# Use web resources to construct ontology concept hierarchy

HeYu[1], Xueqiang Lv[1], Liping Xu[2]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University,Beijing, 100101, China

[2]Beijing Research Center of Urban System Engineering, Beijing, 100089, China

**Keywords:** Ontology construction; terminology extraction; reverse speech rules; reference corpus; domain relevance degree

**Abstract.** Domain terminology with high quality are the fundamental component in ontology construction and domain terminology automatic extraction is the basis of domain ontology construction. Proposed an approach of achieving domain concepts hierarchies from the Web data. Used the clue words to product queries containing hierarchical relation to get corpus rich in concepts hierarchical relation through the search engine from Web. Then we got target concepts explanation from the online encyclopedia of knowledge corpus such as Wikipedia. By combining the previous two corpus with the domain news documents and HowNet, we constructed the concepts graph model. After applying pruning algorithm on the graph, we proposed a modified hierarchical concepts tree building method. Experimental data on auto car proved the efficiency of the proposed method.

## 1. Introduction

Domain ontology is a shared conceptual model specific areas clear, standardized description of to the relationship between the term set and terms reflect the field of knowledge systems can be used to achieve semantic information between human-computer interaction and machines [1].Currently, the domain ontology information has been widely used in many application areas, such as information retrieval, information extraction, and question and answer system. However, how fast and efficient access to areas of the body remains a pressing problem. Buitelaar[2] propose a hierarchical ontology construction model, the main building is divided into terminology recognition , synonyms digestion, concept acquisition, get the concept of hierarchical relationships, relationships and axioms six levels. Concept hierarchy relationships can be broadly divided into three categories: classes and instances of "IS-A" relationship, the whole - part relationship, ATTRIBUTE-OF and other implications associated with hierarchical relationships. This article does not detail the relationship between the classification to distinguish, only to find the whole concept of a hierarchical relationship between the angle of the right, and the establishment of tree concept.

Conceptual level, there are two aspects of the acquisition: the first is the concept of hierarchy rich corpus of the acquisition, which directly affects the quality of the conceptual level of acquisition; the second is to get the concept of hierarchy from the corpus.

Rich resources of the Internet is increasingly used to obtain the concept of hierarchical relationships. Wu Jiedeng [3] describes the six kinds of different forms an integral part of the relationship between knowledge, thus obtaining such knowledge of grammar patterns, establish grammar pattern library, then get sentences based grammar model and derive the concept of relationship. Xin et al [4] proposed a method for constructing a query based on the intention to obtain from the Web corpus contains part of the overall relationship. In the context of the query by adding the word describes an integral part of the relationship, to obtain an integral part of the relationship between the corpuses from the Web. Xuli Heng et al [5] to obtain network knowledge base entry as a relationship corpus.

This paper analyzes the characteristics of the existing conceptual level and inadequate methods of acquisition, the concept of a level of resources and utilization of web graph model acquisition methods. With the clue word corpus get rich hierarchy from the web, to establish the concept of text

 

vector model, using the cosine distance between the concept and the concept of vector-based text HowNet building concept semantic similarity graph model, and finally through the map model pruning operation to get the concept of hierarchical tree structure[6].

## 2. Corpus Acquisition and Modeling

The quality of the relationship between corpuses directly determines the relationship between the concepts of obtaining results from the two aspects of the quality of the relationship between the evaluation corpuses[7]: the relationship between the degree of enrichment of the concept and the concept of the relationship between the degrees of difficulty to obtain from the corpus. Based on the traditional hierarchy corpus acquisition method is restricted due to the content of the concept of the relationship between the statement corpus is not high.

