RESEARCH ARTICLE OPEN ACCESS

The Study of Web Text Processing Base on Cyber Retrieval

FU Danlong

(School of Information science and technology, Computer Architecture, Jinan University, 510632)

--************************----------------------------------

Abstract:

As a rising number of information online, we can easily obtain a large quantity of information, but to process them with manpower will be divorced from reality. Targeted the accounting reports and the contracts which are clientele oriented and scientific, this study tried to discuss a solution based on Information Extraction(IE) to obtain such kind of information from Internet by cyber retrieval, and then format these information by Natural Language Processing(NLP), the essence of this work is to process Internet text into structured data. This study obtained more than 20000 pieces of PCT information from the WIPO website server by using web crawler, and took the valuable information at particular locations. This study also tried to process quoted company announcements into valuable structured data by NLP, this study proposed the solution to build a system that to process professional text, which is written by natural language (Chinese) and from the CNINF website server, into structured data, on the base of the Domain Ontology Knowledge Lib. This study also did a simulation experiment, to prove the feasibility of the system.

Key words: **Information Extraction; cyber retrieval; web crawler; Domain Ontology.**

--************************----------------------------------

On how to transform text to structured data, domestic and foreign research have gone through several stages. IE was first initiated by the Message Understanding Conference (MUC) in 1987, but the main work of the year was mainly for news and reporting on terrorist attacks. In the last century, from 80 to 90s, the IE technology driven by MUC mainly completed the entity name recognition, CO reference resolution, template relation extraction, template filling, CO referential relationship determination and so on[1]. At present, information extraction is mainly used in the social aspects of the Internet, but often in a specific field.

At present, the main research work in the field of information extraction technology in China is the recognition of Chinese entities. A number of literatures [2-6] introduced the research of current domestic scholars to improve the extraction performance and accuracy. The ideas of these studies provide guidance for the future research of information extraction technology. Another important research on information extraction is the research of domain ontology. Research [4, 7] pointed out that the current Chinese information extraction is still dependent on the construction of domain ontology, and the field is targeted. Cai [4] built the disease naming domain into domain ontology information and propose the "maximum entropy model of Chinese disease named phrase recognition method", which improved the accuracy of the disease named phrase recognition and recall rate.

Chinese domain ontology construction methods have been done by scholars. Research [8, 9] has studied the construction and maintenance of domain ontology in greater depth.

This article mainly establish a set of automated financial text mining system, through the construction of web crawler system to obtain text information, with the text processing technology to access to the main content of financial reports.

1. Web Search For Text

The data users see or get from the Internet are the response of server to a certain request. We can send this request by clicking the link or button in the web page. But when we already know how these scripts sending request to the server, we can use the computer program to send the same request to the server without clicking on the button. We do not even need to load the page in a browser to get the resource.

There are many forms of request, such as links, POST method, GET method, AJAX technology, etc. Different types of requests will return different types of data, and the browser will process the data and present it to the user.

In essence, a crawler program is a computer program for batch requests. The basic process is to access links - get links, the loop can theoretically traverse every open web page on the internet. Using the crawler to send requests to the server, we can search all the information of the Internet with the high speed of the computer, which provides the technical foundation for the large-scale Internet information retrieval and analysis.

In order to study the information disclosure of commercial enterprises, we choose CNINF website as sources of information disclosure of listed companies. CNINF is the first website to disclosure the company information and market data of more than 2500 listed company in Shanghai and Shenzhen. It is one of the earliest securities information professional website. To research for financial information, we select the CNINF website as sources of information disclosure. Some more targeted approach are taken: the page structure of the website are analyzed and those useful links are identified (through the position of the link on the page, link context or keywords), which will reduce the traversal times and improved the efficiency.

2 Text Processing

Most of the information on the Internet is using the natural language as the carrier, which is easy for human readers to understand. And from the perspective of interaction, this kind of information is friendlier. With the explosive growth of Internet information, it is impossible to analyze all the information by manpower. For example, the number of patent application with HUAWEI since 2000 is more than 25000 (source: World Intellectual Property Organization public data), and the announcement number released by China high-speed rail (A shares 000008) since 2000 is more than 1000 copies (source: CNINF public data).To extract a small amount of entity relationship from that by manpower, the efficiency is too low.

So the work must be done by the computer. However, such as "domestic downtown pressure on the economy increase, commodity prices fall depth, international financial market turmoil intensified, face three superimposed situation and adverse environment, will have a direct impact on the company's future performance" such a statement, not to mention the computer, is the ordinary readers may not be able to understand what it means.

The way people understand the statement is as follow: human beings first learn to speak (ie, the accumulation of words, the accumulation of grammatical sentences) and then contact the field of knowledge (learning, access to terms, learning the relationship between entities), and finally combined with human's excellent association and knowledge migration ability, people can read a highly specialized sentence containing a variety of data.

