Advances in Engineering Research, volume 162
3rd International Conference on Advances in Materials, Mechatronics and Civil Engineering (ICAMMCE 2018)

Research on Geographic Information Extraction Based on Knowledge Graph

Ren Zhenyang[1,a],Yu Hongzhi[2,b] ,Wan Fucheng[3,c,*]

[1] Key Laboratory of National language intelligent processing, China, Gansu Province, Lanzhou,730030

[2] Key Lab of China's National Linguistic Information Technology，China,Lanzhou,730030

[3] Key Laboratory of National language intelligent processing, China, Gansu Province, Lanzhou,730030

[a]1031439279@qq.com, [b][2]wanfucheng@126.com, [c]306261663@qq.com

* Wan Fucheng

Keywords: Knowledge Graph, Web information extraction, Placename extraction, Web page annotation.

Abstract. In the rapid development of the Internet, in order to achieve the effective extraction of geographical names information, the study of location information extraction of web on the basis of the Chinese knowledge map CN-DBpedia is carried out. The Chinese geographical names are obtained from Wikipedia based on statistics. And the attribute value dictionary is built on the basis of CN-DBpedia, which can be used to automatically annotate the data items of the web pages. Compared with manual annotation, this method has higher accuracy and recall rate, which can significantly save human resources and greatly improve the efficiency of information acquisition.

1. Preface

At present, about 20% of data queries on the Internet through search engine websites are related to geographically related information, and 80% of these geographically related queries are specific place name searches. At the same time, most human behavior activities are closely related to a certain geographic space. This leads to all kinds of data generated by the activities have a geographical context, and most of them are expressed in the form of place names. According to relevant investigations, at least 20% of web pages are associated with geographical indications, such as zip codes, Internet Protocol addresses (IP addresses), and so on. With the advent of the mobile Internet era and the proliferation of location-aware devices, many web content directly provide coordinates or geotagging. Therefore, the importance of geographical information is particularly important.

On the other hand, with the rapid development of the Internet, the number of websites is growing exponentially. According to the statistics of Internet Live Statas in 2014, the number of global Internet sites has exceeded the one billion mark. In 2016, only 1.25 million new online websites were launched in China throughout the year, and the average number of websites opened every month was as high as 105,000. The era of the Internet has already arrived. However, the huge amount of webpages not only contain many false and meaningless content, but also increase people's difficulty in obtaining geographic information. How to effectively obtain relevant geographic information becomes very important. Based on the existing Knowledge Graph and information extraction research status, this paper proposes applying Knowledge Graph to web page information extraction to promote the application of Knowledge Graph. This article is based on this background to start research.

2. Based on Wikipedia's place name extraction

2.1 Pre-processed corpora

Wikipedia is an online global encyclopedia website. The entire website covers a very large area of knowledge. Wikipedia has been welcomed by many countries for its free and comprehensive features. As of 2015, the number of English entries in Wikipedia has exceeded 4.9 million, and the total number of registered users is more than 30 million. From 2002 to the present, Wikipedia covered a large number of Chinese-named entities, including Place names.

After downloading the Chinese offline data package from the Wikipedia website, after removing a series of work such as removing unnecessary XML tags, unifying the fonts, and performing word segmentation, the content is preprocessed.

2.2 Place Named Entity Recognition

The method of geographical name identification mainly includes two categories based on rules and statistical methods. Both have their own advantages and disadvantages, and now it is common to use statistical methods. Because the rule-based method is extensible and easy to understand, it is not easy to transplant, and the description of the rules is relatively complicated. Even a professional linguist is not easy to complete.

Statistically based methods are low in terms of language and have good portability, but require manual annotation of corpora and selection of appropriate statistical learning models and parameters. Here, using a multi-conditional random field model to process pre-processed text, a Chinese place-name set is obtained.

