Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (18): 3620-3625, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3620.3625

Improvement of Method of Concept Extraction Based on Bootstrapping in Ontology Learning

[1]Wu Junyun and [2]Xie Caiyun
[1]School of Information and Engineering, Nanchang University, Nanchang, China
[2]Department of Information Science, Nanchang Teachers College, Nanchang, China

Abstract: This study introduces method of concept extraction based on Bootstrapping in Ontology learning. On the basis of the method for concept extraction based on Bootstrapping, it increases the compound words extraction and improves the statistics method of frequency, which can be more scientific to extract the domain concept. At the meanwhile, aiming at the weakness which the concept extraction method based on Bootstrapping uses statistical methods and ignores the semantic affect to the result, the accuracy of concept extraction is improved.

Key words: Ontology learning, bootstrapping, concept extraction

INTRODUCTION

Bootstrapping is an unsupervised machine learning method, which is widely used in knowledge acquisition. It starts from the collection of a small amount of seed concepts, automatically learns new concepts through continuously learning to acquire knowledge (Chen *et al*., 2003). The method is independent of the specific areas and has the property of portability.

The model of automatically getting domain concept based on Bootstrapping is a departure from the concept of artificial selection of seed and this model acquires new vocabulary automatically through learning in the large unlabeled corpus. Learning the field of the concept is an iterative process: Firstly, using the seed concept as an important concept to learn unlabeled corpus to generate a set of candidate concepts; secondly, evaluating candidate concepts, then adding the candidate concepts consistent with the evaluation criteria to the collection of domain concepts and important concepts; if the collection of important concepts need to update, there will be a new round of learning process.

PROCESS OF BOOTSREAPPING

Bootstrapping-based domain concepts extraction automatically is a departure from the concept of artificial selection seeds and then to learn from a lot of files, the model structure is shown in Fig. 1. And we can see from the figure, the domain concepts of the learning process is an iterative process. And the algorithm is described in the next chapter.

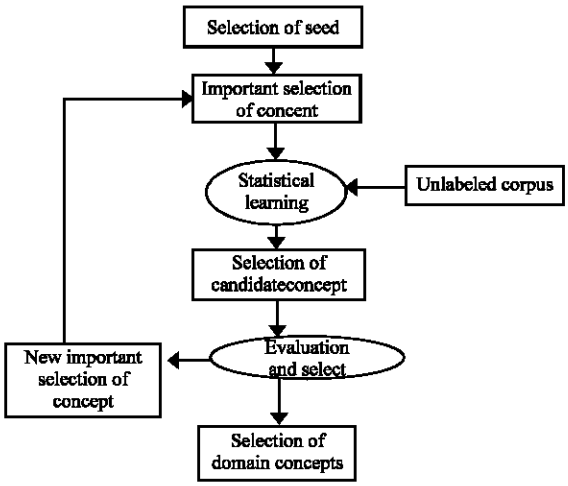

Fig. 1: Model structure of bootstrapping

To determine whether an ordinary concept can constitute the domain concept by using statistical methods, there are common statistical parameters including mutual information, Dice coefficient, correlation, co-occurrence frequency (Liang and Wu, 2006).

PROCESS OF CONCEPT EXTRACTION

Extracting domain concepts based on the Bootstrapping method is based on this assumption: If the frequency which certain words and terms in a sentence appeares together reached a certain value, then this word may be a domain concept in this field, so the statistical parameters selected is co-occurrence frequency. Abstract definitions are as follows.

Corresponding Author: Wu Junyun, School of Information and Engineering, Nanchang University, Nanchang, China

Definition 1: Frequency of words: The number of sentences containing the word 'w' in corpus.

Definition 2: Frequency of sets: The number of sentences containing any element in the set 'X'.

Definition 3: Frequency of co-occurrence F (w, X): The number of sentences containing the word 'w' and any element in the set 'X'

Satisfying condition of adding candidate concepts from the ordinary concept extracted in the corpus set is shown as Eq. 1:

$$\begin{cases} F(w) \geq F_{min} \\ F(w,IW)/F(w) \geq R_{min} \\ w \notin wordstoplist \end{cases} \quad (1)$$

Among them, F_{min} represents the minimum frequency as domain concepts must appear in the corpus, R = F (w, IW)/F(w) represents a support of concepts in the corpus, R_{min} is the minimum support degree concept must be achieved, wordstoplist represents the list of stop words.

