Domain Concept Extraction and Ontology Visualization

Xiao-fei Li¹,a
¹Nanchang Normal University, Nanchang, China
aEmail:jxlixiaofei@126.com

Abstract: A semantic navigation method based on ontology was proposed to solve the problem of digital library information trek. In this paper, the domain concept was extracted based on the knowledge base, linguistics, combined word length ratio, domain correlation and domain consistency. Finally, the domain ontology visualization scheme was put forward, and the related updating technology, wide-angle and aggregation technology were analyzed. The process of using Touch Graph to realize the navigation of ontology visual information was also analyzed.

1. Introduction
Domain ontology has the potential to solve the problem of digital library information trek in order to conduct visual semantic navigation. The domain ontology is an ontology that describes concepts and their relationships in a particular domain of a professional discipline[1]. Therefore, the extraction method of domain concepts and ontology visualization method were discussed below.

2. Domain concept extraction

2.1 Manually capture the core domain concepts
Firstly, the most representative domain literature should be obtained and the core domain concepts should be selected with the help of domain experts. Get the initial glossary and customize the glossary. The initial glossary is manually obtained from the field literature, such as medical Microbiology literature, financial glossary, epigenetics - terminology, Software Engineering guide SWEBOK V3, SNOMED CT medical terminology, etc. Chinese word segmentation adopted NLPIR Chinese word segmentation system, and the initial glossary was used as a custom dictionary[2].

The maximum frequent K term items were extracted for candidate domain concepts. Calculating the support of the concept, \[ \text{support}(t) = \frac{f_c}{|T|} \], \( f_c \) represents the total number of occurrences of the concept \( T \), \(|T|\) represents the total number of domain terms. When the degree of support is greater than the threshold, frequent 1-term items are extracted. Adjacent terms are merged according to the adjacency relation matrix to form T2, and then \( \text{support}(T2) \) is calculated to obtain frequent 2-term items. Frequent K-term items are cycled obtaining. Omitting a subset of the maximum frequent K term items from the domain term set to update the candidate domain concept set.

2.2 Combined with generic knowledge base to extract domain concepts
The extraction of domain concepts relies on WordNet, HowNet, Wikipedia and other common knowledge bases. WordNet is a semantic thesaurus developed by Princeton University in
2021 International Conference on Big Data and Intelligent Algorithms (BDIA 2021) IOP Publishing
Journal of Physics: Conference Series 1994 (2021) 012014 doi:10.1088/1742-6596/1994/1/012014

1985[3]. HowNet is a semantic organization structure of Chinese words similar to Word Net[4]. When the candidate concept is extracted from the field literature, the frequency F1 used in WordNet is found, and then the frequency index F2 of the candidate concept in the field literature is calculated. The RTD(Domain Relation Degree) of the candidate concept is \( F2/ F1 \). The concepts whose domain relevance are higher than the threshold are taken as concepts in the domain ontology. The concepts with hyponymy position relation, Part-whole relation, inheritance relation and near-synonymy relation are found from WordNet, and the domain concepts are extracted according to the RTD.

2.3 Extraction of domain concepts in conjunction with linguistics

To extract domain concepts, some filtering rules is needed by linguistic methods. The typical methods include part-of-speech filtering and ordinary word filtering. The rule of part-of-speech means that the candidate domain concept does not include interjection, phrase, onomatopoeia, locative word, pronoun, status word, does not start with auxiliary word, conjunctions, and post-conjunctions, and does not end with positional word, conjunctions, auxiliary word, or pre-conjunctions, but must have noun, verb, or quantifier in the concept[5]. Common word filtering is based on the common word list, filtering out the common words in the segmentation of domain corpus.

2.4 Domain concept extracting by combining word length ratio, domain relevance and domain consistency.

Word length ratio (WL) refers to the ratio of the lexical length of a domain concept to the average length of all words in the corpus. The larger the word ratio of the candidate domain concept, the more likely it is to be a domain concept. Domain correlation (Dr) refers to the probability distribution gap between candidate concepts in domain corpus and balanced corpus. The larger the value is, the more likely it is to be a domain concept. Domain consistency (DC) refers to the uniform distribution of candidate concepts in the domain. The more domain texts the concept appears, the more likely it is to be a domain concept. The calculation formula DD(t,C) is as follows:

\[
DD(t,C) = WL(t,C) \cdot (\alpha \cdot DR(t,C) + (1-\alpha) \cdot DC(t,C))
\]

(1)

\[
WL(t,C) = \frac{\sum_{i \in C} l_i}{\sum_{i \in C} f_i}, \quad WL_c = \frac{\sum_{i \in C} l_i}{\sum_{i \in C} f_i}
\]

(2)

\[
DR(t,C) = \frac{dr(t) - \min(dr(t_i))}{\max(dr(t_i)) - \min(dr(t_i))}
\]

(3)

\[
DC(t,C) = \frac{dc(t) - \min(dc(t_i))}{\max(dc(t_i)) - \min(dc(t_i))}
\]

(4)

\[
dr(t) = \sum_{d_i \in C} P(t, d_j) \log \frac{P(t \mid C)}{P(t \mid C_b)} , \quad P(t \mid C) = \frac{f_{t,c}}{N_c} , \quad P(t \mid C_b) = \frac{f_{t,b}}{N_b}
\]

(5)

\[
dc(t) = \sum_{d_j \in C} P(t, d_j) \log \frac{1}{P(t \mid C)} , \quad P(t, d_j) = \frac{f_{t,d_j}}{f_{t,c}}
\]

(6)

\( l_i \) represents the word length of the concept \( T \). \( WL_c \) represents the average word length in domain corpus \( C \). \( \sum_{i \in C} l_i \) is the sum of the word-length of all words in \( C \). \( f_{t,c} \) is the sum of word frequency of all words in \( C \). \( P(t \mid C) \) is the probability that \( t \) occurs in \( C \). \( P(t \mid C_b) \) represents the probability of \( t \) appearing in the balanced corpus \( C_b \). \( N_c \) represents the total number of documents in \( C \), \( f_{t,b} \) represents the word frequency of \( t \) in \( C_b \), and \( N_b \) represents the total number of documents in \( C_b \).
$f_{t,d_j}$ represents the frequency of $t$ in document $d_j$. $0<\alpha<1$, $\alpha$ is the empirical constant.

