Automatic Acquisition and Semantic Annotation of Web Tourism Information

Hui PENG¹ and Wen-qi QU²

¹School of Tourism Sciences, Beijing International Studies University, Beijing, China
²Department of Basic Education, Beijing International Studies University, Beijing, China

Keywords: Tourism information, Word segmentation, Keywords, Semantics, Semantic annotation.

Abstract. Data collection and semantic annotation is often the basic of information processing such as semantic relation analysis of data, big data mining and semantic information search et.al. A method which collects data from tourism web site and annotates these data with semantic tags automatically is promoted in this paper. The crawler which collects data from web site automatically is introduced firstly. Then the Chinese word segmentation tool and a classic key word extraction algorithm TF/IDF are introduced. With the help of a crawler, we collection tourism information about 247 sight spots in Beijing and 4198 sight spots in other area of China from the web sites of elong and ctrip. Then with the help of the ICTCLAS and TF/IDF, we abstract keywords from the information as semantic tags to annotate the sight spots.

Introduction

Since the 21st century, with the rapid development of information technology and the development of smart tourism supported by the new generation of information technology, the development of smart tourism information integration project is an important foundation and symbol of the modernization of tourism services. After years of exploration and accumulation, there are many domestic tourism websites with certain service capabilities, which can provide more comprehensive tourism information. The constant enrichment of tourism information provides a data basis for smart tourism. How to make effective use of such data and how to dig the relationship among users, tourism products and tourism services from existing data accumulation to make tourism services more accurate and effective has become an important topic of smart tourism research.

Wisdom of tourism research is an important part of tourism information semantic relations research, through the establishment of tourist information semantic relation model, the computer system can automatically analyze the tourism potential relationship between data, the correlation degree between tourism data, scattered at the heart of the hot spots implied in the tourism data, etc., thus showing the laws between the data in order to be used by people.

In recent years, there are some researches on semantic relation of tourism information at home and abroad. Reference [1] takes the official website of Tibet tourism, the website of tourism enterprises and the website of tourism guide as the main research content, and adopts the method of combining semantic network analysis and semantic network clustering to analyze and study the tourism image of Tibet spread by Chinese Internet. Reference [2] utilizes ontology knowledge, combines existing geography and tourism ontology, extracts core concepts of traditional tourism geography subject knowledge, and constructs tourism geography ontology model based on new concepts related to tourism geography in the era of big data. Reference [3] proposed a knowledge annotation scheme from the perspective of multiple fields by combining multiple domain subject word lists for semantic annotation, laying a foundation for the construction of knowledge base. Reference [4] USES Bayesian algorithm for automatic tourism classification, including concept tags related to tourism. The results show that Bayesian algorithm is very effective for tourism classification. Reference [5] proposes an automatic image semantic annotation method based on ontology knowledge base of tourism domain.

Tourism information acquisition and semantic annotation are the research basis and important components of the semantic relationship of tourism information. Only after obtaining enough data...
Acquisition of Web Information

Web crawlers are a common and effective means to obtain information from Internet pages. A web crawler, as the name implies, is a program that starts from one or more network addresses (URLs), automatically obtains information on the URL and continuously transfers to the URL with which it is linked to obtain information.

Generally, there are two types of web crawlers: general purpose web crawlers and directional web crawlers. General web crawler is a particular topic, general procedure of all effective web page collection, as long as through the crawler control policy setting seed URL to crawl through the search engine with a seed URL for the entrance, on the Internet all the web pages crawled, constantly new URL can be drawn from the page into the queue, until meet system must stop condition, to stop work.

The second type of directional web crawler is also called the theme web crawler. It crawls according to the theme given in advance when crawling the web page and crawls selectively. Analyze hyperlinks and downloaded web content, predict the next URL to be grabbed and the topic relevance of the current web page, crawl and download as many pages related to the theme as possible, and reduce the number of page downloads unrelated to the theme as far as possible. The webpage crawled by a targeted web crawler is highly correlated with the target topic, so when the user retrieves information, the accuracy is higher than that of the data crawled by a general crawler [6].

WebMagic[7] is a crawler framework that requires no configuration and is convenient for secondary development. It provides a simple and flexible application programming interface, and a crawler can be realized after programming. In this paper, WebMagic is used to realize a targeted web crawler tool, which can obtain webpage information from elong and other tourism websites, and store the information in the database. The information includes relevant text information of 247 tourist attractions in Beijing and 4198 tourist attractions in other parts of China. Some data are shown in Figure 1.

