

Research on Domain Ontology Construction in Digital Library

Xiao-fei LI

Nangchang Normal University, China Nangchang

Email: 826907721@qq.com

Abstract. Learning resource navigation is a major problem that restricts learners' learning behavior. This paper proposed semantic approach based on ontology, which is helpful to solve the knowledge organization problem of digital library. The corpus was collected according to the 22 major categories of the Chinese Library Classification. It was divided into domain corpus and balanced corpus. Based on the knowledge base, information entropy and variance, linguistics, combined word length ratio, domain correlation and domain consistency, the domain concept was extracted. The cohesion hierarchical clustering method was used to extract the classification relationship between concepts. The log-likelihood ratio algorithm was used to extract the non-classification relations of domain concepts.

1. Introduction

It is easy for learners to get lost in the disordered and incoherent information resources on the internet. The automatic aggregation of information resources has become the focus area of library science and computer science research. Many clustering algorithms are only used for automatic classification of network documents and do not apply them to learning resources. And actually along with the book resources digitization advancement, the electronic book resources quantity increases year by year. Faced with the massive book resources in digital library, it is necessary to establish an aggregation mechanism to gather and integrate the knowledge units in the book contents from the logical semantics, so that they can help learning and become a meaningful resource structure. Resource aggregation is a knowledge system that organizes fragmented and scattered knowledge into logic. Therefore, it is very suitable to build semantic aggregation model with ontology. Ontology is a formal specification of Shared conceptual model[1,2]. Ontology is the key to achieve semantic interoperation.

2. Define domain ontology

Domain ontology refers to the ontology that describes the concepts and the relations between concepts in a specific domain of a specialized subject[3]. The ontology components include classes, attributes, relationships, and instances. The domain concept scope was defined first, then the class and its attributes were defined, the class relationships were defined, instances of each class were created, and property values were assigned. Domain ontology is the description of domain concept, so first is to determine the domain concept scope before the domain concept was extracted, and then is to collect related domain literature. Different ontologies can be built according to different disciplines with reference to the Chinese library classification. We can set up the law, literature, automation technology, mathematical science, health care, transportation, arts, history, education, management, environmental science, chemistry, language, computer technology, chemical industry, economic and so on a total of 22 domain ontology.



2.1. Domain concept extraction

Domain concept is not a collection of concepts in a book, but a collection of concepts in all literatures of a certain field. First, the domain corpus is acquired and 22 domain corpus are established according to Chinese library classification. Each category corresponds to a domain corpus, while other corpora are called control corpus or balanced corpus. The balanced corpus refers to a corpus of other domains in a certain domain. If 22 ontologies are established, 22 domain corpora need to be established. Each domain corpus corresponds to 21 other balanced corpora. The Chinese word segmentation tool ICTCLAS is used to pre-process the word segmentation of the domain corpus to obtain the set of terms and word frequency of each document. Adjacent related terms are recorded in the unit of sentences and stored in the matrix $SNT[M][M]$. M is the number of corpus terms in the whole field, and when two terms are adjacent then set $SNT[M][M]=1$, otherwise the value is 0.

2.1.1. Combine the algorithm of information entropy and variance to extract domain concepts

Currently, there are two difficulties in domain concept extraction based on statistics (such as TFIDF, conditional random field and mutual information). First, the sparsity of terms is such that the low-frequency terms cannot be extracted. Second is the term integrity problem, a complete terminology constitutes a domain concept. The domain concept has the characteristics of uneven field distribution, so the variance can be used to reflect this fluctuation. The larger the variance value is, the more likely it is the domain concept. Left-right information entropy (NE) can reflect the completeness of a term. By analyzing the boundary uncertainty of a string, the higher the left-right information entropy is, the stronger the independent integrity is, and the more likely the term is a domain concept. The formula to determine the domain concept that DV (s) is as follows:

$$DV(s) = NE \cdot \sigma \quad (1.1)$$

$$NE = LE(s) \cdot RE(s) \quad (1.2)$$

$$LE(s) = - \sum p(ls | s) \log_2 p(ls | s) \quad (1.3)$$

$$RE(s) = - \sum p(rs | s) \log_2 p(rs | s) \quad (1.4)$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (tf'_i(s) - \overline{tf'}(s))^2} \quad (1.5)$$

$$tf'_i(s) = \frac{tf_i(s)}{L_i} \quad (1.6)$$

$$\overline{tf'}(s) = \frac{\sum_{i=1}^N tf'_i(s)}{N} \quad (1.7)$$

N is the number of domain corpus and N is the variance. $LE(s)$ and $RE(s)$ represent the entropy of left and right information comentropy, and $p(ls | s)$ represents the conditional probability that s 's left adjacency is ls when s occurs. $p(rs | s)$ denotes the conditional probability that the right adjacency of s is rs when s occurs, and $tf_i(s)$ denotes the frequency of s in the i corpus. Considering the different document sizes in different fields, in order to eliminate the impact of domain size, L_i represents the total number of documents of corpus i and $tf'_i(s)$ represents the average frequency.

As the variance only reflects the differences of different fields of the concept, it cannot reflect the internal distribution of the fields. Therefore, this paper introduced another index of internal distribution consistency -- information entropy[4]. The greater the information entropy, the more domain documents that reflect the candidate concept, the more uniform the distribution within the domain of

the concept, and the more likely it is to be a domain concept. The field information entropy (DE(t)) formula is as follows:

$$DE(t) = - \sum_{i=1}^n P_t(d_i) \log_2 P_t(d_i) , P_t(d_i) = \frac{f_{t-d_i}}{\sum_{j=1}^n f_{t-d_j}} \quad (1.8)$$

$P_t(d_i)$ represents the distribution probability of the candidate domain concept t within the domain document d_i , and n represents the total number of documents within the domain. $P_t(d_i)$ represents the frequency of concept t in domain document d_i , and $\sum_{j=1}^n f_{t-d_j}$ represents the total frequency of t in domain document set. On the basis of the above mentioned variances and left-right information entropy, the candidate domain concept is obtained, and the domain information entropy is used to further select the domain concept.

3. Domain concept relationship extraction

The relations between domain concepts include classification relations and non-classification relations. Two relations between concepts are extracted separately.

3.1. Classification relationship extraction

In this paper, the relation of classification is extracted by the method of agglomeration hierarchy clustering. Suppose the concept set of clustering is N , and the $sim(i, j)$ of concept similarity is calculated by formula (2.2). The larger the similarity between concepts is, the closer the distance is. Formula (2.1) replaces it with the matrix $D[dis(i, j)]_{N \times N}$ of concept distance. Because the agglomeration hierarchical clustering algorithm depends on the initial concept similarity, this paper adopts the context plus dictionary method to calculate the similarity between concepts from a broader level according to the concept context, the wikipedia document and the Chinese thesaurus[5,6]. The formalized algorithm is as follows:

(1) Get the context words of a concept in a domain document in sentence units. The explanation documents of the concept are obtained from the encyclopedia entry, words and frequency of co-occurrence with the concept are obtained after word segmentation, and synonyms are extracted from the Chinese thesaurus to combine into a set T_i . Forming the vector $S_c = \langle T_i, W_i \rangle$ of concept C , W_i is word frequency.

(2) Calculate the cosine similarity $sim(i, j)$ between each concept to form the concept distance matrix $D[dis(i, j)]_{N \times N}$.

$$dis(i, j) = \log \frac{1}{sim(i, j) + 1.01} + 0.5 \quad (2.1)$$

$$sim(i, j) = \cos(S_{c_i}, S_{c_j}) = \frac{\sum_{W_i \in S_{c_i}, W_j \in S_{c_j}} W_i W_j}{\sqrt{\sum_{W_i \in S_{c_i}} W_i^2} \sqrt{\sum_{W_j \in S_{c_j}} W_j^2}} \quad (2.2)$$

(3) Set the $L(0) = 0$, $m = 0$, Scanning matrix D to find two clusters r, s with the smallest distance.