According to this point, this research is necessary to use computer program to analyze the professional articles for human readers to a certain extent, to identify and sort out the useful structure of the data group, the research work must contain three parts: 1. Lexical analysis And syntactic analysis; 2. domain ontology knowledge base construction; 3. intelligent pattern matching and semantic relation extraction.

The purpose of this study is to try to deal with a section of the professional text (listed companies annual report), the annual report is a public announcement of listed companies, although written by natural language, but professional, rigorous, less ambiguous. The article intends to convert the information expressed in natural language into structured data.

2.1 Natural Language Processing

An ordinary reader has the flexible language ability to understand the interpretation of a sentence, which is based on a person's long-term use of the corresponding language. In the field of computer science, many scholars have given a feasible solution to divide natural language processing [10] (here refer to natural Chinese language processing) into three levels:

Lexical analysis: including word segmentation, partnered annotation, named entity recognition, word meaning disambiguation and other research.

Syntactic analysis: split the structure of the sentences, analyze relationship between the between the words and the relationship between the interpretation, explain the various components of the sentence. As the focus of Chinese natural language processing, syntactic analysis is a hot topic in natural language processing research in recent years. At present, there are two kinds of syntactic analysis labeling system: Phrase structure syntactic system [11] and Dependent structure syntactic system [12]. In contrast, dependency syntax has a more concise, more efficient and flexible advantage [13].

Semantic analysis: Exploring the true meaning of extracting sentences, preserving semantic content as structured information. The current research community is still arguing the general form of semantics. Semantic Role Labeling (SRL) [14,15] is a relatively mature shallow semantic analysis technique.

2.2 Pattern Matching

Before the pattern matching, the domain ontology knowledge base is also needed, which should include the necessary ontology corpus in the field, enough patterns, enough semantic interpretation rules and self-expansion and learning functions

Compared to the process of reading and understanding of human beings, computer systems do not have some flexibility in natural language processing. The main way is through the pattern matching to access to the main components of the sentence. By comparing the relationship between the predicate and the number of subject and object, we can further understand the emotional tendencies of sentences and carry on semantic analysis.

The text of this paper is the company's annual report, whose content and form have certain structure. After the primitive accumulation of syntactic structure, the main information can be extracted by pattern matching to lay the foundation for the following analysis.

3 Experiments

3.1 CNINFO Website

The CNINFO website is designated by the China Securities Regulatory Commission to list company information disclosure website. It is the earliest professional website of securities information.

This study tries to extract some public online announcement of listed companies which meet certain conditions in CNINFO website, and download all of them for subsequent analysis. CNINFO set up a powerful search function, which can precise position the announcement we want. But CNINFO does not allow batch download these announcements. only to a human click on the download link, in the face of a large number of download analysis announcement, this approach is clearly too inconvenient. We use crawler program to solve this problem.

3.2 Crawler Program

The information structure of CNINFO website is as follow:

Homepage (search page) → search result page(Bulletin list) → Download link → announcement PDF (text)

In order to obtain the text of the announcement in batches, the procedure is executed as follows:

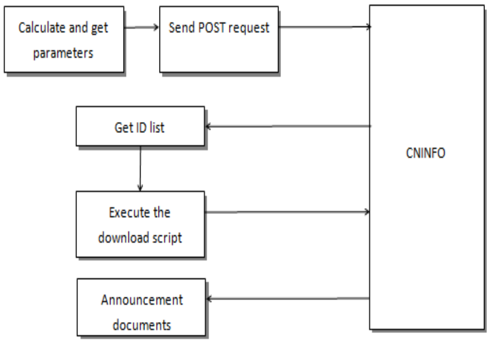

Figure 3.1 The information retrieval process

The search conditions of CNINFO website is integrated in the POST request header, and the format is basically as follows:

```
{
    stock=[ Stock code]
    searchkey=[ keyword]
    category=[ 21 types of announcements ]
    pageNum=[ Results page number ]
    pageSize=[ The number of single page items, the maximum value is 50]
    column=[ Exchange information]
    tabName=[ Specifies the lookup policy, ”latest” or “fulltext”]
    sortName=[ Sort name]
    sortType=[ Sort type]
    limit=[ the number of results]
    seDate=[ Time interval setting]
}.
```

3.3 Text Processing

First, we need to establish domain terminology corpus, including termbase, predicate termbase, physical library. The storage of professional vocabulary corpus should be structured. Through the establishment of professional lexical corpus, the system can then follow the syntax analysis to identify which is the entity and the entity type is what.

