ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 116-122
doi: 10.18178/wcse.2017.06.020

An Algorithm of News Topic Time Extraction Based on the Optimal Parameter Hierarchical Tree Model

Yang Weixin[1], Tang Lingli[1], Wang Shengxiang[1], Wang Chaoliang[1], [+]Li Chuanrong[1]

Key Laboratory of Quantitative Remote Sensing Information Technology, Chinese Academy of Sciences, Beijing 100094, China

Abstract. The time of news topic event is the representation of the temporal characteristics of news, which plays an important role in news information retrieval and mining. According to the current news topic time especially low accuracy of implicit time extraction problem, starting from the characteristics of news analysis, news category feature and web structure mining news topic - time relation model, considering the frequency and position of the theme time, put forward the theme of time extraction algorithm of optimal parameters of the model hierarchical tree model. Randomly selected samples from the Internet for test, the results show that the proposed algorithm has higher accuracy than other methods.

Keywords: Topic time, Information extraction, Web pages, Hierarchical tree, Explicit time, Implicit time.

1. Introduction

The rapid development of Internet technology provides technical support for the rapid dissemination of information, which, as a very important media, is rich in a lot of useful information. Because of the rapid update of the network news and the content of the web pages, how to extract the information from the huge network news quickly and accurately is an important application in the research of natural language information processing.

In the study of Web news statistics, more than 80% of the web pages contain spatio-temporal data[1]. Topic news time is a very important information, which reflects the time features of news events, if topic news events lack of time, It will lead the user to understand the whole process of news development. At the same time, the accuracy of time extraction is not only directly related to the efficiency of news extraction and retrieval, but also has an important impact on the information monitoring of web news[2].

There is a close relationship between the news topic and topic time. In order to obtain the news content accurately, it is very important to get the news topic time accurately. Search news topic time usually contains two algorithms， one is the news release time[3,4], such as Yahoo, Google and so on. Other algorithm is the statistical time[5-8], it is to appear as the highest frequency of web news time. However, the two methods can’t find the news topic time accurately, there is usually a deviation, so it is can't satisfy the demand of the news topic time retrieval. Therefore, it is necessary to find a new algorithm which can extract topic time from Web news accurately. Therefore, in this paper, accord to the characteristics of web news[9], a new web news topic time algorithm is proposed, which takes full account of the frequency and structural features of the temporal information in the news text, eliminates the influence of the noise time by the non topic time, and improves the accuracy of the topic time extraction.

[+] Corresponding author. Tel.: +86-10-82178616; fax: +86-10-82178600.
E-mail address :crli@aoe.ac.cn.

2. News Feature Analysis

News writing is special, in order to attract the readers, the result of the news would show in the title or the first paragraph. Thus, by analysis of the structure of news, the key of important information often appear in the news title and the first paragraph [9]. Strengthen the use of title, first sentence and paragraph can effectively extract topic news information.

The title is the essence of news, which has the function of revealing, clarifying and evaluating news content. In the light of the genre of online news, Yan [10] has studied the characteristics of title in form, content and language. From the formal point of view, due to the layout effect, the network news headlines for sentence length limited. The statement of the news headlines is dominant, and a small number of titles are composed of verb phrases and noun phrases. In terms of content, most news headlines can be accurately summed up what happened, but view from the linguistic; it is very simple and maybe has some grammatical ambiguity.

3. Time Resolution Algorithm

Time is an important part of news information [11], the analysis of time information in the news is a branch in the field of Natural Language Processing. In the news, the time information includes explicit time and implicit time, and the explicit time can be founded in the time axis directly. But as the implicit time, it is related to the context semantics, the reference time is different, so how to accurately analyze the implicit time is very difficult. In order to accurately analyze implicit time, for dynamic context, the dynamic reference time is a kind of high precision method [12], as to static context, by setting certain rules, build a static reference time can obtain good results [13]. There are two main analytical methods for time resolution, one is a time rule matching method [14], which is constructed by the time lexicon and time information description pattern library to parse time [13]; the other is a machine learning method, it is the use of semantic roles to parse time[15].

