ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Research and Implementation of TCM Knowledge Acquisition Based on Open Data Source

Yonghong Xie, Yanxuan Qian, Shuang Ha and Dezheng Zhang*

School of computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Beijing Key Laboratory of materials science and Engineering, Beijing 100083, China

*Corresponding author

***Abstract*—With the development of online information resource, open data sources are becoming an important channel for domain knowledge acquisition. To establish a comprehensive knowledge base of TCM, this paper choses Baidu Encyclopedia as data source by comparing the data and web structure of the main open TCM related data sources on the Internet. Combining web crawler and pattern matching, this paper studied the related technologies of acquiring TCM knowledge based on open data source and implemented an efficient automatic extension of TCM attributes and attribute values.**

Keywords-open data source; TCM; knowledge acquisition; attribute extraction

I. INTRODUCTION

With the development of online information resource and Internet technology, the Internet gathers data, information and knowledge from all walks of life. How to obtain effective knowledge for specific domain services from open data sources is a hot research topic in recent years.

From now on, the extraction of entity attributes abroad is extensive. Poesio et al [1] used syntactic patterns to extract candidate conceptual attributes from Web, after that, a directed two meta classifier is used to identify attributes by classifying attribute discrimination as a classification problem; Yoshinaga et al [2] proposed an unsupervised method for extracting attributes and attribute values from HTM documents. Those two methods that mentioned above are only suitable for semi-structured or structured documents, and the portability is poor.

In term of Chinese information acquisition, XianYi Cheng and Qian Zhu [3] obtain seed relationship from the information box in Wikipedia Chinese, find strong counterexamples by linear classifier, then train the classifier to find more counterexamples. The semi supervised learning is used to obtain the relation candidate instances, and then the clustering is used to determine the relation categories; JianYi Guo and Zhen Li et al [4] use domain text as the data source, and propose a collaborative classifier to solve the extraction of domain concepts, attributes and attribute values, and to predict the corresponding relationship between the three; Wang et al[5] get concepts, concepts, hierarchies, and conceptual attributes from the Interactive Encyclopedia and the Baidu Encyclopedia's classification system and semi-structured information box.

In order to make full use of the rich knowledge in the field of TCM on the Internet, promote the electronic and automation process of TCM and establish a knowledge base of TCM Ontology, this paper takes the names of TCM as an example, selects suitable open data sources and efficient knowledge acquisition technology to obtain and supplement the 27 attributes and attribute values of TCM.

II. ACQUISITION OF ATTRIBUTES AND ATTRIBUTE VALUES OF TCM

A. The Selection of Open Data Sources

According to the definition of Open Knowledge Foundation of the UK [6], openness requires the following 3 basic elements:

- Non-discrimination: If the data is open, it is open to everyone.
- Machine-readability: If the data is open, it should be machine-readable. (e.g. the table data should be in .csv instead of in .pdf)
- Open license: If the data is open, the license should make sure that the user have rights of access, obtain, use, addition, deduction, copy, dissemination freely.

Meanwhile, the selection of open data source should pay attention to the normalization of web pages, the rationality and the comprehensiveness of Web Information. In other words, the selection of optimal open data sources mainly includes document specification, structure specification, naming specification, layout specification and content specification [7].

First of all, this paper compared the most famous five open data sources for the comprehensiveness of Web Information on TCM by searching the 1000 kinds of TCM that randomly selecting from 11115 kinds of them in Baidu Encyclopedia, Chinese Pharmacopoeia, TCM Encyclopedia, Interactive Encyclopedia and 360 Encyclopedia individually. The hit rate of searching traditional TCM in these five data sources is shown in Figure 1. Apparently, Baidu Encyclopedia has more TCM information than the other four open data sources.

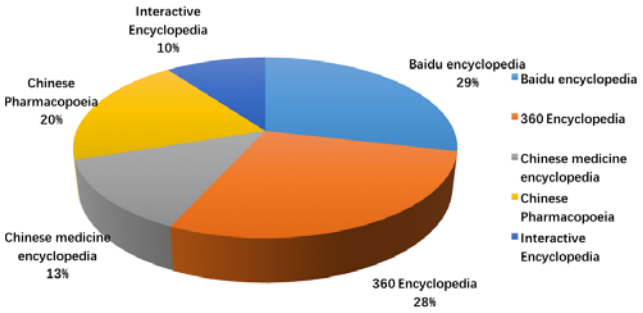


FIGURE I. COMPARISON OF THE HIT RATE OF SEARCHING THE SAME TCM IN FIVE DIFFERENT DATA SOURCES.

