Domain Ontology Concept Extraction Method Based on Text

Yuefeng Liu
School of Computer Science
Communication University of China
Beijing, China
yuefeng_liu@foxmail.com

Minyong Shi, Chunfang Li
School of Computer Science
Communication University of China
Beijing, China

***Abstract*—This paper propose a new method to extract ontology concepts from multiple text of the same type. This method uses mutual information and document frequency. According to the mutual information tend to choose the low frequency words, combining mutual information and document frequency to avoid this problem. In this paper, a number of financial and economic reports are reported as the corpus. First of all, do the text preprocessing, and then based on the N-gram algorithm to generate a set of candidate phrases, and finally use the statistics and the rules to screen for the concept of ontology from candidate phrases.**

Keywords—ontology concept; mutual information; document frequency; N-gram

I. INTRODUCTION

As a conceptual model of knowledge, ontology has become a hot research topic in the current research. With the continuous expansion of the scale of ontology construction, relying on the construction of a large-scale ontology is not realistic, using computer technology to replace manual operation, so as to shorten the construction cycle and cost, become the focus of research[1]. Ontology mainly consists of two parts, concepts and relations. How to extract the concept from the data source is the first step in the construction of the domain ontology. Before extracting the concept of ontology, it is needed to determine the source of data to obtain the concept. Text is an important source of data, the text contains all the documents that can reflect the knowledge of the field, including journals, books, newspapers, etc..

Domain ontology concept extraction, in essence, is to cut sentence in accordance with the given algorithm, and then extract it out. Although the concept of ontology extraction algorithm is a lot, but there are relatively fixed steps and can be roughly divided into three steps[1]: (1) text segmentation, (2) phrase extraction, and (3) phrase selection.

Phrase extraction or term extraction plays a very important role in many applications, there have been many scholars have studied. Reference [2] puts forward the concept of basic noun phrase in Chinese from the angle of linguistics. He thought The recognition know ledge includes the basic construction templates which specify the syntactic composition of baseNPs(static knowledge) and the context-sensitive transformative rules(dynamic knowledge)which reflect the context features . Based on the above knowledge , a transform action-based model for recognizing Chinese baseNP is put forward, which incorporates the static knowledge and the dynamic knowledge into an organic whole to recognize the baseNPs in Chinese texts. Reference [3] puts forwards an automatic method for extracting Chinese Multiword Expressions(MWEs) with help of statistics and linguistic rules . Seed words of high frequency in special domain are selected to ex tract candidate strings. By means of statistical measures and linguistic rules, noises in candidate strings are filtered .After filtering, Chinese MWEs are obtained finally. Reference [4] at first test the performance of nine widely adopted statistical measures of such kind in Chinese word extraction on the individual basis, then try the possibility of improving the performance by properly combining these measures. Genetic algorithm is explored to automatically adjust the weighting of combination. Result suggests that these measures could not supplement well each other, and the simplest and effective way in Chinese word extraction would be using mutual information directly.

The proposed methods can be applied to the extraction of ontology concepts. But the input of these methods is a set of text in a field, and there is no distinction between the texts and do not consider a number of the same type of text as input.

II. THE DIFFERENCE AND RELATION BETWEEN ONTOLOGY CONCEPTS AND TERMS OF SPECIFIC FIELD

The term is used in a subject area that represents a concept or relationship within the subject area and can be a word or phrase. The term can exist only in a subject area, but also in a number of subject areas. The Common Words refers to a word that is in the field of a subject, in addition to the term[5]. Domain ontology concepts defined in this article, not only include terms, but also Common Words.

III. ABOUT CORPUS

In this paper, we use five hundred corporate quarterly earnings collected from the network as corpus. Algorithm proposed in this paper can obtain the ontology concepts of earnings. This is of practical significance, such as certain types

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

of sports news, health report, weather forecast, can get the number of the text of the same type. The proposed approach is also suitable for other areas, but needs a sufficient number of text of the same type.

IV. TEXT PREPROCESSING

In Natural Language Processing, the Chinese word segmentation can be seen as a problem has been solved. Jieba word segmentation is an open source Chinese word segmentation on GitHub, with a higher accuracy of the word segmentation. The general steps of Jiaba are as follows[6]:

- Based on Trie tree to achieve efficient word map scanning.
- Generate sentence into directed acyclic graphs formed by all possible Chinese characters.
- Dynamic programming to find the maximum probability path, find the maximum combination of segmentation based on word frequency.
- For unknown words, use the HMM and Viterbi algorithm for segmentation.