Conditional Random Field Models (CRFs) are theories proposed by J. Lafferty et al. in 2001. With the rise of artificial intelligence in recent years, this theory has been widely used in the field of natural language processing. The principle: For a given observation sequence $X=\{x^1, x^2, \ldots, x^n\}$. The state sequence corresponding to the conditional random place is $Y=\{y^1, y^2, \ldots, y^n\}$, and its conditional probability is defined as:

$$P(Y|X)=\frac{1}{Z(X)}exp\left(\sum_{i,j} l_j\, t_j\left(y_{i-1}, y_i, x, i\right)+\sum_{i,k} m_k\, s_k\left(y_i, x, i\right)\right). \quad (1)$$

In the formula, Z(X) is a normalization factor, the probability sum of all state sequences is 1; $t_j(y_{i-1}, y_i, x, i)$ is the state transition function of the observation sequence $i-1$ and i; $s_k(y_i, x, i)$ is the state characteristic function at the mark of the observation sequence i.;l_j and m_k are the weight of the corresponding eigenfunction, obtained through training estimation. After establishing the probability model of $P(Y|X)$, the solution of state sequence marker Y can be transformed into solving Y^* when $P(Y|X)$ is maximum.

$$Y^*=argmax_Y P\,(Y|X). \quad (2)$$

The probability map model used by the conditional random field can effectively express long-distance and interdependent features, and all features can be globally normalized to obtain a global optimal solution.

The Chinese geographical name recognition based on conditional random fields mainly consists of five steps: (1) feature template generation, (2) feature selection, (3) parameter training, (4) data processing, and (5) place name identification. The feature template includes three types of features, which are syntactic, geographical name elements, and part of speech. The overall structure is shown in Figure 1.

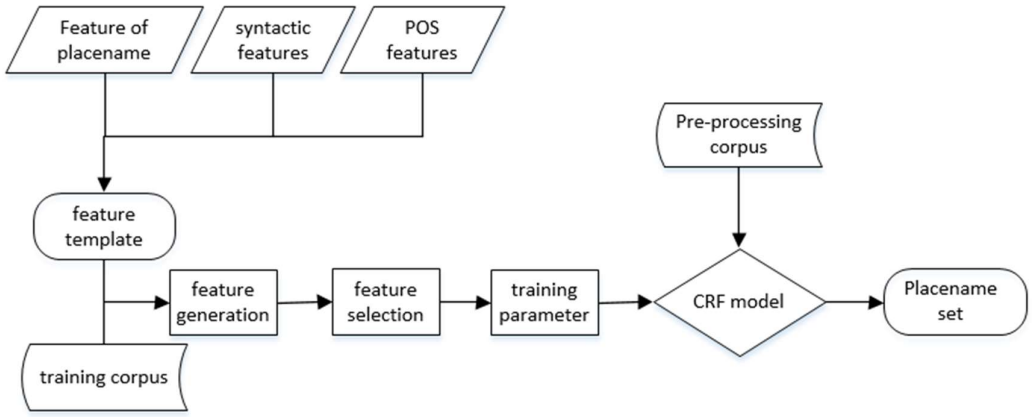

Fig. 1. Chinese geographical name recognition structure based on conditional random field model

3. Generation of attribute value dictionary based on Knowledge Graph

Knowledge Graph are also called scientific Knowledge Graph. They are modern frontier theories combined with applied mathematics, information science, and graphics, and a knowledge of triplets constitutes the main body of the Knowledge Graph. Using Knowledge Graph can significantly improve search efficiency, not only can quickly find the most desired information, but also make the search more depth and breadth. At present, there are a large number of knowledge bases on the Internet, such as Freebases, DBpedia, and so on. Among them, the largest domestic and the earliest published Chinese Knowledge Graph is CN-DBpedia proposed by the GDM laboratory of Fudan University. The map is free to use, so it is used by nearly 100 enterprises and organizations now.

Here, we obtain the triplet through CN-DBpedia and generate an attribute value dictionary. The specific process is shown in the figure 2 below.