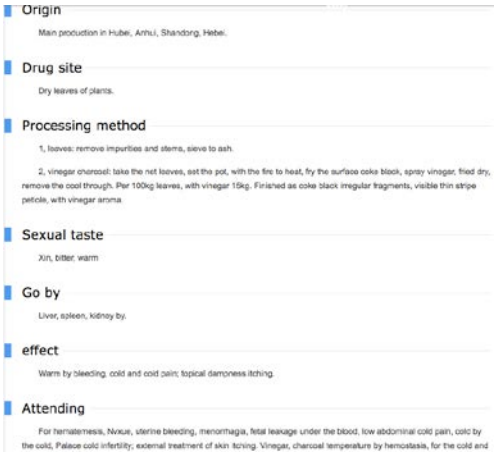
It can be shown in Figure 2 that there is an information box in the Baidu Encyclopedia page of Leaves. This is the semi-structured data which contains a large number of attribute relationships. For example, "Chinese scientific name", "alternative name" and "boundary" are the attributes of "Leaves", meanwhile, "Leaves", "Leaves or Artemisia Argyi," and "Ai", "plant kingdom" are the corresponding attribute values.

Obviously, the information box of Baidu encyclopedia is an important source of access to the attributes and attributes values of traditional TCM.

Chinese name	Leaves	Genus	A genus
nickname	Ai Ye, Ai Hao, home Ai	Species	Ai
Sector	Plant community	distribution a...	Northeast, north, east, southwest and Shaanxi,
door	Angiosperm door		Gansu and other places
Outline	Dicotyledon	Harvest time	Summer flowers are not open when picking
Purpose	Platycodon head	Dosage	3 ~ 9g
Section	Asteraceae	Toxicity	There is a small poison
		Storage	Cool and dry place

FIGURE II. THE INFORMATIN BOX OF “LEAVES”

However, the content of Baidu Encyclopedia is mainly based on unstructured text data. So, it’s very important to achieve the extraction of attributes and attribute values of TCM from unstructured text. For example, Figure 3 shows the Part of the text in Baidu Encyclopedia page about “Leaves”.


Origin

Main production in Hubei, Anhui, Shandong, Hebei.

Drug site

Dry leaves of plants.

Processing method

1, leaves: remove impurities and stems, sieve to ash.

2, vinegar charcoal: take the net leaves, set the pot, with the fire to heat, fry the surface coke black, spray vinegar, fried dry, remove the cool through. Per 100kg leaves, with vinegar 15kg. Finished as coke black irregular fragments, visible thin stripe petiole, with vinegar aroma

Sexual taste

Xin, bitter, warm

Go by

Liver, spleen, kidney by.

effect

Warm by bleeding, cold and cold pain; topical dampness itching.

Attending

For hematemesis, Nvxue, uterine bleeding, menorrhagia, fetal leakage under the blood, low abdominal cold pain, cold by the cold, Palace cold infertility; external treatment of skin itching. Vinegar, charcoal temperature by hemostasis, for the cold and

FIGURE III. PART OF THE TEXT IN BAIDU ENCYCLOPEDIA PAGE ABOUT “LEAVES”

Apparently, “properties and taste” and “channel tropism” are the attributes of “Leaves ”, meanwhile, “xin, bitter, warm” and “channel tropism of liver, spleen and kidney” are the corresponding attribute values.

According to comparing the content in information box and text, it comes to conclusion that:

- Most of the pages about TCM in Baidu Encyclopedia have information boxes that contains attributes and attribute values.
- Different TCM has different kinds of attributes in their information boxes. So it’s hard to unified data.
- Because of the lack of uniform standards, the information boxes and texts of pages about TCM in Baidu Encyclopedia that written by different users have different problems about text structure and content.