(4) The cluster $m = m + 1$ was merged with r . S was denoted as (r, s) , the clustering level was denoted as $L(m) = \{m, LM', \{r, s\}\}$, and LM' denoted the sub-layer number mark extracted from the sub-layer before the merge.

(5) Delete the row and column corresponding to r and s in the distance matrix D , add the merged

cluster (r,s) in the matrix D, and update the matrix according to $D(k,(r,s)) = \min(D(k,r), D(k,s))$.

(6) Repeat (4)-(6), set a distance threshold, calculate the distance over the threshold when the algorithm terminates.

After the hierarchical clustering, it is assumed that M clusters are finally obtained, and the central word of each cluster is extracted as the upper concept of the cluster. For larger clusters, the upper concept of K layer (K is a constant, generally taking 3 layers) is extracted. The formalized algorithm is as follows:

(1) Set $m = M$. The concept set $\{C_i | i = 1, 2, \dots, n\}$ of the cluster is extracted from $L(m)$, and the average distance $avgd[i]$ from each concept in the cluster to other concepts is calculated. Take the concept of minimum $avgd[i]$ as the central concept, as the upper concept of other concepts.

$$avgd[i] = \frac{1}{n} \sum_{j=1}^n D(C_i, C_j) \quad (2.3)$$

(2) If the number of concepts in the cluster is greater than the preset constant N, the upper concept of K layer is extracted, and sublayers $L(m_1)$ and $L(m_2)$ are obtained from $L(m)$, transfer to step (1).

3.2. Non-classification relationship extraction

Firstly, the candidate domain concept pairs with high co-occurrence rate of the domain documents were obtained through the association rule mining algorithm Apriori[7]. According to the rule support degree formula (3.1) and (3.2), the candidate concept pair set $\{FT1_i\}$ is formed by obtaining the concept pairs larger than the minimum support degree and larger than the minimum trust threshold value. The formula of support degree and trust degree are as follows:

$$\sup port(C_x \Rightarrow C_y) = p(C_x \cup C_y) \quad (3.1)$$

$$confidence(C_x \Rightarrow C_y) = \frac{\sup port(C_x \cup C_y)}{\sup port(C_x)} \quad (3.2)$$

Then, the log-likelihood ratio is used to calculate the correlation of the concept pairs, and the concept pairs with greater similarity are obtained. The logarithmic likelihood ratio has two assumptions:

Hypothesis H_1 : it is assumed that the concepts are independent of each other in the domain corpus, and the probability is p, that is $p(c_2 | c_1) = p(c_2 | \neg c_1) = p$.

Hypothesis H_2 : it is assumed that the occurrence of the concept pair in the domain corpus is not accidental, and the probability is respectively p_1, p_2 . $p(c_2 | c_1) = p_1 \neq p_2 = p(c_2 | \neg c_1)$.

Calculation formula of logarithmic likelihood ratio score:

$$\begin{aligned} SC_{c_1, c_2} &= -2 \log \lambda = -2 \log \frac{L(H_1)}{L(H_2)} = \\ &2 \log L(n_{c_1, c_2}, n_{c_1}, p_1) + 2 \log L(n_{c_2} - n_{c_1, c_2}, N - n_{c_1}, p_2) \\ &- 2 \log L(n_{c_1, c_2}, n_{c_1}, p) - 2 \log L(n_{c_2} - n_{c_1, c_2}, N - n_{c_1}, p) \end{aligned} \quad (3.3)$$

$$p = \frac{n_{c_2}}{N}, p_1 = \frac{n_{c_1, c_2}}{n_{c_1}}, p_2 = \frac{n_{c_2} - n_{c_1, c_2}}{N - n_{c_1}}, \quad L(k, n, x) = x^k (1-x)^{n-k}.$$

$$CPR(c_1, c_2) = SC_{c_1, c_2} * sim(c_1, c_2) \quad (3.4)$$

The frequency of concept occurrence in $FT1_i$ is counted in the unit of sentence. n_{c_1} and n_{c_2} represent the frequency of concept c_1 and c_2 . n_{c_1, c_2} represents the co-occurrence frequency of

concept pairs and N represents the total number of sentences containing concepts. As the log-likelihood ratio only considers the possibility of forming concept pair from word frequency statistics and lacks semantic analysis, this paper integrates the above mentioned formula (2.2) on the basis of the log-likelihood ratio. Calculate the relevance of the concept pair in terms of the concept context and dictionary, as in formula (3.4). Finally, the concept pair's CPR(c_1, c_2) greater than threshold was selected.