We can get the evaluation of each candidate concepts of value by evaluating the set RW of candidate concepts generated set. According to the evaluation results,we choose the concepts which are in line with the concept of evaluation standard as the domain concept, the add to the domain concept in the collection DW and then select the new important concepts to adding to the important concept in the collection. The evaluation formula is as follows:

$$m_w = Log_2 F(w,IW) \times \frac{F(w,IW)}{F(w)} \quad (2)$$

The larger The value of m_w is, the greater the possibility that W is domain concepts is.

ANALYSIS AND IMPROVEMENT OF BOOTSTRAPPING

The Bootstrapping method is machine learning method for automatically learning new concept in the field in the premise of inputting seed concept (Liao and Grishman, 2010; MacLean *et al*., 2013). Using the statistical concept of seed and word co-occurrence frequency to determine the domain concept, ignoring the semantic information it contains. It is actually a kind of methods based on statistics. Because Ontology is realized at the semantic level of domain knowledge description, therefore, using the Bootstrapping method of concept extraction in ontology learning has the inevitable defects. This study gives the corresponding improvement for the Bootstrapping method, on the basis of statistical method of adding the semantic factors.

The improvement of Bootstrapping method in this study is mainly from the following three aspects: (1) Adding compound word extraction on the basis of the original module, so that it can be more scientific when extracting domain concept, (2) Changing the frequency statistics method to extract domain concepts with lower frequency; (3) Improving existing evaluation methods to adapt to the improved Bootstrapping method.

Extraction of compound word: Domain concept is usually the expression characteristics of a field of words, not only has a single word, also contains some compound words which consists of a plurality of words by words. Therefore, before the extraction of domain concepts, we extract the compound words in the text in the field.

Compound words can not be obtained by word segmentation system automatically. Because compound word usually consists of more than one word or word combination, they can be obtained by computing the word and the word compactness. There are many kinds parameters when computering the compactness of word and word. After studying the combination of several commonly used statistical parameters and different in automatic keyword extraction, Luo Shengfen and Sun Maosong (Luo and Sun, 2003) give the conclude: there is not good complementarity among several statistical parameters commonly used, the simple and effective method of keyword extraction is to use mutual information for word extraction directly, the accuracy rate is higher. There, the mutual information is calculated between the word and the word compactness, to extract a compound concept. The calculation formula of mutual information is as follows:

$$MI(w_1,w_2) = \log_2 \frac{P(w_1,w_2)}{P(w_1)P(w_2)} = \log_2 \frac{\frac{F(w_1,w_2)}{F}}{\frac{F(w_1)}{F}\frac{F(w_2)}{F}} = \log_2 \frac{F(w_1,w_2)F}{F(w_1)F(w_2)} \quad (3)$$

In this equation, P (w_1, w_2) is the probability of words w_1 and w_2 occurrs in the corpus at the same time. P (w_1) is the probability of the word w_1 appears. P (w_2) is probability of word w_2 appears. F (w_1, w_2) is the frequency of words w_1 and w_2 occurrs in the corpus at the same time. F (w_1) is frequency of the word w_1 appears. F (w_2) is frequency of word w_2 appears. F is the total number of sentences.

In this study, the steps of extracting compound words are as follows:

- Make segmentation of texts in the corpus and determine the choice of words window. The window size refers to the given word as the center, the longest distance to the left end and to the right end respectively. The shorter the distance between two words is, the more relevant these two words may be
- To generat the candidate words of compound word. Considering the length of compound word is less than 5 in general, this study generates the candidate words of compound word according to two yuan grammar, three yuan grammar and four yuan grammar, respectively. Take "Sino-foreign\cooperate\operate\ enterprises" as an example. According to two yuan grammar,we can get the candidate words "Sino-foreign cooperating, cooperative operating, operative enterprises"; According to three yuan grammar, we can get the candidate words "Sino-foreign cooperate, cooperative enterprises"; According to four yuan grammar,we can get the candidate words "Sino-foreign cooperate enterprises"
- To calculate the mutual information of candidate words, select the values of words which are greater than the threshold of candidate word as compound words. The mutual information of multiple words is calculated by the extended formula of mutual information of two yuan terms. The method of extending is: Mutual information is first calculated by two yuan terms, the value of this mutual information will be taken as the value of mutual information in three yuan term if it is more than the threshold of mutual information, mutual information values of four yuan terms will be obtained by the same method. For example: If MI ("Sino-foreign cooperation")> threshold, then:

$$MI(\text{"Sino-foreign Cooperative operating"})=\log_2\frac{P(\text{"Sino-foreign Cooperation","operating"})}{P(\text{"Sino-foreign Cooperation"})P(\text{"operating"})} \quad (4)$$

Frequency statistics of weighted word: The Bootstrapping method sets out concepts statsfying certain conditions as basical concepts through calculating the frequency and the co-occurrence frequency of words (Kaneishi and Dohi, 2011; Singh and Sedory, 2011), such as Eq. 1 shows. It can extract words with certain frequency in the field, but cannot find the words of lower frequency in this field. This study obtains that in two cases the field words with low frequency will be generated by observing the corpus: One is the synonym phenomenon in Chinese. If the text uses different words to express the same meaning, it will cause that some words have lower frequency (Maedche and Staab, 2001a). It uses the semantic similarity computation to recall the domain concept with low frequency and synonyms. The details is in the next section; The other one is getting the compound words according to the mutual information, these compound words may include field words with low frequency, so compound words will also have lower frequency. This study obtains composite concept includes low-frequency words by impoving the method of calculating the frequency of words.

The frequency of presence of compound words in the text in the field is usually low (Maedche and Staab, 2001b), but if it contains the word domain concept, this compound word is also very likely domain concept, this study will change the frequency statistics method for compound words. The calculation formula of frequency of words is as follows:

$$F(w)=\sum_{i=1}^{n}F(w_i) \quad (5)$$

In Eq. 5, n is the number of words contained in w. w_i is the word contained in w in the location of i. Therefore, the conditions of candidate concept extracted from the corpus are shown in Eq. 6:

$$\begin{cases} F(w)\geq F_{min} \\ R=\dfrac{\sum_{i=1}^{n}F(w_i,IW)}{F(w)}\geq R_{min} \\ w\notin \text{wordstoplist} \end{cases} \quad (6)$$

Method of evaluation: Using the evaluation formula in the Bootstrapping method to select some important domain concepts to update the key concept set. Because it improves the statistical parameters on the foundation of the Bootstrapping method and introduces a semantic factors, so its evaluation methods also should make corresponding improvement. The evaluation formula is defined in this study as follows: Makeing the sum of mutual information and semantic similarity, as is shown below:

$$E(w)=\alpha\bullet m_w+\beta\bullet\max(\text{sim}(w,IW)) \quad (7)$$

In Eq. 7, the calculating method of m_w can be seen in Eq. 2. Max (sim (w, IW)) represents the maximum similarity for word w and important concept set concept IW. There is no common theory to select general values of