To solve those problems that mentioned above, different schemes are adopted to obtain the attributes and attribute values of TCM from Baidu Encyclopedia information boxes and texts.

B. Web Crawler Driven by TCM Names

Web Crawler is the kind of program that automatically crawling information on the Internet according to certain rules [8]. This paper makes full use of 11115 kinds of existing TCM that have not stored their attributes and attribute values and proposes a kind of web crawler that driven by TCM names. The program automatically searches those 11115 kinds of existing TCM in Baidu Encyclopedia and saves the pages about them. Then, it parses those pages’ content in different ways, according to the differences of data structure. This process can be shown in Figure 4:

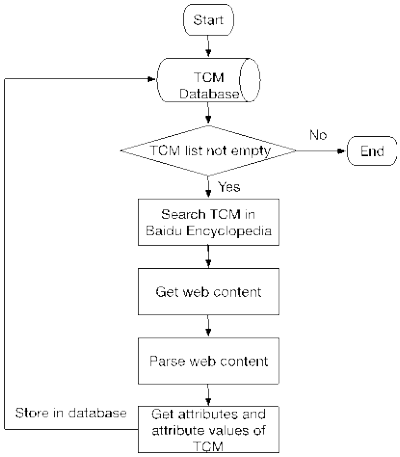


FIGURE IV. FLOW CHART OF WEB CRAWLER THAT DRIVEN BY TCM NAMES

C. Analysis of Baidu Encyclopedia Data Source

To locating important attributes and attribute values in the page of TCM, this paper needs to analyze the page of data source. Normally, the web page should be parsed into syntax tree by parser before extraction, then the extraction of text is transformed into the operation of syntax tree to extract

information by formulating some extraction rules. This process can be shown in Figure 5:

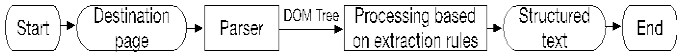


FIGURE V. FLOW CHART OF INFORMATION EXTRACTION BASED ON STRUCTURAL FEATURES OF WEB PAGES

In order to filter out irrelevant content in HTML text, this paper needs to analyze the web page. If the structure of this web page is simple, regular expression can be directly used to match what is required. However, if it's too complicated, as it is shown in Figure 6 and Figure 8:

Chinese name	Leaves
nickname	Ai Ye, Ai Hao, home Ai
Sector	Plant community
door	Angiosperm door
Outline	Dicotyledon
Purpose	Platycodon head
Section	Asteraceae

FIGURE VI. THE PARTICIAL INFORMATION BOX IN THE WEB PAGE ABOUT LEAVES

```
<dt class="basicInfo-item name">Chinese name</dt>
<dd class="basicInfo-item value">
艾叶
</dd>
<dt class="basicInfo-item name">nick    name</dt>
<dd class="basicInfo-item value">
Ai Ye,Ai Hao,home Ai
</dd>
<dt class="basicInfo-item name">Sector</dt>
<dd class="basicInfo-item value">
Plant community
</dd>
<dt class="basicInfo-item name">door</dt>
<dd class="basicInfo-item value">
Angiosperm door
```

FIGURE VII. INFORMATION BOX CORRESPONDING THE PARTICIAL HTML SOURCE CODE

To solve the problem that mentioned above, first, XPath is used to locate the HTML tag nodes that need to extract the content, and then the content is extracted directly, or the regular expression is used to extract the content accurately.

D. Obtaining Attributes and Attribute Values of TCM Based on XPath

1) The concept of XPath: XPath locates the specified node from the document node tree, thus obtaining nodes, atomic values, node sets, and so on in the form of an expression language.

2) Path expression for XPath: There are two kinds of the path expression for XPath: stepping expression and relative path expression "/". There are three kinds of the stepping expressions: axis stepping expression, node test stepping expression and predicate stepping expression. For example, Figure 8 shows the HTML source code about the effect of Leaves:

```
</div><div class="para-title level-2" label-module="para-title">
<h2 class="title-text"><span class="title-prefix">Leaves</span>effect</h2>
</div>
<div class="para" label-module="para">Warm by bleeding, cold and cold pain; topical dampness
    itching.</div>
```

FIGURE VIII. THE HTML SOURCE CODE ABOUT THE EFFECT OF LEAVES

The relative path expression of the effect of Leaves is like this:

//h2[@class='title-text'][contains(text(),'effect')]

"//h2" means to the entire HTML document for all of the <h2> tags. "[@class='title-text']" means that this "h2" tag contains the "class" attribute, and its value is "title-text". [contains(text(),'effect')] means that this "h2" tag contains the "effect" text, in this way, the "h2" tag contains the "effect" text is located.