### 2.1 Corpus acquisition method based on hierarchical relationships clue word

Before describing the clue word acquisition method, the first formal definition is given some symbols:

We define that $C_i$ stands for conceptual *i*. The $C=\{c_i \mid 1 \le i \le m\}$ shows the concept of a collection. The $r_i$ expressed some hierarchical relations. We define that the $R=\{r_i \mid 1 \le i \le n\}$ is the set which consists of a hierarchical relationship in which $n$ is the number of hierarchical relationships. We define $P=\{(c_i,c_j) \mid (c_i,c_j) \in R,\ c_i \in C, c_j \in C\}$ which is consisted by $c_i$ and $c_j$ . $c_j$ is hierarchical relationship constituted. We define $cw$ as the clue word. The $w_i$ stands for the word. $W=\{w_i \mid 1 \le i \le q\}$ is the vocabulary set. $T_1$ is the number of k text. $context_1$ stands for the relation grammar atmospheres which come from $T_1$. $T_2$ stands for the number of search engine return results. $context_2$ is gotten from $T_2$ contextual . $k$ is the parameters.

The procedure to get the clue word:

Submit the $(c_i,c_j)$ to the search engine to get top k results and the set of abstract text $T_1$ and the recommend text set $T_2$;

Search the $(c_i,c_j)$, $T_1$ and $T_2$ to get the concept of $c_i$ and $c_j$ in same sentence to store the them in $context_1$ and $context_2$;

Individually marked the contextual to get the segmentation and POS tagging which is in $context_1$ and $context_2$. Then store the nouns, verbs, adverbs and conjunctions into $W_1$ and $W_2$. $W=W_1 \cup W_2$ . $W$ is the frequency of each word in the right conceptual;

Calculate the clue word relevance level ($relevance(w_i,r_j)$) between $w_i$ and $r_i$ in $W$.Because of $r_i$ can create many query String and $w_i$ can appear many times in the documents returned by search engines. Then it has more associate degree with $r_i$. So the candidate word $w_i$ and hierarchical $r_i$ **:**

$$relevance(w_i,r_j)=(1+f_2(w_i))\times f_1(w_i)+f_q(w_i)\Big/\underset{w_i \in W}{\text{Max}}\{f_q(w_i)\} \qquad (1)$$

Where, $f_1(w_i)$ is the frequency that $context_1$ contains $w_i$ , $f_2(w_i)$ is the frequency that $context_2$ contains $w_i$ , $f_q(w_i)$ is the number that query string including hierarchical relationships of $r_i$ contains $w_i$ and Max is the function to get maximum.

For hierarchical relationships $r_i$ corresponding candidate clue word set, select the top 10 that relevant degree $relevance(w_i,r_j)$ is maximum as clue word.

### 2.2 Conceptual vector space model based on VSM

In addition, this paper also gets to explain the concept of the target corpus entries from Baidu Encyclopedia and Chinese Wikipedia, for the concept set $C=\{c_i \mid 1 \le i \le m\}$ , Its corresponding Wikipedia corpus can represent as $D_2=\{d_i \mid 1 \le i \le m\}$ ,it also use Sohu news corpus that contains the target concept, it represents as $D_3=\{d_i \mid 1 \le i \le q\}$ .

In the vector space model[8], documentation set of document show as feature and document

matrix. Similarly, this paper construct conceptual vector space model $A$ using the concept of frequency and document matrix.

$$A=[a_{i\times k}]_{m\times p} \quad (2)$$

$$a_{i\times k}=f_k(c_i) \quad (3)$$

Where, $f_k(c_i)$ is the frequency that the concept $c_i$ appears in the document $d_k$, $m$ is the number of concept, $p$ is the number of documents in the documentation set.

For three different document corpus $D_1$, $D_2$ and $D_3$, this paper establish corresponding conceptual vector space model $A_1=[a_{i\times k}]_{m\times m}$, $A_2=[a_{i\times k}]_{m\times m}$ and $A_3=[a_{i\times k}]_{m\times q}$.