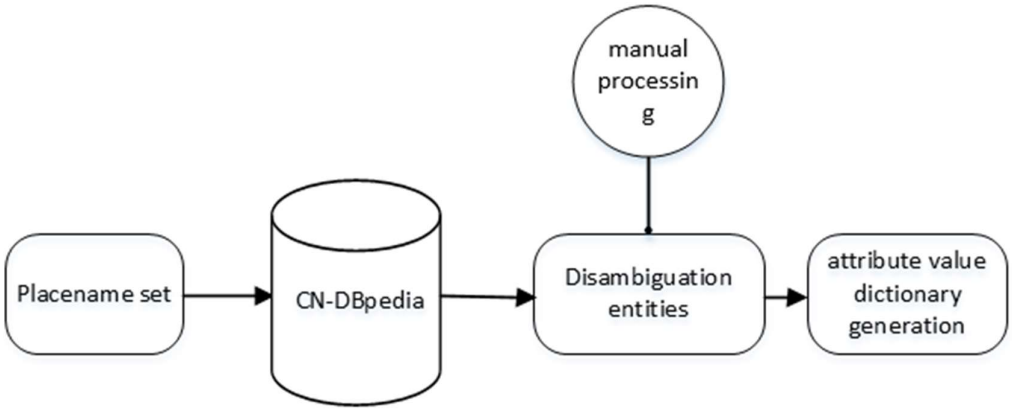

Fig. 2. attribute value dictionary generation flow chart

3.1 Chinese Knowledge Graph

CN-DBpedia, like the well-known knowledge base in the world, uses an entity - centric format of triples, namely, < entity, attribute, attribute value >. A triple can be a description of an entity category or ontology, or a description of an attribute of an entity. For example, the description of the location of Dunhuang is shown in the table below.

Table1. The description of the location of Dunhuang

entity	Attribute name	Attribute value
Dunhuang	Chinese name	Dunhuang
Dunhuang	Alias	Shazhou
Dunhuang	Location	Northwestern Gansu
Dunhuang	Code	620982
Dunhuang	famous scenery	Mogao Grottoes
Dunhuang	Postal code	736200
Dunhuang	area	31200 square kilometers

The set of placenames obtained in the above section is input data, and the relevant entities can be obtained through CN-DBpedia. After disambiguation entities are eliminated, all the triplets of all place entities are counted, and attribute and attribute value data are obtained.

3.2 Disambiguation entities

Considering the ambiguity of place names, for example, searching CN-DBpedia for the word "Dunhuang", multiple entities with the same name will appear, such as the movie "Dunhuang", the song "Dunhuang" and the book "Dunhuang", etc. Obviously these entities The corresponding attributes and attribute values do not meet the characteristics of the place, so the work of disambiguating entities must be done.

Since the location entity has fixed attributes such as alias, geographic location, postal code, famous attraction, and area code, the common attributes of the site can be summarized in advance and these attributes can be made into a property template. Then use this template to compare all attributes corresponding to entities found in CN-DBpedia, perform similarity calculations, set a threshold, and filter out entities below this value. This effectively eliminates ambiguous entities. It is worth noting that this method may not be able to filter out similarly named entities with similar attributes. This requires further manual processing. Due to the previous filtering process, manual processing workload is significantly reduced, which greatly improves the accuracy of the entity selection.

3.3 attribute value dictionary generation

Based on the processing of the previous section, you can get the exact location entity. For all the attributes of all the location entities, build the attribute value dictionary Dict() in the format of <entity, attribute value of attribute p>. To operate, simply take the triplet containing the attribute p under location A from the Knowledge Graph. Because multiple triples are used to store different attribute values of an entity attribute in the Knowledge Graph, multiple values need to be combined when generating the attribute value dictionary.

4. Web pages are automatically labeled

4.1 Web Preprocessing

There are many webpages on the Internet today that do not conform to the W3C standard. Therefore, before the webpage is annotated, all the webpages to be extracted are subjected to HTML code cleanup, the labels that have not been correctly ended are filled in the webpage, and the wrong labels are processed. Fix and other operations. We can parse HTML-encoded webpages into DOM trees, and construct Xpath-to-text mappings for all DOM trees containing textual leaf nodes. This facilitates webpage labeling and final execution of the wrapper to complete the information extraction. This conversion converts an HTML page into a set of XPath-to-text node mappings. The subsequent page annotations only need to be compared with the text nodes in <xpath,text_node>, and the corresponding XPath data items can be marked as specific. Attributes.

4.2 Web Entity Mapping

The article uses the document detection technology in the traditional search engine to complete the mapping of web pages and entities. This kind of mapping relationship is the degree of correlation between the computing entity and the web page. We treat it as a document detection problem. That is, given an entity, the related web page w is returned from the web page collection W.

For those webpages to be annotated, HTML tags are removed, word segmentation is performed, web pages are converted into word vectors, an inverted index of these web pages is established, and the web pages are sorted according to the TF-IDF policy. After the index of the web page is established, the entity in the attribute value dictionary is traversed, and the web page associated with each entity is queried. In addition to the corresponding entity name in the page describing the entity name, the entity name may also appear in other pages. Therefore, using the entity name as a keyword to perform the query is not desirable. Instead, an <entity name, attribute value> is used. Search as a keyword so that you can find the actual relevant web page. Obviously the description of an entity has at most one web page in the same type of web page. Therefore, only Top1 is taken as the relevant web page of the entity, and finally it is converted into a mapping relationship of <web page, entity>. It should be noted that a web page may correspond to multiple entities, and a large number of web pages may not have corresponding entities.