Posseg module in the Jieba is responsible for word segmentation and part of speech tagging, and provides a Cut interface. Cut interface accepts input text, outputs a result sequence of word segmentation and part of speech tagging. In the process of phrase extraction, we should distinguish the sentence. Otherwise, the last word of the first sentence and the first word of the second sentence will be combined into a phrase into the candidate set, which of course is not allowed. Text preprocessing steps are as follows:

- A sequence of sentences with clauses.
- Cycle for each sentence with the posseg module for word segmentation and part of speech tagging.
- Each sentence returns a list of words and parts of speech.

Chinese punctuation usually use regular expressions to separate text in accordance with the division of the meaning of punctuation marks (such as a period). The regular expression used in this paper are as follows:

[（） 《》？()、：；，,\[\]。“ ”‘ ’ ！【】]+

The following structure after word segmentation are as show in Fig.1, where *word* is a real word, *flag* is part of speech.

[
[(*word*, *flag*),(*word*, *flag*),•••],
[(*word*, *flag*),(*word*, *flag*),•••],
•••
]

Fig. 1. Struct after posseg

The Table I is part of speech and its abbreviation:

TABLE I. PART OF SPEECH

short	part-of-speech
a	adjective
b	distinguished
c	conjunction
d	adverb
e	interjection
f	directions
g	morpheme
h	enclitic
i	idioms
j	short
k	behind the word
m	numeral
n	noun
o	mimetic
p	preposition
q	quantifiers
r	pronoun
u	auxiliary
v	verb
y	int

V. PHRASE EXTRACTION

In this paper, we use the N-gram algorithm to generate candidate phrase set. Steps are as follows:

a) TempQueue.put(InputQueur.pop()) .

b) Repeat step a) N times.

c) OutputQueue.put(TempQueue).

d) InputQueue.hasnext(), if True jump to step e), else jump to step f).

e) TempQueue.pop(); TempQueue.put(InputQueue.pop()); OutputQueue.put(TempQueue); jump to step d).

f) Output OutputQueue; Program termination.

Explain: Queue represents a queue; Queue.put () said add elements in the end of the queue; Queue.pop (), return and delete the first element of the queue; Queue.hasNext (), to determine whether there are elements in the queue. If we want two-word-phrases, N should be 2.

N-gram will not miss the field phrases, but will also extract a large number of obvious error phrases[1]. In this paper, 80686 two-word-phrases extracted by 2-gram algorithm based on the five hundred quarterly earnings. These two-word-phrases are the input for the next phrase filter step.

VI. PHRASE FILTER

A. Mutual Information

MI(mutual information) in information theory is a useful measure, it can measure the correlation between a set of two events. The MI of the two events, X and Y, is defined as (1)[7]:

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

$$I(X,Y)=H(X)+H(Y)-H(X,Y) \quad (1)$$

Among them, $H(X,Y)$ is the joint entropy, defined as (2):

$$H(X,Y)=-\sum p(x,y)\log\left(p(x,y)\right) \quad (2)$$

Two grams mutual information refers to the probability function of the two events, the definition is as (3):

$$MI(w_1,w_2)=\log\frac{p(w_1,w_2)}{p(w_1)\times p(w_2)} \quad (3)$$

In (3), $p(w_1)$ is the probability that w_1 is concentrated in the whole training text; $p(w_2)$ is the probability that w_2 is concentrated in the whole training text; $p(w_1,w_2)$ indicates the probability of the simultaneous emergence of w_1 and w_2. The higher MI indicates the larger possibility that the correlation between X and Y is to be an ontology concept; Conversely, the lower the MI, the smaller the correlation between X and Y [3].

Calculation of MI and TF(term frequency) and sort the results according to the mutual information in descending order. The Table II is part of the results:

TABLE II. MI IN DESC ORDER

Word1	Flag1	Word2	Flag2	MI	TF
凡	d	国君	n	18.92957	1
国电	j	千万元	m	18.92957	1
移动网	n	游乐	n	18.92957	1
沉重	a	哀悼	v	18.92957	1
因有	c	世界杯赛	nz	18.92957	1
有待	v	监管部门	n	18.92957	1
755	m	股麦	n	18.92957	1
攻击	v	而已	y	18.92957	1
核高	n	基	n	18.92957	1
未升	v	反跌	v	18.92957	1
应付	vn	票据	n	18.92957	1
藏族	nz	神话	n	18.92957	1
前沿	s	新闻报道	n	18.92957	1
抵减	v	项	q	18.92957	1
百变	nz	金刚	nr	18.92957	1
0.123	m	R	eng	18.92957	1
现四	t	足	a	18.92957	1
性价	n	产出	v	18.92957	1
新闻出版	n	总署	n	18.92957	1
美丽	ns	童行	nr	18.92957	1

As shown in the Table II:*Flag1* and *Word1* is the first word and part of speech of the two word phrase; *Flag2* and *word2* is the second word and its part of speech; *TF* is the number of times the two word phrase appears in all texts. The *TF* of these two-word-phrases is very small. Because *MI* does not take the frequency into account. This is a very big disadvantage of *MI*, it leads to MI evaluation function often tend to choose rare words[8]. These low frequency words do not conform to the requirements of extracting ontology concept. So we can not directly use the MI to filter out the concept of ontology.