| id  | feature          | value | history                      | intro                                      | url                    |
|-----|------------------|-------|------------------------------|-------------------------------------------|------------------------|
| 1   | Xiang Mountain  | OK     | history                      | climate of Xiang mountain, price          | The Xiang Mountain     |
| 2   | NULL             | OK     | NULL                         | national evening pool is called water    | Water Cube             |
| 3   | NULL             | OK     | NULL                         | the street of bar is located in a         | Audito                 |
| 4   | NULL             | OK     | NULL                         | thousands of miles, spectacular          | The Great Wall         |
| 5   | NULL             | OK     | NULL                         | the famous stuff in beijing              | XiJian                 |
| 6   | NULL             | OK     | NULL                         | The travel guide of tibet park           | Tibet Park             |
| 7   | Happy Valley    | OK     | history                      |_HOST reservation around happy valley     | Happy Valley           |
| 8   | NULL             | OK     | NULL                         | Lama Temple is located in beijing        | Lama Temple            |
| 9   | NULL             | OK     | NULL                         | Large comprehensive museum               | Capital Museum         |
| 10  | NULL             | OK     | NULL                         | The travel guide of beijing zoo          | Beijing Zoo            |
| 11  | NULL             | OK     | NULL                         | introduction of discount, hiking         | Shichahai              |
| 12  | NULL             | OK     | NULL                         | introduction of ticket, open time        | railway museum         |
| 13  | NULL             | OK     | NULL                         | UCCA is an art center, is located in     | The China Center for Contemporary Art (UCCA) |
| 14  | NULL             | OK     | NULL                         | price, open time, best travel time       | sharing hot spring     |
| 15  | NULL             | OK     | NULL                         | Olympic Park                              | the Olympic Park       |
| 16  | NULL             | OK     | NULL                         | price, open time, best travel time       | Museum of modern      |
| 17  | NULL             | OK     | NULL                         | price, open time, best travel time       | National museum       |
| 18  | NULL             | OK     | NULL                         | introduction of changshan temple         | changshan temple       |
| 19  | NULL             | OK     | NULL                         | price, open time, best travel time       | Beijing Art Museum     |
| 20  | NULL             | OK     | NULL                         | price, open time, best travel time       | Beijing Art Museum     |

Figure 1. Crawling travel information.

The main functional parts of WebMagic crawler include: URL management, page download, page analysis and link extraction, and the processing of collected data, and the web crawler organizes them together to form the core part. The functions of the core part are: first, the page download component is responsible for downloading travel-related pages from the Internet. Then the page analysis and link extraction component is responsible for analyzing the page, extracting travel-related information and discovering new links. Secondly, the URL management component is responsible for managing the travel-related urls to be fetched and removing some duplicate urls. Finally, the processing component of data collection is responsible for the extraction of results. The core of WebMagic (WebMagic-core) is a compact, modular piece of content that requires some very practical functionality to be extended. Extension the functions are: first of all, in such as elong travel web crawl we need travel information, we use the Xpath parsing and regular expression matching method, to tourist attractions name (title),
scenic spot of the link (URL), scenic spots (type), the scenic spot historical background (history) and attractions (feature) for extracting the characteristics of, will go to extract the content of heavy, finally will extract content in the MySQL database.

**Text Content Processing**

Tourism information obtained from web pages is often large blocks of text. In order to carry out semantic annotation on such information, the computer needs to understand the content of the text and select appropriate annotation words for semantic annotation. Words and phrases are the smallest unit to understand the semantic meaning of text. Due to the characteristics of Chinese language, the first basic link of Chinese language analysis and processing is word segmentation. ICTCLAS, language cloud of Harbin Institute of Technology, Paoding Chinese word segmentation tool is currently several commonly used Chinese word segmentation tools. Among them, ICTCLAS[8] developed by institute of computing science of the Chinese academy of sciences is the earliest open source Chinese word segmentation software, which has the characteristics of fast word segmentation speed (word segmentation speed of single machine 996KB/s), high accuracy (98.45%), wide use of platform (WINDOWS, LINUX, FREEBSD), less system resources (API not more than 200KB, various dictionary data compression less than 3M), and constantly updated version.

ICTCLAS word segmentation software includes Chinese word segmentation; Part of speech tagging; Named entity recognition; Neologism recognition; User dictionary definition and other functions; It also supports GBK encoding, UTF8 encoding and BIG5 encoding. Therefore, ICTCLAS word segmentation tool is finally selected in this paper to process the obtained tourism information word segmentation.

Using the ICTCLAS word segmentation tool, we successively completed the following tasks :(1) to construct a user dictionary about tourism, extract special nouns related to tourism, such as place names, scenic spots and names, and generate a custom user dictionary; (2) add the custom user dictionary to the word segmentation tool and conduct word segmentation for the text information in the database; (3) building stop using dictionary word segmentation software to get rid of the text analysis of the follow-up work meaningless pronouns, adverbs, prepositions, conjunctions, modal particles, particles, interjection, onomatopoetic words, prefix, suffix, keep nouns, verbs, adjectives, numerals and quantifiers, word denoting time and the place word and so on, and add the space between the word after word; (4) store the processed text data in the form of .txt and excel, as shown in figure 2.