4.3 Webpage annotation

The following uses webEntityMap to represent the <webpage, entity> mapping relationship <webpageName, entitySet> established in the previous section. In order to facilitate labeling the data item of the attribute p in the web page, each page in the map is labeled with the value of the attribute p of its corresponding entity. The algorithm is described as follows:

Algorithm Create an automatic page labeling algorithm for webpage entity mapping

Enter The next set of webpages to be annotated in the domain D. The webpage is represented in the format of <xpath,text_node>, and the attribute value dictionary Dict() of the attribute p to be annotated, the webpage entity mapping webEntityMap

Output Label the data item corresponding to the attribute p

```
propery Value Set = { } ;
// Iterate over the collection of web page entity mappings, labeling the pages within them
for ( webpageName，entity Set) in webEntityMap:
webpage = get Page( W，webpageName) ;
//Acquire a set of attribute values for the webpage corresponding entity
for entity in entity Set:
current Property Value = get Value(Dict(),entity) ;
property Value Set.add( current Property Value) ;
// Use a property value to tag a page
for( xpath，textNode) in w ebpage:
if text Node in property Value Set
Label(xpath，Text Node，p) ;
```

In Algorithm , it is known that the current page w describes the entity e, and the value of the attribute p of the entity e is pValue. Therefore, all the text nodes in w that have the same value as the pValue are labeled as data items of the attribute p.

5. Conclusion

This paper applies the Chinese Knowledge Graph CN-DBpedia to the automatic annotation of training data in Web geographical information extraction. First, get geographical name data from Wikipedia and get the original place name set. Then all the entity objects corresponding to the original place-name set are obtained by CN-DBpedia, the ambiguous entities are filtered out, and a small amount of manual processing is performed. Then, all the triplets of the location entity are read out to generate the attribute value dictionary, and the automatic labeling of the web page is further

completed.

Although the study in this paper only requires a small amount of manual processing, in practice, the amount of geographical name data is too large, and there are still problems with the extraction of place names throughout the Internet. Moreover, when generating the attribute value dictionary, the size of the attribute value dictionary directly affects the efficiency and accuracy of the web page annotation. The next step will be to conduct more in-depth research on these two issues.

Acknowledgements

This work is supported by "the National Natural Science Foundation of China (Grant: 61602387)"

References

[1] Meng Xiaofeng. A Survey of Web Data Management Research[J]. *Journal of Computer Research and Development*, 2001, 38(4):385-395.

[2] Chen Yi, Zhang Dongmei. A Survey of Web Information Extraction Techniques[J].*Application Research of Computers*,2010,27(12):4401-4405.

[3] Wei Yong, Li Hongfei, Hu Danlu, Li Xiang, Ma Leilei. A method of Chinese geographical names recognition based on compound features[J]. *Journal of Wuhan University (Information Science Edition)*, 2018, 43(01):17-23.

[4] Fudan University Knowledge Factory Laboratory. CN-DBpedia [EB/OL].[2016-04-11]. http://gdm.fudan.edu.cn/CKGraph/.

[5] Liu Ye, Li Yang, Duan Hong et al. Review of Knowledge Graph construction techniques[J]. *Journal of Computer Research and Development*, 2016, 53(3): 582-600.

[6] Xu Zhihao. Research on Chinese Named Entities Corpus Construction Based on Wikipedia[D]. *Soochow University*, 2016.

[7] Wang Zhibao, Xia Xin, Wang Chengbo. Review of key technologies for geographic information retrieval [J]. *Computer Engineering and Science*,2018,40(03):533-543.

[8] Banko M，Cafarella M J，Soderland S，et al．Open Information Extraction from the Web[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence．*Hyderabad，India:[s.n.]*，2007: 2670-2676．

[9] Muslea I，Minton S，Knoblock C．Hierarchical Wrapper Induction for Semistructured Information Sources[J]．*Autonomous Agents and M ulti-Agent Systems*，2001，4(1):93-114．

[10] Soderland S．Learning Information Extraction Rules for Semi-structured and Free Text [J]．*Machine Learning*，1999，34(1-3):233-272．

[11] Song Dandan, Wu Yunpeng, Liao Lejian, et al. A Dynamic Learning Framework to Thoroughly Extract Structured Data from Web Pages Without Human Efforts[C] //Proceedings of ACM SIGKDD Workshop on Mining Data Semantics．*New York，USA: ACM Press*，2012:1-8．