In order to extract attribute values, the context of the page also need to be taken into consideration. The axis stepping expression of XPath selects the tree structure relationship between the current node and context node. As it is shown in Figure 8, after locating tag "<h2>", the expression of XPath should be formulated to get the attribute value below that corresponding to "effect".

./following-sibling::div[1]

In the expression above, "." means the current tag node "<h2>", "following-sibling" means all of the sibling nodes of the current node below. "div[1]"means the first "div" node. In this way, the text containing specific effects like "Warm by bleeding, cold and cold pain; topical dampness itching." is located. In other words, it's the attribute values of "effect". The results can be shown in Figure 9:

alias	jie	men	gang	mu	ke	shu	zhong	poisonous
艾叶、艾蒿、家艾	植物界	被子植物门	双子叶植物纲	柏楸目	菊科	艾属	艾	有小毒

FIGURE IX. THE PARTIAL ATTRIBUTES AND ATTRIBUTE VALUES OF LEAVES

E. Obtaining Attributes and Attribute Values of TCM Based on Pattern Matching

Considering that not all of the HTML pages about TCM satisfy the XPath path expressions above. For example, the "nature and flavour" of some TCM may appear only in unstructured text.

The "properties and taste" is shown in table I and table II. Meanwhile, there are some modifiers that can modify the "nature and flavour", such as "slightly" or "very".

TABLE I. TCM NATURE

Warm	Hot	Cold	Cool	Mild

TABLE II. TCM FLAVOUR

Sweet	Sour	Salty	Bland	Acrid	Pucker	Bitter

The word " nature and flavour" is used as an indicator to solve this problem. This paper sets a length range, and thinks that the corresponding attribute values definitely will show up following the words "nature and flavour" within a certain range.

The entire process of information extraction can be shown in Figure 10:

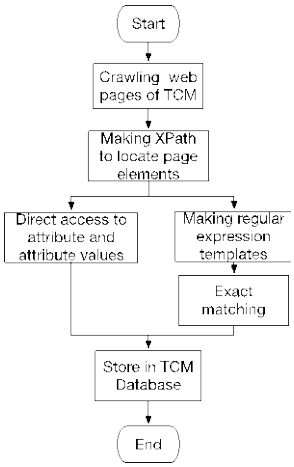


FIGURE X. EXTRACTION FLOW CHART OF ATTRIBUTE AND ATTRIBUTE VALUE OF TCM.

1) Formulate the regular expression [9] *to obtain the TCM property*

"nature\\s*flavour[\\s\\S]{0,10}?([slightly very][warm hot cold cool mild])"

The regular expression above means that the word "nature and flavour" itself or within 10 characters after the word "nature and flavour" matches several qualifying words describing the TCM property.

2) Formulate the regular expression to obtain the TCM taste

"nature\\s*flavour[\\s\\S]{0,10}?(slightly?sweet)"

In this way, this paper can avoid the fact that it's impossible to use a uniform XPath to locate the desired location in pages that with different structure. Meanwhile, it makes matching and extracting the information scattered in the text of the web page much easier than before.

F. The Result of Experiment and Evaluation

The Recall and Precision [10] are used to evaluate information extraction results. Assume that X represents all extraction results, and Y represents all possible extraction results, and Z is the intersection of X and Y which means the result of correct extraction. It can be shown in Figure 11:

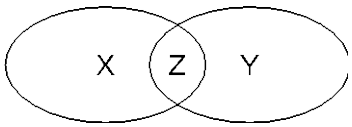


FIGURE XI. LOGICAL RELATION OF EXTRACTION.

So that the Recall $R = |Z| / |Y|$, the Precision $P = |Z| / |X|$. Recall describes the degree of understanding of the information extraction, while Precision describes the accuracy of information extraction.