### 2.3 The establishment of the concept map

Using the cosine distance calculate similarity of the concept[9]. Concept $c_i$ and $c_j$ of document feature vector can represent as,

$vec(c_i)=(f_1(c_i),f_2(c_i),...,f_p(c_i))$ and $vec(c_j)=(f_1(c_j),f_2(c_j),...,f_p(c_j))$ ,its computational formula of similarity $sim(c_i,c_j)$ as follows:

$$sim(c_i,c_j)=\frac{\sum_{t=1}^{p} f_t(c_i)\times f_t(c_j)}{\sqrt{\sum_{t=1}^{p} f_t(c_i)^2\times\sum_{t=1}^{p} f_t(c_j)^2}} \quad (4)$$

For three different document sets $D_1$, $D_2$ and $D_3$ ,its Corresponding concept vector space model are $A_1=[a_{i\times k}]_{m\times m}$ , $A_2=[a_{i\times k}]_{m\times m}$ and $A_3=[a_{i\times k}]_{m\times q}$ ,use the cosine distance calculate conceptual similarity matrix: $Sim_1=\left[sim(c_i,c_j)\right]_{m\times m}$, $Sim_2=\left[sim(c_i,c_j)\right]_{m\times m}$ and $Sim_3=\left[sim(c_i,c_j)\right]_{m\times m}$.

This paper takes $Sim_4=\left[simH(c_i,c_j)\right]_{m\times m}$ to represent the concept and the concept of similarity matrix obtained by HowNet. Using polynomial addition way for fusion four similarity matrix, the final similarity matrix use $S=s(c_i,c_j)_{m\times m}$ to represent,

$$S=K_1\times Sim_1+K_2\times Sim_2+K_3\times Sim_3+K_4\times Sim_4 \quad (5)$$

## 3 Level domain concepts acquisition

The current study used many methods to obtain hierarchical clustering concepts, such methods need to manually specify the number of hierarchical clustering[10], and cannot get a clear idea of a hierarchical tree structure presents a conceptual level of the property acquisition method, the method cannot get the concept of centralized any clear concept hierarchy affiliation.

### 3.1 Acquisition method based on the concept of hierarchical graph

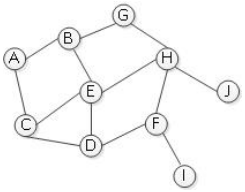


Figure 1 node topology diagram

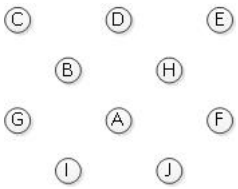


Figure 2 node hierarchy

The use of pruning algorithm uses the figure of Internet topology node classification. Starting from the outer node graph, according to the degree of the size of the map, from small to large, to the same degree nodes pruning operation step by step, and finally get the hierarchy diagram nodes.2, respectively, and the original structure of the node hierarchy obtained after reverse pruning FIG 1 and FIG 2.

### 3.2 Reverse pruning algorithm

Section 2 for the construction of conceptual graph model $G=(C,E)$. Collection $C=\{c_i\mid 1\le i\le m\}$ A conceptual diagram of the set of nodes, $E=\{(c_i,c_j)\mid s(c_i,c_j)\ge\alpha\}$ represents the edge set of the graph.

With $degree(c_i)$ a conceptual diagram concept $G$ the degree to $c_i$ Conceptual which level $rank(c_i)$, $H=\{(c_i,rank(c_i))|1\le i\le m\}$ as conceptual level in which the number of nodes to its collection of tuples, $G_1=(C_1,E_1)$ represents $G=(C,E)$ The same type of set. Reverse pruning algorithm as follows:

Input: $G=(C,E)$ $E=\{(c_i,c_j)|s(c_i,c_j)\ge\alpha\}$ $C=\{c_i|1\le i\le m\}$

Output: $H=\{(c_i,rank(c_i))|1\le i\le m\}$

**Step 1**: Initialization , $G_1$ assigned to $G$ ,so as to $C_1$ assigned to $C$ , $E_1$ is assigned to $E$. The initial value $t$ is 1.