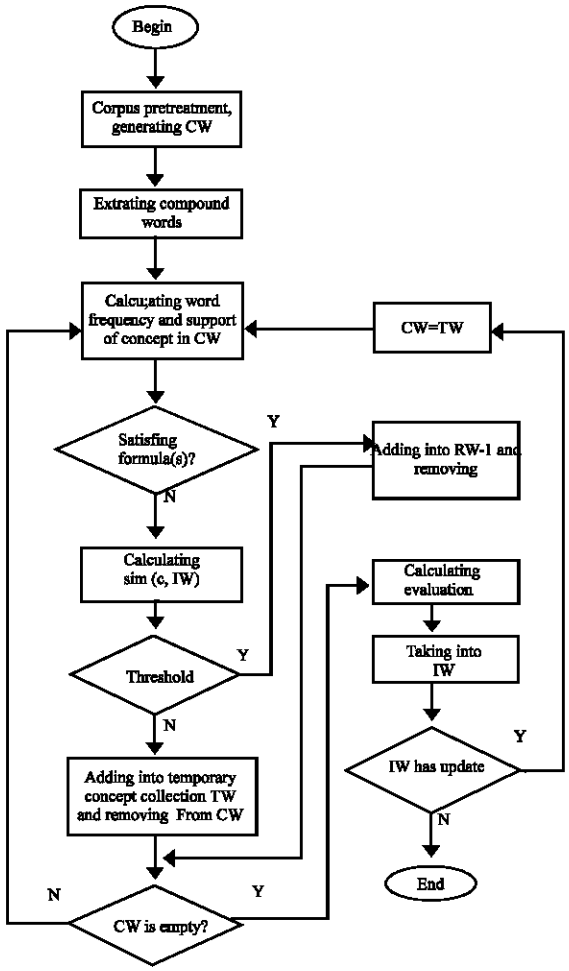

Fig. 2: Flow chart of concept extraction

coefficient α and β. According to the test results and appropriate regulation in the reference (Yamaguchi, 1999), it was found that when $\alpha = 0.55$, $\beta = 0.45$, we can get good experimental effect relatively.

Algorithm of concept extraction: In this study, we use the impoved algorthm of Bootstrapping, the dicription of this algorthm is as follows:

Input: Set of seed concepts, unlabeled corpus
Output: The domain concepts in the corpus
Process:

- Make segmentation on Corpus sentence, remove stop words and generate a common set of concepts
- Extracted compound words, put the words in the collection
- A collection of important concepts IW = SW
- If CW is not empty, go to Step 5; otherwise go to Step 7
- From the collection CW to select a word c to calculate the frequency, support, if they satisfies (6), add into the candidate set of concepts and remove them from CW, then go to step 4
- To calculate semantic similarity sim (c, IW), if it is greater than the threshold value, add ino the candidate set of concepts and remov them from CW, go to step 4, otherwise add them into the provisional set of concepts
- Add the concepts in RW-1 into DW, calculate evaluation value according to the Eq. 7 at the same time, take the highest evaluation value into IW
- If IW updated, add temporary concept TW into concept CW, Empty TW and go to step 4, otherwise end the algorithm

After analysing and improving the Bootstrapping method, we can get the flow chart of concept extraction in this study. It is shown in Fig. 2.

ANALYSIS AND IMPROVEMENT OF BOOTSTRAPPING

Experimental data: According to the concept extraction method described above, it takes the Sogou corpus of text set as experimental data sources. This study selects 100 texts respectively in four fields: Legal, economic, traffic and computer in experiment and the extraction results are analyzed. The values of learning parameters in the experiment are given as follows: $F_{min} = 10$, $R_m = 0.5$. The number of each cirle added to the important concept in the set is N = 10. Each domain concept of seed is selected by hands. As is shown in Table I.

Results and analysis: The accuracy of experiment (precision), the recall rate (recall) and F-value (F-measure) are three kinds of evaluation index (Villaverde *et al.*, 2009; Nakaya *et al.*, 2004), they are defined as follows:

$$\text{Accurate rate} = \frac{\text{Concept}}{\text{correct extracted No.}}$$

$$\text{Recall rate} = \frac{\text{Correct concept of the concept of No.}}{\text{Actual}}$$

$$\text{F-value} = \frac{\text{Accurate rate} \times \text{Recall rate}}{\text{Accurate rate+Recall rate)}} \times 2$$

According the experimental data, when window size is 3 and the threshold is 3, the effect of concept extraction

Table 1: Seed concept in the field

Name of field	List of seed concept
Law	crime, illegal, trial, litigation, case, legal, law enforcement, parties, court, judicial
Ecolomics	Stock, finance, investment, banking, stock market, foreign exchange earnings, loans, securities, finance
Tranportation	Automobile, vehicle, driving, overloading, road, vehicle, the traffic police, transportation, civil aviation, traffic accident
Computer	Computer, software, hardware, database, chip, embedded, program, electronic commerce, video, notebook

Table 2: Experimental results

Name of field	No. of concepts rate extracted	The accurate (%)	The recall rate (%)	F-value (%)
Law	126	79.4	59.9	68.3
Ecolomics	131	77.1	63.5	69.6
Tranportation	107	72.9	58.2	64.7
Computer	119	80.7	61.5	69.8

Table 3: Comparison of experimental results

Name of field	No. of concepts rate extracted	The accurate (%)	The recall rate (%)	F-value (%)
Bootstrapping method before improving	113	71.7	49.7	58.7
Bootstrapping improving method after	126	79.4	59.9	68.3

is the best. Therefore, we choose these set of parameters to continue the experiment, the extraction results in various fields is shown in Table 2.