How to determine the time of news topic, some people use news release time as topic time[16], the method is simple and direct, but the accuracy is very low. The most of the topic time contain in news context. In consideration of news title, the first section contains a large amount of information, the topic time probability is greater than others, therefore, some people applied this feature extract topic time[16], however，different news topic time expression in thousands of ways, extract topic time accuracy is not high. In order to avoid the impact of news headlines and the first paragraph, a new approach time frequency features topic time is proposed[17], but this method only consider frequency feature, ignore the relevance of news and topic time, while non-topic time and topic time frequency very close, this method is invalid. To be more accurate get topic time, based on the relationship between topic time and news topic, the topic time mapping model is constructed to determine the topic time, and then using machine learning method get topic time[18], the effect is dependent on the effect of the extracted news topic and the mapping model.

4. Topic Time Extraction

4.1. Time Resolution

Regular expression matching method [19] has many advantage, such as simple, extensible, good recognition (F measure can reach 90.15%[13]) and so on. So we can use this method to analyzing news time. In the process of news time resolution, the reference time largely determines the accuracy of time resolution, it is divided into global and local reference time. Explicit time usually based on global reference time and can be resolved accurately. With the change of context, implicit time's reference would be changed, so it must choose local reference time, therefore it is very difficult to resolve. News time analysis algorithm as follows:

Algorithm-1

Input: a list of time expressions arranged in chronological order

Output: a standardized list of time expressions

Initial global reference time and local reference time

For TE_i time expression TE_i

If TE_i is an explicit time expression

Analysis of TE_i based on global reference time

Modify the local reference time to TE_i

Elseif TE_i is global time

Analysis of TE_i based on global reference time

Modify the local reference time to TE_i

Else

Analysis of TE_i based on local reference time

End if

End if

4.2. Time Extraction

Taking full account of news feature, time frequency and different times correlations, constructed of hierarchical tree model of time information in figure 1. The hierarchical tree model contains the root node and some sub-trees, time information stored in the node, and time node through the branches connected, so as to establish a layered structure relationship.

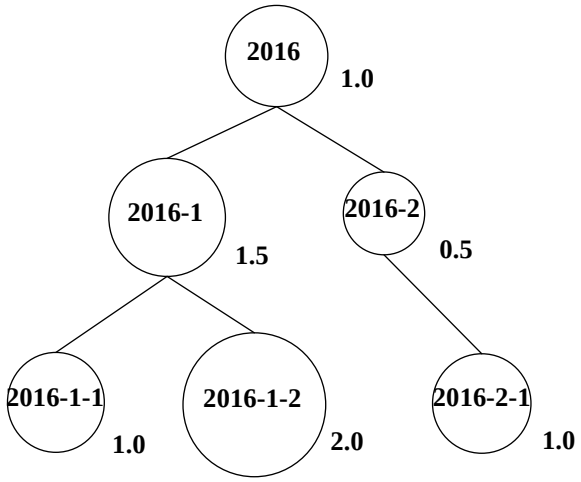

Fig. 1: Time hierarchical tree model

Time frequency weight

The hierarchical tree node contains weight, it is representative important degree. The weight of node should takes into account the position and frequency of time. The calculation rules are as follows:

Score(T_i) representative T_i weight, consisting of explicit weight ES(T_i) and implicit weight IS(T_i).

$$\text{Score}(T_i) = ES(T_i) + IS(T_i) \quad (1)$$

Explicit weight ES(T_i) include both explicit time frequency $TF_{ETE}(T_i)$ and implicit time frequency $TF_{ITE}(T_i)$. In some case, explicit time and implicit time are the same time, and not certain. So in order to improve weight accuracy, we set weight factor d, explicit time weight as formula (2).

$$\text{ES}(T_i) = TF_{ETE}(T_i) + d \cdot TF_{ITE}(T_i) \quad (2)$$

Implicit time weight IS(T_i) is related to the sub time node in the hierarchical tree, the method as follows:

$$\text{IS}(T_i) = \alpha \cdot \sum_{i=1}^{n} Score(C_i) \quad (3)$$

$C_1, C_2, \ldots, C_n$ on behalf of sub time node T_i, α time contribution factor.

Structure weight

News genre determines the title, the first paragraph contains important information[13]. Therefore, it is necessary to increase the weight when calculating the weight information of the structural position.

Weight adjustment principle is as follows, if T_i in title, first sentence and second sentences, we will increase weight T_i.

$$\text{Score}(T_i) = Score(T_i) + \mu \cdot SUM \quad (4)$$

SUM is all nodes weight in a hierarchical tree, μ is adjust factor. Algorithm for computing node weights of hierarchical tree as follows.