**Key Word Extraction**

After the processing of tourism text, it becomes more standard text data. For each scenic spot text, extract the key words that can represent the characteristics of each scenic spot. In this paper, TF-IDF feature is used to extract keywords. TF-IDF is a scientific measure of the weight of search keywords and a common weighted technique used for information retrieval and information exploration. The main idea of TF-IDF is that if a certain word or phrase appears frequently in one article and rarely in other articles, it is deemed that this word or phrase has good category differentiation ability and is suitable for classification. In literature [9], based on the traditional TF-IDF and combining the position feature and length feature, the improved TF-IDF calculation formula was proposed to extract key words considering the different part of speech of concurrent words. In literature [10], TF-IDF algorithm was used to calculate the weight of keywords, and the obtained weight was applied to the vector space model and Markov model to obtain the semantic similarity and basic similarity, and finally to obtain the overall text similarity.
The specific steps to extract keywords with TF-IDF are as follows: 1. Formula (1) is used to calculate word frequency: TF is word frequency, and the frequency of each word appearing in the text is calculated by dividing the frequency of a word appearing in the text by the total number of words in the text. 2. Formula (2) is adopted to calculate IDF inverse document frequency. 3. Formula (2) is used to calculate the value of TF-IDF. The value of each word is arranged in descending order to get the key words of the text. The formula is shown as follows:

\[ \text{TF} = \frac{\text{The number of times a word appears in the text}}{\text{Total number of words in the text}} \]  
\[ \text{IDF} = \log\left(\frac{\text{The total number of documents in the corpus}}{\text{Number of documents containing the word}}\right) \]  
\[ \text{TF-IDF} = \text{TF} \times \text{IDF} \]

Finally, we get the key words of Beijing scenic spots.

**Semantic Annotation**

According to the processed tourism data, this paper classifies and marks the scenic spots in Beijing. Firstly, the processed data are classified by the top-down method, which is mainly divided into three categories: parks, venues and sites. Each major category is divided into several small categories, such as the park category includes four subcategories: the city park, forest park, picking park and playground; Under the venue category, there are five subcategories: museum, art gallery, memorial hall, former residence and gymnasium. Under the sites category, there are temples, royal gardens, the Great Wall, Hutongs, ancient villages and ruins.

Under each subclass are instances of attractions, and this article USES Protege5.0 to build ontologies. The description of Protege in literature [11] is that "Protege is an ontology development software developed by researchers in Stanford university according to the needs of ontology construction. In order to achieve the compatibility and openness of the tool software to other languages, the development of Protege software adopts object-oriented language -- Java language". Protege’s simple and convenient operation interface and some help functions provide a high degree of convenience for the construction of ontology. It also supports the interaction between multiple systems and enhances the universal applicability of software. In this paper, Protege is used to construct ontology, as shown in Figure 3.
Summary

Tourist information acquisition and semantic tagging is the premise and foundation of tourism information intelligent processing. On the basis of semantic annotation, the semantic processing including semantic relation analysis of tourism information, semantic retrieval, information management and recommendations, information evaluation and a series of intelligent management method, through the effective management and utilization of information, can provide tourism planning and better services.

Acknowledgement

This research was financially supported by the National Social Foundation (14BTQ031).

References

[1] Wang Xiaohui. Research on the tourism image of Tibet via the Internet—semantic network analysis based on the text of Chinese tourism websites [J]. Guizhou ethnic studies, 2014, 10:165-168. (in Chinese)

[2] Hong Ye, Kang Mingjuan, Li renjie, et al. Design of tourism geographic ontology model and case study of zhangjiajie [J]. Geography and geographic information science, 2016, 32(3): 95-99. (in Chinese)

[3] Li Peng. Research and implementation of knowledge annotation from the perspective of multiple fields [J]. Information engineering, 2016, 1:013. (in Chinese)

[4] Peng hui. Annotating Tourism Resources with Social Tags and Classifying Them with Bayesian Algorithm [J]. International Journal of u- and e-Service, Science and Technology, 2016, 9 (5): 419-428.

[5] Zhang P, Du J, Fan D, et al. Automatic Image Semantic Annotation Based on the Tourism Domain Ontological Knowledge Base[C]//CCF Chinese Conference on Computer vision. Springer Berlin Heidelberg, 2015:61-69.

[6] Yang Jingtao, Chen Huigu. Research on web crawler technology [J]. Journal of science and technology entrepreneurship, 2010(10):170-171. (in Chinese)

[7] http://webmagic.io/

[8] http://ictclas.nlpir.org/
[9] Niu Ping, Huang Degen. Research on automatic extraction of Chinese keywords combined with tf-idf and rules [J]. Miniature microcomputer system, 2016, 37(4): 711-715. (in Chinese)

[10] Zhou Lijie, Yu Weihai, Guo Chen. Text similarity algorithm based on improved tf-idf method [J]. Journal of Taishan University, 2015 (3): 18-22. (in Chinese)

[11] Li Qingsai. Research on ontology construction in the tourism field [D]. Zhengzhou University, 2015. (in Chinese)