The evaluations include the F value [11] that combines recall R and precision F. The formula is:

$$F = 2/(1/P + 1/R) \quad (1)$$

The formula shows that F only increases with the increase of both of the recall R and precision F. When R and F decrease or the difference is too large between them, F will decrease. Therefore, F can reflect the comprehensive effect of R and F.

In this experiment, F is used to evaluate the performance of crawler program. This paper takes 1000 TCM that randomly selecting from 11115 kinds of them as samples to test the crawler program by comparing the attribute value supplement of each TCM in 27 attribute values we predefined. The results shows that the F value of each drug was maintained above 0.8.

It can be concluded that the information in the field of TCM that based on open data source extraction can basically meet the general needs and ensure the accuracy and timeliness of information. Part of the result can be shown in Figure 12 and13:

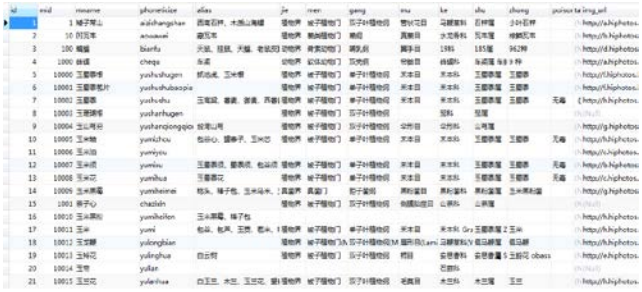


FIGURE XII. EXTRACTION RESULTS OF ATTRIBUTES AND ATTRIBUTE VALUES OF TCM (1)

FIGURE XIII. EXTRACTION RESULTS OF ATTRIBUTES AND ATTRIBUTE VALUES OF TCM (2)

III. SUMMARY AND PROSPECT

Although this paper solves the problem of extracting the attributes and attribute values of traditional TCM from the open data source, the crawler is still based on automatically inputting the existing traditional TCM names. With the development of NLP, it hopes that the crawler can extract more comprehensive information about traditional TCM through self-learning with only a few seed words.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61175048) and we would like to express our gratitude to Bing Sun for he helped us conduct experiments.

REFERENCES

[1] Baldwin T, Korhonen A, Villavicencio A. Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition[J]. Acl Workshop on Deep Lexical Acquisition, 2005:67--76.

[2] Yoshinaga N, Torisawa K. Open-Domain Attribute-Value Acquisition from Semi-Structured Texts[C]// The Workshop on Ontolex -The Lexicon/ontology Interface Held at the Fifth International Semantic Web Conference. 2007.

[3] XianYi Cheng, Qian Zhu. Research on semi supervised learning framework for relation extraction of undefined type[J]. Journal of Nanjing University (NATURAL SCIENCE), 2012, 48(4).

[4] JianYi Guo, Zhen Li,ZhenTao Yu, et al. Domain ontology concepts, instances, attributes and attribute values extraction and relationship prediction[J]. Journal of Nanjing University (NATURAL SCIENCE) 2012, 48(4):383-389.

[5] Wang Z C, Wang Z G, Juan-Zi L I, et al. Knowledge extraction from Chinese wiki encyclopedias[J]. Journal of Zhejiang University Science C, 2012, 13(4):268-280.

[6] Mc Kinsey Global Institute. Open data: unlocking innovation and performance with liquid information. http:// www. Mckinsey. Com /insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information, 2013.

[7] SiAcrid Tang. Design and manufacture of web page based on Web standard[M]. Tsinghua University Press, 2009.

[8] JinHong Liu, YuLiang Lu. Research review of topic crawler[J]. Computer Application Research, 2007, 24(10):26-29.

[9] GongMing Wang, HuaRui Wu, ChunJiang Zhao, et al. Application of regular expression in verification of e-government client[J]. Computer Engineering, 2007, 33(9): 269-271

[10] J. Makhoul, F. Kubala, R.Schwartz, R.Weischedel.Performance measures for information extraction[J].DARPA Broadcast News Workshop,1999.

[11] WeiHui Zeng, Miao Li. A survey of deep web crawler research[J]. Computer System Applications,2008.