We need to obtain the semantic relations of concept pair to enhance the understanding of concept pair. The transitive verbs in general sentences express the semantic relations between concepts[8]. Therefore, the verbs between concepts were extracted as relational labels. In order to find verbs those are closely related to concept pairs, we use the method of VFICF (Verb frequency-inverse Concepts Frequency), which refers to verbs those are cooccurring with a certain specific concept and are weakly related to most Concepts. The formula is as follows:

$$VFICF(c_i c_j, v) = cp_{c_i c_j} - f_v \times \log \left(\frac{\left| \{ (c_i, c_j) \mid i \in m, j \in n \} \right|}{cp_{all} - f_v} \right) \quad (3.5)$$

$cp_{c_i c_j} - f_v$ refers to the frequency of $c_i c_j$. $\left| \{ (c_i, c_j) \mid i \in m, j \in n \} \right|$ denotes the total number of pairs of concepts, $cp_{all} - f_v$ denotes the number of occurrences of verbs and all pairs of concepts, and only once if the verb occurs more than once in a concept pair. The concept pairs $c_i c_j$ and the co-occurrence verb v are extracted from the sentences of the domain corpus. Calculate the $VFICF(c_i c_j, v)$ value of each verb. Several verbs with a larger weight of $VFICF(c_i c_j, v)$ are used as candidate labels for the concept pair $c_i c_j$. The concept pair was used as the preceding and the verb as the latter, and formula (3.2) was used to calculate the degree of trust to further filter the verbs with low trust. Finally, the verb is used as a relation tag to form a triple (c_x, v, c_y) .

4. conclusion

The domain ontology reveals the semantic relation between knowledge points and can realize the effective aggregation of knowledge. The domain ontology was constructed according to several major categories of the Chinese library classification. After further classifying knowledge from the overall situation of knowledge, the concepts were organized with semantic correlation technology, and the connotation connection of scientific knowledge was revealed in depth. Only with the capability of semantic navigation can the digital library truly go from information service to knowledge service.

References

- [1] Gruber T R. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition (S1042-8143), 1993, 5 (5) : 199-220.
- [2] Xiao Ximing, Wanyan Dengdeng. Research on Ontology-Based Semantic interoperability of public digital cultural resources integration. Journal of the National Library of China, 2015, (3): 43-49.
- [3] LI Zhi-long, WANG Dao-ping, GUAN Zhong-xing. Research on the Method of User Interest Model Construction Based on Domain Ontology. Information Science, 2015, (11): 69-73.
- [4] LIU Tao, LIU Bing-quan, XU Zhi-ming, WANG Xiao-long. Automatic Domain-Specific Term Extraction and Its Application in Text Classification. Electronica Sinica, 2007, 35(2): 328-332.
- [5] Mei Jiaju. Chinese thesaurus. Shanghai Lexicographical Publishing House, 1996.
- [6] Fu Jibin, Liu Jie, Jia Keliang, Mao Jintao. Ontology Relationship Extraction Research Based on HowNet and Term Relevancy Degree. New Technology of Library and Information Service, 2008 (9): 36-40.

- [7] He Haitao. Research on the concept extraction technology of domain ontology based on association rules. Changchun University Of Technology, 2015.
- [8] REN Fei-Lian, SHEN Ji-Kun, SUN Bin-Bin, ZHU Jing-Bo. A Review for Domain Ontology Construction from Text. Chinese Journal of Computers, 2017,(26): 1-26.