too abstract or refined places. For example, we want to get “Hubei Province” in “Hubei Province Talent Market held a job fair for college graduates”, but the “Talent Market” will also be extracted which is abstract to block the news. In another case, some local news contains the refined places, such as “Qingcheng Mountain” and “Forbidden City”, which generate blocks containing few news events. These two types of places lack of application value, therefore the word segmentation is inapplicable to extract places from the news.
In order to unify the places, we constructed a word dictionary composed of provinces and cities in China and the world's major countries. For those over 200 place words, it is unrealistic to match them out one by one on each news because of the time consuming. To get the keywords of the news effectively, Figure 5 shows our method which is to generate bit strings corresponding to the length of the keywords and store the strings as hash set. Then the text will be traversed quickly to get characters, each one of them will be matched with the word dictionary and the corresponding position in the bit strings, the character is changed to 1 if it is matched. If some bit strings are all 1, the corresponding keywords are returned. For example, the keyword is “CHINA”, the matching result of “CHI” is 2 and the binary is “10”, which means “CHI” is in the first position of the keyword. Then the following characters are compared until the end of the text, if the bit string of “CHINA” is “11” it means the text includes “CHINA”.

4.2.3 Type Extraction
The news type is the last blocking layer to narrow the block sets. Given the elements of the news, the object and action are very crucial factors to define news apart from the time and place. However, on the one hand, these two factors will let the final blocks too fine grained and make the process too difficult to realize. On the other, extending the E-SKB needs to recognize which news belongs to the emergency, so the type extraction works.

In order to get the emergency news, the final categories should be extracted particularly. We chose 8 categories that may contain the emergency, together with the news corpus in E-SKB to form a new training set. Then the training set was used to reclassify the news datasets to get the emergency and the corresponding types.

4.3 Matching news
Through the extraction method mentioned above, the news datasets are divided into several blocks based on the date, place and type subsequently. Then the news in each fine-grained block is compared with others by the news headline. Since the headline covers the major information of the news, we simply use the cosine similarity to match the same news out.

5. Experiments

5.1 Matching Numbers
The Hierarchical Blocking method we put forward can reduce the matching numbers greatly, since this method will divide the datasets into several blocks by the properties of the news, only the news in the same blocks will be compared instead of the Cartesian product on the whole datasets. To evaluate the
matching numbers, we use three cluster baselines under different numbers of news, K-Means, DBSCAN and GSDMM.

Figure 5 shows the matching numbers of the four methods under different numbers of news in two websites. Our method in the case of different amounts of data, the matching numbers are much lower than the other three methods. The matching numbers of the K-Means and the DBSCAN present a huge increase with the augment of the data volume. Since they use TF-IDF in documents distance calculation, so they compare with each other between two news datasets and the result is the Cartesian product between two datasets. As for the other baseline, the matching numbers of the GSDMM relates to the number of words in each document. Since we use the title of the news to cluster, the average number of words in titles is set to 6.

From the result, we can find that the comparison times on our method is less than the baselines. In addition, the time consuming of the Hierarchical Blocking is shorter than the basic clustering methods, it takes only 4 seconds to process about 450 news articles, but the basic methods will take several minutes or even longer. Because the parameters are needed to be very specific (such as the cluster numbers k in K-Means is more than a few hundred) so that each cluster contains the same news or not, instead of containing the similar news or series of events which cannot help us to figure out the same news events exactly.

5.2 Matching Results on the Same News
In this part, we try to compare the performance of Hierarchical Blocking with K-Means, DBSCAN and GSDMM on matching the same news. Since extracting the same news events belongs to the fine-grained clustering problem, we chose other three effective clustering methods to evaluate ours on 450 news articles (300 in CHINANEWS and 150 in 163NEWS).

Based on the manual statistics on the news articles, we set the k in K-Means to the true number of the clusters which is 250. For DBSCAN, we set the density radius to 1.8 and the minimum number to 230. For GSDMM, we set the number of iterations at 100, α = 0.1, and β = 0.1 for the dataset. The Hierarchical Blocking does not need to set any parameters. Then, we use F1-score to evaluate the performance, the recall represents the number of the same news pairs found divided by the total number of the same news pairs, the precision means the number of the same news pairs found divided by the clusters which contain more than 1 articles.

Table 1. Performance of four methods on matching the same news

| Methods  | Recall (%) | Precision (%) | F1 (%) |
|----------|------------|---------------|--------|
| K-Means  | 64.15      | 43.17         | 51.61  |
| DBSCAN   | 62.26      | 42.32         | 51.58  |
| GSDMM    | 88.68      | 68.05         | 77.01  |
Table 1 shows the performance of Hierarchical Blocking with K-Means, DBSCAN and GSDMM on matching the same news. We can see Hierarchical Blocking performs better than K-Means and DBSCAN in recall and it reaches the precision over 90%. The GSDMM can reach a high level of recall but the precision is common in this task. Although GSDMM can deal with sparse and high-dimensional problem of short texts, like the other two baselines, the clusters will contain some irrelevant news or similar news (like series of events) to reduce the precision. On the one hand, the parameters required in the baselines will affect the results deeply. On the other hand, the high-degree sparseness of the data makes it hard to get the desired results with coarse-grained clustering. We used two news websites to prevent omission, so the ideal result is that each block contains 2 news (refer to the same event) or only 1 news (has no repetitions). Hierarchical Blocking can extract the news pairs only referred to the same events and will not consider the irrelevance, but other methods can not accurately divide the same news into the unique clusters, which will influence the precision and the F1-score finally.

6. Conclusion

This paper uses a Hierarchical Blocking method to complete emergency news matching, which is used to extend the original emergency knowledge base by converting the event matching problem into a clustering problem. At the same time, the event attributes are extracted in the process of Blocking. The experimental results show that the proposed method can reduce the number of matches compared with other clustering methods, and has higher accuracy and F1 value. However, the matched emergency needs to be further integrated into the knowledge base, which is the future work.

References

[1] Chang W, Yu L, Gu J, et al. E-SKB: A Semantic Knowledge Base for Emergency[C]// China Conference on Knowledge Graph & Semantic Computing, 2016.

[2] Naughton M, Stokes N, Carthy J. Sentence-level event classification in unstructured texts[J]. Information Retrieval, 2010, 13(2):132-156.

[3] Zhang Chuanyan, Hong Xiaoguang, Peng Zhaohui, et al. Web Entity Activity Extraction Based on SVM and Extended Conditional Random Fields[J]. Journal of Software, 2012, 23(10): 2612-2627

[4] Cecile B, David C J, Matteo M, et al. Clustering attributed graphs: Models, measures and methods[J]. Network Science, 2015, 3(3):408-444.

[5] Chandel A, Nagesh P C , Sarawagi S . Efficient Batch Top-k Search for Dictionary-based Entity Recognition[C]// International Conference on Data Engineering. IEEE, 2006.

[6] Jingyang Li, Maosong Sun. Scalable Term Selection for Text Categorization. Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 2007, pp. 774-782.

[7] Qiu T, Qiao R, Wu D O. EABS: An Event-Aware Backpressure Scheduling Scheme for Emergency Internet of Things[J]. IEEE Transactions on Mobile Computing, 2018, PP(99):1-1.