Second, build a model library of known patterns as follow:

```
IF{
    Subject: [The name of the organization];
    Core predicate: “Signed”;
    Object: [Modified protocol type];
    Modifier of the core predicate:[ Time phrases];
}THEN{
    The party of the transaction =subject;
    Transaction Type =” sign the agreement”;
    Protocol = object;
    Signing time = Modifier of time phrase of the core predicate;
}END;
```

(Note: This is only the contents of the mode use case, the actual storage structure is not the case)

Through the above pattern, we can process the data in the sentence into structured data to extract the information.

The extraction of information from text to structured data is the focus of this research. The Chinese text involved in the experiment is for human readers, and human readers have very good flexibility, and it is difficult for computer programs to do this. But we can still do semantic analysis of a field (such as disclosure of information from listed companies).

Its core work is to calculate the semantic vector of the statement of the entity in the preprocessed text. Then find the domain ontology knowledge base, match the vector values close to the example of the statement and match the corresponding pattern. At last, the program will know the internal semantic relations of the source statement. The main work of calculating the semantic vector to match the corresponding pattern is explained by the domain ontology knowledge base

The ontology knowledge base provides the ability to identify entities and terms when the lexical syntax is analyzed. When the matching pattern is needed, the semantic vector of the source statement is calculated according to the semantic interpretation rule, and the pattern closest to 1 is found from the pattern library.

The specific process is shown below:

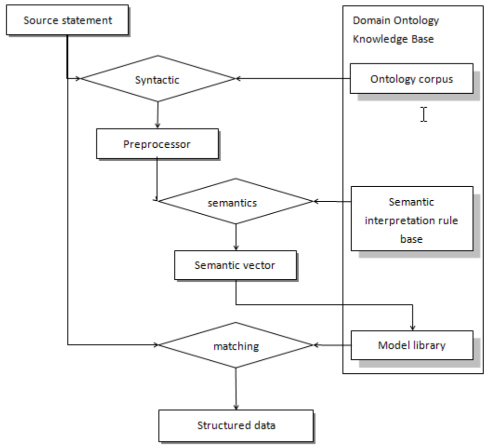

Figure 3.2 Text Processing Process Based on Domain Ontology Knowledge Base

4. Conclusions

According to the characteristics of the customer and the science of the contract legal documents for the accounting report and the patent information, this study attempts to explore a solution based on information extraction technology. The crawler system is used for data acquisition. Through the text processing technology, the network text is transformed to the structured data. On the basis of reasonable construction of domain ontology model, automatic data extraction is completed. Finally, through the simulation experiment of CNINFO website data, the feasibility of the system is demonstrated.

References

[1] Yu Yongbo, Research on Some Key Problems in Web Information Extraction[D]. China University of Science and Technology, 2015

[2] Huo Yage, Huang Guangjun. Recognition Method of Chinese Phrase Structure Based on Maximum Entropy [J]. Computer Engineering, 2011, 37(16)

[3] Yu Hongkui, Zhang Huaping. etc, Chinese named entity identification using cascaded hidden Markov model[J]. Journal on Communications, 2006, 20(5): 40-50.

[4] Cai Xiaobai, Fan Xiaozhong. Maximum Entropy Method in Recognizing Disease Named Phrase in Chinese[J]. Journal of beijing institute of technology, 2006, 26(6):517-520.

[5] Zhu Qiaoming, Li Peifeng. Chinese information processing technology tutorial [M]. Tsinghua University Press, 2005, 254-255.

[6] Li Lei, Zhou Yanquan, Wang Jinghua. Comprehensive Information Based Chinese Information Extraction System and Application [J]. Journal of Beijing University of posts telecommunications, 2005, 28(6):48-51.

[7] Zhao Yanyan, Qin Bing, Che Wanxiang. Research on Chinese Event Extraction[J]. Journal of Chinese Information Processing, 2008, 22(1): 3-8.

[8] Liu Baisong, Gao Ji. A Study on Ontology Learning for the Knowledge Grid [J]. Computer Engineering and Application. 2005, 41(20):1-5

[9] Wang Xue. Research on the Method of Ontology Construction in Chinese [D]. Huazhong University of Science and Technology, 2012.

[10] Li Zhenghua. Research on Key Techniques of Chinese Dependence Syntax Analysis [D]. Harbin Institute of Technology, 2013.

[11] Collins M. Head-driven statistical models for natural language parsing[D]. Univiersity of Pennsylvania, 1999.

[12] Mel’cˇuk I A. Dependency Syntax: Theory and Practice[M]. State University of New York Press, 1988.

[13] Ma Jinshan. A Corpus - based Analysis of Chinese Dependence Based on Statistical Methods [D]. Harbin Institute of Technology, 2007.

[14] Che Wanxiang. Research on Semantic Role Marking Based on Nuclear Method [D]. Harbin Institute of Technology, 2008.

[15] Li Junhui. A Semantic Analysis of Chinese Syntax and Its Joint Learning Mechanism [D]. Suzhou University, 2010.