B. IDF and DF

IDF(inverse document frequency) is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number by the number of documents containing the term, and then taking the logarithm of that quotient.

While the DF(documents frequency) is the inverse of IDF. It represents the universality of things. This is in line with our ontology concept.

The definition of DF are as (4)[9]:

$$DF(t,D)=\frac{\left|\{d\in D:t\in d\}\right|}{N} \quad (4)$$

D, the set of all documents; N, total number of documents in corpus N=|D|; $\left|\{d\in D:t\in d\}\right|$, number of documents where the term t appears.

This is the most simple evaluation function, it is characterized by a small amount of computation. But in the actual use of the general does not directly use DF[10]. Usually, we taking the logarithm of that quotient as (5):

$$\log\left(DF(t,D)\right)=\log\frac{\left|\{d\in D:t\in d\}\right|}{N} \quad (5)$$

Two grams mutual information reflects the correlation between the two words, while the logDF reflects the universality of the two word phrase in the document collection.

The logDF of 80686 two-word-phrases are sorted in descending order as Fig. 2, The horizontal coordinate is the number of words, and the vertical coordinate is logDF value:

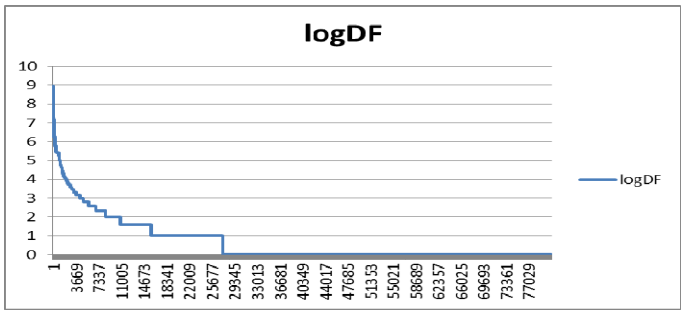

Fig. 2. LogDF of 80686 terms

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

The MI of 80686 two-word-phrases are sorted in descending order as Fig. 3, The horizontal coordinate is the number of words, and the vertical coordinate is MI value:

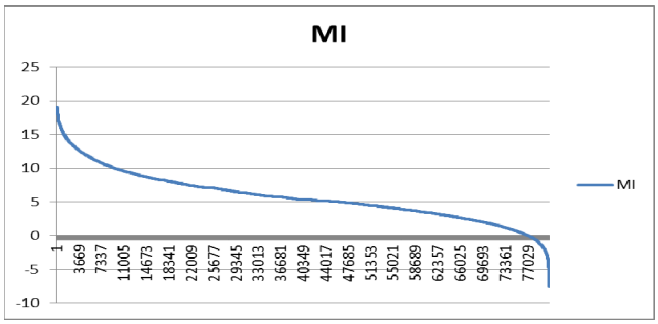

Fig. 3. MI of 80686 terms

We can set a threshold for each logDF and MI and fiter the 80686 two-word-phrases[11]. Through repeated tests, it was found that when the logDF was 6, the MI was 2, the result was the best. After filtering, the number of the rest two-word-phrases is 244.

C. *Filtering Based on Linguistic Rules*

Although the statistical method can scale for researchers for screening, but the statistical method is not a panacea. Meet the statistical value of words is not necessarily the expression of words with semantic completeness. Therefore, we consider the introduction of linguistic rules for screening.

To save storage space and improve searching efficiency, some words or words can be automatically filtered before or after the processing of natural language data (or text). These words are called Stop Words. Words Stop roughly divided into the following two categories[12]:

- These words are widely used in the Internet can be seen everywhere, such as the word "Web" on almost every site will appear. For such a word, search engine can not guarantee the real relevant search results, it is difficult to help narrow the scope of the search, but also reduce the efficiency of search.
- This class is more, including particles, adverbs, prepositions, conjunctions etc. and usually has no clear meaning, only to put it in a complete sentence in a certain role, such as the common “of”, “in” and so on.

In this paper, the Stop Words is the second kind. With the help of part of speech tagging, filter out the two-word-phrases that the head flag or the tail falg included in f, p, c, u, y .