**Step 2** Determine if $C_1$ is empty, go to step 3; otherwise continue;

**Step 3** Calculated the degree of $G_1$ to find the smallest degree of $md$ ; As the various $c_i$ ,if $degree(c_i)$ equals $md$ ,then assign $t$ to $rank(c_i)$, finally store $(c_i,rank(c_i))$ in $H$ , then get $c_i$ from $G_1$ and delete it from $C_1$. Delete $c_i$ from $G_1$ the edge set $E_1$. $t=t+1$ return 2;

### 3.3 Hierarchical tree structure algorithm

Hierarchical tree structure is to determine the level of explicit dependencies between concepts and concept based on the concept of node pruning operations which the level of access to information as well as the concept of node concept vector space model showing the distance.

Extension of the previously declared variables to $H=\{(c_i,rank(c_i))|1\le i\le m\}$ said the primary concept hierarchy by pruning algorithm to $S=s(c_i,c_j)_{m\times m}$ Conceptual similarity matrix by conceptual graph modeling method obtained two to $Hy=\{(c_i,c_j)|1\le i,j\le m,j\ne i\}$ Clear hierarchical subordination relation between a conceptual, i.e. $c_i$ Subordinate $c_j$ By 3.1 shows, $c_i\in C$ , $c_j\in C\bigcup\{Null\}$ ,when $c_j$ equal $Null$. When the show $c_i$ for the root node. Set up $C_1$ is a conceptual node set.

input: $H=\{(c_i,rank(c_i))|1\le i\le m\}$ , $S=s(c_i,c_j)_{m\times m}$

output: $Hy=\{(c_i,c_j)|1\le i,j\le m,j\ne i\}$

**Step 1**. Initialize $C_1$ and $Hy$ empty;

**Step 2**. If $H$ is not null, continue, otherwise go to 3;

**Step 3**. Get the smallest value of $rank$ from $H$ as below:

$C_{\min}=\{c_k|\forall j,rank(c_k)\le rank(c_j)\}$ , $C_{\min}$ concept $c_k$ ,get $rank$ value bigger than $rank(c_k)$ to choose the similarity $c_p$ ,as below:

$\forall j,s(c_k,c_j)\le s(c_k,c_p)\wedge rank(c_k)<rank(c_p)$ , make $(c_k,c_p)$ into $Hy$ , and put concept to the double tuple $(c_k,rank(c_k))$. Finally delete it from $H$ go to step2;

Output $Hy=\{(c_i,c_j)|1\le i,j\le m,j\ne i\}$ .

## 4 Experiment and Analysis

### 4.1 Experimental Design

This article assumes that due to the concept of centralized each concept there is a unique concept father, therefore, the whole concept of the number of hierarchical subordination set contains $relationNum$ Is the total number of concepts, i.e. $relationNum$ 109.With $relationNum$ Indicates the number of the correct level of affiliation identified, $validNum$ Indicates the number of valid relationship identified, representing the number of effective relationships include the correct affiliation and other non-subordination relation between the presence of the identified concept.In summary, this paper P and VP As the evaluation index, P represents the correct rate, VP is correctly represents the effective rate, calculated as follows:

$$P=\frac{rightNum}{relationNum}\times 100\% \qquad (6)$$

$$VP=\frac{validNum}{relationNum}\times 100\% \quad (7)$$

### 4.2 Results and Analysis

Choose this method under different parameter combinations of circumstances, to obtain an accurate level of affiliation rate than baseline; entry Use clues to explain the concept of the relationship between language materials and get the word encyclopedia knowledge acquired in building Graph Model proportion, the higher accuracy is obtained, and the influence of the former than the latter slightly larger weights;