3. Construction and storage of the domain ontology
For the domain concepts extracted above, domain experts are invited to manually check and modify some errors, then Protege can be used to construct the domain ontology. The established domain ontology is expected to export as an RDF document and stored in a database. The database that stores ontology generally should be the database system that supports SPARQL language, such as NEO4J.

4. Visual retrieval of domain ontology
Visual information retrieval refers to the transformation of invisible internal resources and the semantic relations between resources into visible graphics, which are displayed in the form of 2D and 3D spatial graphics. It includes two processes: resource rendering and user interaction[6]. The basis of visual information retrieval is ontology representation of resources, computer graphics and cognitive psychology. The visualization technology of information retrieval includes correlation updating technology, wide Angle and aggregation technology. Associated update refers to a linkage change between two Windows. Wide Angle and aggregation means that there is always a specific individual information that the user is currently focused on that needs to be magnified, the rest of the profile information is associated with the aggregation information, and the aggregation can be switched. The requirements of resource depth visualization are: user-led, real-time visualization, easy to operate, intuitive image, multiple sources, and full interaction[7].

The visualization application tools of ontology can be divided into visualization plug-ins based on Protege and general visualization tools independent of ontology construction tools. Protege visualization plug-ins include OWLViz and OntoSphere. OWLViz shows the hierarchical relationship of ontology in a tree structure, while OntoSphere shows the three-dimensional structure of ontology. Such a visualization tool is not easy to be associated with the information retrieval system, while a Touch Graph visualization tool independent of ontology tool is easy to be integrated with the information retrieval system, and it can also support the plug-in TGViz of Protege, so it is very suitable to be used as a visualization tool of domain ontology.

Prefuse is another tool that looks like Touch Graph. First, Jena is used to extract triples from the ontology OWL file. This is followed by a series of mapping procedures that mapping classes and instances to nodes and mapping object relationships to edges. The labels on the edges represent relationships between nodes. Nodes can be represented by different colors and shapes. Edge lines can be either thick or thin, long or short, depending on the degree of association between ontology classes. At the same time, user interaction is supported. When a user clicks a node, the current node immediately gathers, and a series of operations complete data mapping and feedback from the background to redraw nodes and edges. This technique can be implemented with Ajax, and the module is responsible for real-time data exchange between the client and the server.

You can also visualize it in the JSON data format, which is a lightweight data exchange format. Jena can directly generate JSON formatted output from RDF data, and then visually present the data using JavaScript visualization tools. The visualization of ontology is also connected with the information retrieval system. When a patron clicks a graph node, the ontology display area immediately generates a new graph. At the same time, the associated update technology is adopted to update the information retrieval interface. Keywords in the ontology are automatically entered into the retrieval system, and the retrieval system returns the retrieval results to the retrieval interface. The navigation process of ontology visual information retrieval is shown in Figure 1:
5. Conclusion
In the future, we will continue to study several problems in depth. First, the orderly aggregation of individual knowledge resources requires the study of the relationship between the forerunner and successor of several individuals to make learning plans for learners. The second is the identification of unregistered words of domain concept. Unregistered words are the difficulty in the current research, but they are important domain knowledge and determine the completeness of domain ontology. Therefore, they need to be mined and filled into the domain ontology library. In the future, we will continue to study several problems in depth. First, the orderly aggregation of individual knowledge resources requires the study of the relationship between the forerunner and successor of several individuals to make learning plans for learners. The second is the identification of unregistered words of domain concept. Unregistered words are the difficulty in the current research, but they are important domain knowledge and determine the completeness of domain ontology. Therefore, they need to be mined and filled into the domain ontology library.

Acknowledgments
This paper is supported by Science and Technology Project fund form Jiangxi province ministry of education. The project name is "Research on Construction and Visualization of Digital Library Domain Ontology", (Grant numbers : GJJ181089).

References
[1] 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.
[2] Huaping Zhang. NLPIR. http://ictclas.nlpir.org.
[3] Zuang Miao, Yafei Zhang. Reusing WordNet for Building Domain Ontology. Journal of System Simulation, 2007, (19): 4583-4586, 4598.
[4] Peng L, Zhang Y J, Zhang T. Research on How Net-based Chinese Word Lexical Semantic Similarity Measurement. Proceedings of the 2009 International Conference on Machine Learning and Cybernetics, 2009: 282-287.
[5] Liu Li, Xiao Yingyuan. A statistical domain terminology extraction method based on word length and grammatical feature. Journal of Harbin Engineering University, 2017, (9): 1-9.
[6] Yan Chen. Information Retrieval Visualization Technology. Information Studies: Theory & Application, 2006, 29(5): 618-621.
[7] Yufeng Zhang, Yitang Zen. The Visualization Framework of In-depth Aggregation Results of Library Resources Under Semantic Environment. Documentation, Information & Knowledge, 2014, (5): 65-71.