Algorithm-2

Input: after analysis time list

Output: hierarchical tree time node weight list

For each T_i time

If T_i is an explicit time

$$\text{Score} = 1$$

Else

$$\text{Score} = d$$

End if

$$T_i.\text{Score} += \text{Score}$$

While T_i is not a year.

T_i= T_i father's time

$$\text{Score} = \alpha \cdot \text{Score}$$

$$T_i.\text{Score} += \text{Score}$$

End while

End for

Calculate SUM, adjust score according to position

4.3. Topic Time Selection

Through the above algorithm, the weights of each time node are calculated, and according node weight select topic time. Principle of selection as follows, first, set up node threshold, from the root node traversed to the child node in the hierarchical tree. If the value of the sub node obtained is greater than the threshold value, the time of the leaf node is topic time, else the topic time is unknown.

Algorithm-3

Input: time information hierarchy tree

Output: theme time

Initialize the theme time TR for layered root node time

while time TR is not leaf node

if TR $< threshold$

return unknown

else

$\text{TR} = \max(TR_i)$

end if

end while

4.4. Model Parameter Optimization

After set parameter d, α and μ, we can calculate the weight of each time node, determine the theme time. These parameters will bring the time node weight change, thereby affecting the accuracy of topic time. Therefore, we need to optimize the parameters, making the topic time accuracy maximum. The algorithm as follows.

Algorithm-4

Input：news sample

Output：d, α, μ, which make the topic time extraction accurately maximum

```
for d = 0 : 0.1 : 1
    for α = 0 : 0.1 : 1
        for μ = 0 : 0.1: 1
            for all sample news
                Calculate the time node weight;
                get topic time；
                if  topic time == get topic time
                    right_times = right_times + 1;
                end
            end
          output right_times;
        end
    end
end
```

through max(right_times), we can get the best parameters d, α and μ.

5. Test And Analysis

In order to test the hierarchical tree model method in extracted news topic time, we crawl 500 web news from internet. Obtain standard values by manually, and through the accuracy rate to evaluate algorithm performance. Compare with statistical method, when set d = 0.8, α = 0.5, μ = 0.2, the results as table 1.

Table 1. Test Results

	Hierarchical tree model	statistical method
Test samples	500	500
Right numbers	360	235
Accuracy rate	72%	47%

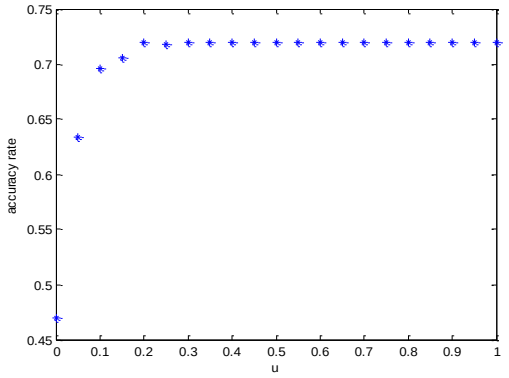

Fig. 2: different μ with the accuracy rate

From Table 1, we can find that the hierarchical tree model method more advantage statistical method. In order to get parameter varying for the hierarchical tree model method, we set μ varying from 0 to 1, the result as Fig. 2. When μ = 0.2, the result get best.

when set μ = 0.2, by change d and α, we can get results as table 2. From table 2 know that d = 0.7 and α = 0.5, the hierarchical tree model method get the best performance.

Table 2. Different d, α, Right Numbers

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	343	347	354	358	364	371	370	370	375	372
0.2	345	351	357	360	364	371	370	371	377	372
0.3	347	352	357	360	363	371	371	374	376	372
0.4	348	352	358	362	371	371	372	374	374	372
0.5	350	357	362	365	373	372	**378**	372	374	373
0.6	352	358	364	364	371	374	371	370	373	372
0.7	352	358	364	365	370	371	370	370	373	372
0.8	352	358	363	366	369	370	369	370	373	372
0.9	352	356	359	367	369	370	369	369	373	372
1	352	360	363	369	373	373	372	372	372	375

6. Conclusion

As to the problem of low accuracy rate in topic time extraction from web news, this paper presents a hierarchical tree model method which can extracted web news topic time. And then, by adjust some parameters, it make accuracy rate maximum. In order to test the method performance, compare with statistical method, out approach has more accuracy rate. By analysis topic time, there is more space can be improved, after check up news sample, we find that most of un-extracted topic time are non- events news, the topic time in the unknown. When using hierarchical tree model to calculate the value of node time, the impact of the score, the applicability of non news event text is poor, so we will focus on the classification of events and non-news events in next step, to improve the accuracy of extract news topic time.