Table 1 level affiliation obtain accurate rate

| No. | $K_1$ | $K_2$ | $K_3$ | $K_4$ | *rightNum* | *validNum* | *totalNum* | P | VP |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0 | 0 | 100 | 0 | 32 | 36 | 109 | 29.36 | 33.03 |
| 0 | 3 | 3 | 3 | 100 | 41 | 54 | 109 | 37.61 | 49.54 |
| 1 | 0.4 | 0.4 | 0.15 | 0.05 | 49 | 61 | 109 | 44.95 | 55.96 |
| 2 | 0.25 | 0.25 | 0.25 | 0.25 | 51 | 64 | 109 | 46.79 | 58.72 |
| 3 | 0.4 | 0.3 | 0.2 | 0.1 | 55 | 66 | 109 | 50.46 | 60.55 |

Table 1 experimental data prove that:

(1)Choose this method under different parameter combinations of circumstances, to obtain an accurate level of affiliation rate than baseline;

(2) Use clues to explain the concept of the relationship between language materials and get the word encyclopedia knowledge acquired in building Graph Model proportion.

(3)General news document set and based on the "CNKI" concept similarity computing the overall relationship recognition the relative contribution of the first two corpus low, intermediate analysis of the experimental data was found, based only on the coverage of the association concept conceptual graph model and news documents established between the concepts is not high, and the similarity between the concept and the actual situation exists than the much deviation, which is the relationship between the content of the document with the news is not high and there is a certain relationship between the concept of the concept of vector space based on similarity calculation. And because "CNKI" is a common ontology field of linguistics, so the use of "CNKI" conceptual areas there is a certain similarity calculation error;

## 5 Conclusion

This paper presents a use clue words get rich hierarchy corpus from the Web method. And extract concepts from Baidu Encyclopedia and Chinese Wikipedia entry explains. Integrated use of these two news corpus and related fields to establish the concept of a document corpus vector space model, and based on the introduction of the concept of "HowNet" similarity calculation method, constructed concept maps. Then use pruning algorithm, hierarchical tree generation algorithm concept hierarchy tree.


## Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No. 61271304 and Beijing Natural Science Foundation of Class B Key Project under Grants No. KZ201311232037

**References**

[1] Gruber TR. A translation approach to portable ontology specifications.Technical Report, KSL 92-71, Knowledge System Laboratory, 1993.

[2] BUITELAAR P, CIMIANO P, MAGNINIB Ontology learning from text: an overview [M] // BUITELAAR P, CIMIANO P, MAGNINIB Ontology learning from text:. Methods, evaluation and applications. Am sterdam: ISO Press, 2005.

[3] Wu Jie, Roberto, Cao stubs, and so on.Get in touch with the text portion of the page to verify the relationship between knowledge of [J]. East China University of Technology (Natural Science), 2006, 11: 013.

[4] Cao Xinyu, Cao stub.Method to get part of the overall relationship corpus from Web [J]. Chinese Information Technology, 2011, 25 (5): 17-23.

[5] Xuli Heng, Liu Yang, but to Sri Lanka, and so on.Ontology concepts based on multi-mount feature representation [J]. China Research Frontiers in Computational Linguistics (2009-2011), 2011.

[6] Binzabiah R, Wade S. Proposed method to build an ontology based on Folksonomy [C] // Information Society (i-Society), 2012 International Conference on.IEEE, 2012: 441-446.

[7] Ahmed KBS, Toumouh A, Malki M. Effective Ontology Learning: Concepts' Hierarchy Building using Plain Text Wikipedia [C] // ICWIT 2012:. 170-178.

[8] He T, Zhang X, Xinghuo Y. An Approach to Automatically Constructing Domain Ontology [C] // Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, Wuhan, China.2006: 150-157.

[9] warm spring, Shizhao Xiang, Yang Guozheng.A method of use of the property to get ontology concept hierarchy [J]. Mini-Micro Systems, 2010, 31 (2): 322-326.

[10] Ge J, Li Z, Li T. A Novel Chinese Domain Ontology Construction Method for Petroleum Exploration Information [J].Journal of Computers, 2012, 7 (6): 1445-1452.