In order to verify whether the improvement of the method of Bootstrapping is available, this study compares the methods before improving and after improving, as is shown in Table 3. It uses the same data set and the selection of the legal field extraction results to compare with. The texts in the law field are taken as data set, other texts are taken as filtered documents.

From the above experimental results, we can summarize as follows:

- The effect of concept extraction is relevant to the parameters in the algorithm. There is no good method to determine the values of the parameters, in order to obtain the better effect of parameters ,this study sets the values artificially and gradually adjusts in the experiment. This method has some shortcomings
- The method based on the domain relevance and consistency and the method based on Bootstrapping are related to the statistical method, the experimental results with the selected corpus and corpus size are closely related. Considering the factors of workload, the size of corpus used in this study is small. With the increase of corpus size, the accurate rate and recall rate will be both improved
- After the Bootstrapping method is improved, due to the addition of compound word extraction of this module, the number of concepts extracted is also increased. And the accurate rate and recall rate have a certain improvement relatively

CONCLUSION

This study introduces the method of concept extraction based on Bootstrapping and gives analysis of the shortcomings in Bootstrapping method. On the basis of original method, it proposes the improvement: adding the module to the extraction of compound words. This improvement provides a basis for extraction of multi-word terms of domain concepts. At the same time, it improves the way of statistics of word frequency: using the similarity calculation method based on the lexical context to make up the defect of statistics method in the semantic aspect. This is an innovation in this study.

ACKNOWLEDGMENT

The project is supported by the Foundation of Jiangxi Provincial Education Department (No. GJJ12048) and Jiangxi Provincial Department of Science Technology (No. 20122BBE500049).

REFERENCES

Chen, W.L., J.B. Zhu and T.S. Yao, 2003. Automatically getting domain vocabulary based on bootstrapping. Proceedings of the 7th National Conference on Computational Linguistics, (CL'03), China, pp: 67-72.

Kaneishi, T. and T. Dohi, 2011. Parametric bootsrapping for assessing software reliability measure. Proceedings of the IEEE 17th Pacific Rim International Symposium on Dependable Computing, December 12-14, 2011, Pasadena, CA., pp: 1-9.

Liang, J. and D. Wu, 2006. Seed conceptual approach and its appication in the body of the text-based learning. Library Inform. Ser., 9: 18-22.

Liao, S. and R. Grishman, 2010. Filtered ranking for bootstrapping in event extraction. Proceedings of the 23rd International Conference on Computational Linguistics, Volume 2, August 23-27, 2010, Beijing, China, pp: 680-688.

Luo, S.F. and M.S. Sun, 2003. The study of Chinese word extraction based on the combined string internal tightness. J. Chin. Inform. Process., 3: 9-14.

MacLean, M.G., M.J. Campbell, D.S. Maynard, M.J. Ducey and R.G. Congalton, 2013. Requirements for labelling forest polygons in an object-based image analsis classification. Int. J. Remote Sensing, 34: 2531-2547.

Maedche, A. and S. Staab, 2001a. Ontology learning for the semantic web. IEEE Intell. Syst., 16: 72-79.

Maedche, A. and S. Staab, 2001b. The ontology extraction and maintenance environment text-to-onto. Proceedings of the ICDM 2001 Workshop on the Integration of Data Mining and Knowledge Management, November 29-December 2, 2001, Silicon Valley, CA., pp: 1-12.

Nakaya, N., M. Kurematsu and T. Yamaguchi, 2004. A domain ontology development environment using a MRD and text corpus. Proceedings of the Joint Conference on Knowledge Based Software Engineering, August 2004, Protvino, Russia, pp: 908-916.

Singh, S. and S.A. Sedory, 2011. Sufficient bootstrapping. Comput. Stat. Data Anal., 55: 1629-1637.

Villaverde, J., A. Persson, D. Godoy and A. Amandi, 2009. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Expert Syst. Appl., 36: 10288-10294.

Yamaguchi, T., 1999. Constructing domain ontologies based on concept drift analysis. Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5), August 2, 1999, Sweden, pp: 131-137.