7. Acknowledgment

This work was supported in part by the National High Technology Research and Development Program (863 Program) of China (2013AA122904，2013AA122103).

8. References

[1] Palkowsky B, Metacarta I. A new approach to information discovery-geography really does matter[C]//ATCE 2005:*SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers*,2005:3231-3234.

[2] Klaus B, Srikanta B, Thomas N. FluxCapacitor: efficient time-travel text search[C]. *In:Proceedings of VLDB* 07,2007:1414-1417.

[3] Melvin M. News reporting and writing (9th ed) [M]. *Beijing: Tsinghua University press*, 2004.

[4] Omar A, Michael G. Clustering of search results using temporal attributes[C]. *In: Proceedings of SIGIR* 06,2006:597-598.

[5] Li B.L, Li W.J, Lu Q. Topic tracking with time granularity reasoning[J]. *ACM Trans. On Asian Language Information Processing*, 2006,5(4):388-412.

[6] Wang W, Zhao D,Zhao W. Identification of topic sentence about key event in Chinese news[J].*Acta Scientiarum Naturalium Universitatis Pekinensis*,2011,47(5): 789-796.

[7] Lee L. Book Reviews: Foundations of Statistical Natural Language Processing[J]. *Microbiology*, 2015, 144 (pt 4)(3).

[8] Mochihashi D. Machine Learning in Statistical Natural Language Processing(Machine Learning for Media Processing; Key technologies for big-data applications)[J]. *Journal of the Institute of Image Information & Television Engineers*, 2015, 69:131-135.

[9] Klaus B, Srikanta B, THOMAS N,et al.A time machine for text search[C]//SIGIR 2007:*The 30th International ACM SIGIR Conference on Research and Development in Information Retrieval.ACM*,2007:519-526.

[10] Yan H, Yang J. A very efficient approach to news title and content extraction on the web.[J]. 2011:389-390.

[11] Tran M V, Nguyen M H, Nguyen S Q, et al. VnLoc: A Real -- Time News Event Extraction Framework for Vietnamese[C]// *International Conference on Knowledge & Systems Engineering*. 2012:161-166.
[12] Zhao X. Jin P.Yue L. Automatic temporal expression normalization with reference time dynamic-choosing[C]//COLING 2010:*The 23rd International Conference on Computational Linguistics. Association for Computational Linguistics*, 2010:1498-1506.
[13] Lin J, Cao D, Yuan C. Automatic TIMEX2 tagging of Chinese temporal information[J].J *Tsinghua Univ(Sci&Tech)*,2008,48(1): 117-120
[14] Zhang C, Zhang X, Li M,et al. Interpertation of temporal information in Chinese text[J].*Geography and Geo-Information Science*, 2014, 30(6):1-6.
[15] Liu L, He Z, Xing X, et al. Chinese time expression recognition based on semantic role[J].*Application Research of Computers*,2011,28(7): 2543-2545.
[16] Jin L, Zhao,et al.TISE: A temporal search engine for web contents[C]//IITA 2008:*Intelligent Information Technology Application. IEEE*, 2008: 220-224.
[17] Sheng L, Pei Q J, Xu J.Z, et al. Extracting focused time for web pages[C]//WAIM 2012:*International Conference on Web-Age Information Management. Springer Berlin Heidelberg*,2012:266-271.
[18] Krapivin M, Autayeu A, Marchese M, et al. Keyphrases extraction from scientific documents: improving machine learning approaches with natural language processing[C]// *The Role of Digital Libraries in A Time of Global Change, International Conference on Asia-Pacific Digital Libraries*, Icadl 2010, Gold Coast, Australia, June 21-25, 2010. Proceedings. DBLP, 2010:102-111.
[19] Shahbaz M, Mcminn P, Stevenson M. Automated Discovery of Valid Test Strings from the Web Using Dynamic Regular Expressions Collation and Natural Language Processing[J]. 2012, 430(4):79-88.