Finally get 102 two-word-phrases. These can serve as our ultimate domain ontology concept. Table III lists some of the results.

TABLE III. FINAL ONTOLOGY CONCEPT

Word1	**Flag1**	**Word2**	**Flag2**	**MI**	**logDF**
北京	ns	时间	n	10.24952	6.70044
本季度	n	运营	vn	4.635771	6.954196
本文	r	来源	n	9.348284	8.654636
部分	n	抵消	v	10.39489	6.066089
财报	n	显示	v	7.118805	7.219169
财报	n	数据	n	7.011172	6.97728
财报	n	全文	n	8.245734	6.129283
存托	v	凭证	n	9.818438	7.209453
第三季度	mq	财报	n	3.983175	6.169925
非	h	美国通用	nt	8.264496	6.507795
费用	n	支出	v	4.930167	6.72792
付费	v	用户	n	8.060147	6.285402
个	m	百分点	m	9.870066	6.807355
公布	v	截至	v	6.618443	6.285402
公司	n	预计	vn	5.371698	6.149747
广告	n	收入	v	5.276085	6.714246
环比	j	增长	v	4.814172	8.238405
环比	j	下降	v	5.000947	7.149747
环比	j	减少	v	5.095033	6.61471
环比	j	下滑	v	4.941726	6.022368
基本	n	持平	n	9.336981	6.584963
季节性	n	因素	n	10.91875	6.247928

VII. CONCLUSION

Phrase extraction or multi word extraction plays a very important role in many applications. Many scholars have studied it and proposed effective methods. These methods are generally based on a large number of texts that does not distinguish as input. But in some areas, such as weather reports, medical reports, sports events, we can collect a number of separate texts of the same type. This paper propose a method based on the use of statistical and linguistic rules and carried out experiment with taking quarterly earnings as an example. Extract concept of the domain ontology composed of two-word-phrase. The next step is to extend this approach to the extraction of the ontology concepts composed of three-word-phrase and four-word-phrase.

ACKNOWLEDGMENT

This paper is partly supported by “the Excellent Young Teachers Training Project (the second level, Project number: YXJS201508)”, “Key Cultivation Engineering Project of Communication University of China (Project number: 3132016XNG1606 and 3132016XNG1608)”, “Cultural technological innovation project of Ministry of Culture of P.R.China (Project number: 2014-12)”, and partly supported by “The comprehensive reform project of computer science and technology, department of science and Engineering”. The research work was also supported by “Chaoyang District Science and Technology Project (CYXC1504)”.

REFERENCES

[1] F. Yu, Domain ontology construction method and empirical research, pp. 20-21, 2015

[2] J. Zhao, “A Transformation-based model for chinese BaseNP recognition,” JOURNAL OF CHINESE INFORMATION PROCESSING, vol. 13, pp. 1-7, Mar 1998.

[3] R. Liu, “Extracting Multiword Expressions with statistics and linguistic rules,” JOUNAL OF TAI YUAN UNVERSITY OF TECHNOLOGY, vol. 42, pp. 133-137, Mar 2011.

[4] S. F. Luo, “Chinese word extraction based on the internal associative strength of character strings,” JOURNAL OF CHINESE INFORMATION PROCESSING, vol. 17, pp. 9-14, Jan 2003.

[5] F. Xie, Automatic acquisition of domain terms, Central China Normal University, Jun 2006.

[6] W. Tao, “Implementation of Chinese word segmentation algorithm based on bidirectional maximum matching in Police Affairs,” the Application of Computer Technology, pp. 153-155, Mar 2016.

[7] W. Q. Cheng, “A text feature selection method using the improved Mutual Information and Information Entropy,” Journal of Nanjing University of Posts and Telecommunications(Natural Science), vol. 33, pp. 63-68, Oct 2013.

[8] X. Y. Chen, “Document feature slection based on the milimum Term Frequeney Threshold,” PR&AI, vol. 19, pp. 531-536, Aug 2006.

[9] S. A. Tan, “Improved TF-IDF Method in Text Classification,” New Technology of Library and Information Service, vol. 238, pp. 27-30, Jun 2013.

[10] H. Ning, “A Study on Feature Selection in Chinese Text Categorization,” Development research and Design Technology, pp. 774-776, Aug 2007.

[11] X. L. JIA, “A Survey of Ontology Learning from Text,” computer science, vol. 34, pp. 181-185, 2007.

[12] B. L. Hua, “Stop-word processing technique in knowledge extration,” New Technology of Library and Information Service, vol. 8, pp. 48-51, May 2007.