Research on Domain Ontology Generation Based on Semantic Web

Jiguang Wu, Ying Li

Research on domain ontology generation based on Semantic Web

Jiguang Wu[1,2],Ying Li[1,2]

1 Radio and Television Information Security Research Institute
Department of Computing, Communication University of China
Beijing 100024, China
2 Communication University of China, Beijing 100024, China
{j.g.wu, liy}@cuc.edu.cn

Abstract. This paper focuses on the generation of domain ontology. At first, the paper introduces the traditional search engines and pointed out their shortcomings, in addition ,it also illustrats that retrieval should be based on the semantic web, Then the concepts of semantic web, domain ontology and so on are proposed, next it makes a research in the plsa algorithm of extracting domain concepts and the k-means algorithm of clustering thoese concepts, finally, it showes a football ontology constructed by protégé , and makes a prospect to semantic retrieval based on ontology.

Keywords: domain ontology, semantic web, plsa algorithm, k-means, protégé

1 Introduction

At the early stage of the development of the World Wide Web, people can get the data or information they want only through a specific URI. With the continuous development and progress of science and technology, especially the Internet technology, which is particularly prominent, therefore, with the emergence of the search engines, people can access the relevant information through the portal of the search engines. Search engines, such as Bing, Google, HotBot and so on,[9] they have greatly improved the efficiency of people obtaining information.

In the age of information and data explosion, their appearance has provided us with great convenience, to a certain extent, we can't be separated from them. However, search engines have not fully achieved the desired value of human beings, they are based on the keywords' matching or the simple logic and/or relationships between keywords, after that they link to related web pages and return them to users, and then users choose from the results, finally users get useful information for themselves, they do not analyze or process the content input by users at the semantic level, therefore the data they return includes much information we do not want. For example, users input" The songs of Michael Jackson" in the search box, but search engines may not return Jackson's songs to us, just a few relevant web pages.

Therefore, in this situation, the concept of semantic web was proposed in 2000 by Berners-Lee Tim, the founder of the world wide web, he expects that computer can understand the content of the query from the semantic meaning to solve current Internet data searching and sharing problems.

2 Semantic web and ontology

2.1 What is semantic web

The semantic web is not an independent network, which is an extension of the world wide web, simply speaking, [10]the semantic web is an intelligent network capturing meaning. Nowadays most of the web pages are designed for human beings, these include videos, pictures, sound, texts, and other forms of expression, but the machines can't read them directly, they can only read the program or data from the database. As a result, the emergence of semantic web is in order to make up for this deficiency, it can complete the network as much as possible with the semantic information and facilitate interaction between human beings and computers, in this way, [8]the whole Internet may become a universal information exchange medium. Fig.1 gives a hierarchical structure of the semantic web.

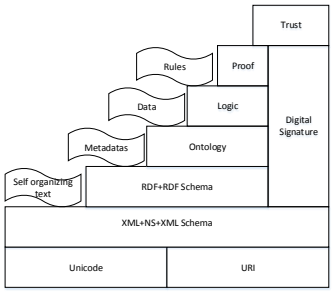


Fig.1: A hierarchical structure of the semantic web

They are Unicode and URI、XML、RRDF and RDFS、Ontology、Logic、Proof and Trust from bottom to top.[12] From the hierarchical structure we can see in the level of Semantic Web:

- Unicode and URI, they are the basic of the Semantic Web, Where Unicode is responsible for handling resource coding, URI is responsible for the identification of resources;
- XML+NS+XML Schema pattern, mainly be used to show the data content and structure;
- RDF and RDFS pattern, mainly be used to describe resources and their types;
- Ontology mainly describes the relationship between the various resources;
- Logic Layer provides the rules for intelligent reasoning， logical reasoning is performed on the basis of its following levels;
- Trust Layer ensures information security and build trust relationships between users;

Ontology is the core technology of semantic web, It describes and defines rich semantic relations between resources, not only so, it makes a complete formal description of the information so that the information can be understood and processed by the computer. To realize the semantic web technology, we need to have enough and rich ontologies to support it， consequently it is very important to establish a perfect domain ontology for the development of semantic web.

2.2 What is Ontology

Ontology originates from philosophy, which is a description of the nature of things, at the same time, it is proposed to avoid ambiguity in the field of philosophy. In the field of science and technology, ontology was first proposed by Neches et al[14], they thought that ontology gives the basic terms and the relationships between the terms in a certain field, what's more, it can be utilized to formulate the rules for the extension of these words. The first widespread definition of ontology was proposed by Tom Gruber[15], a scholar of the Stanford University, that an ontology is an explicit specification of the conceptual model and ontology explains the concepts. Later, many scholars made their own interpretation of the ontology， what Studer et al. proposed[16] was the most representative among them, they considered that ontology was a formal specification of a shared conceptual model, which has made the four features of the ontology show out, that are: Conceptualization, Explicit, Formal and Sharing.

Conceptualization is a model that can be abstracted from some phenomenas in reality; Explicit is that these conceptual models are abstracted and the constraint conditions that use these concepts are not ambiguous; Formal indicates ontology should be able to be processed by computer; Sharing refers to what ontology reflects is the common knowledge shared by the public, rather than just one individual.

In 1999, Perez et al proposed[17] that ontology consists of 5 basic elements, they were as follows: class, relation, function, axiom, instance. Class, also is known as the concept, can refer to any concepts that exist in reality and also be able to refer to a number of functions, processes, behaviors, strategies, etc; relation represents the relationships or interactions between classes in a particular field; function is a kind of special relation, the last element in this relationship can be uniquely determined by the first n-1 elements, formal definition is:C1 ×C2 ×C3 ×… ×Cn-1-> Cn; the axiom represents a true assertion, it is used to define the restrictions and rules between concept and attribute; an instance is a specific entity of a concept. Analysing from the aspect of semantic web, an instance represents a object, however, the concept is a collection of objects or a set of objects correspond to an object. There are four basic relationships between concepts, as shown in Table 1.

Relation name	Meaning
Part of	The relationship of the concept between the part and the whole
Kind of	The inheritance relationship between concepts, which is similar to the relationship between parent and child
Attibute of	A concept is a property of another concept
Instance of	The relationship between an instance of a concept and a concept

Table.1: Basic relationships between concepts

Nevertheless, [11]the construction of ontology is not limited to these five elements in practical applications and relations between concepts are not limited to the four types as shown in table, developers can define themselves according to their needs. Figure 2 represents a simple football body.

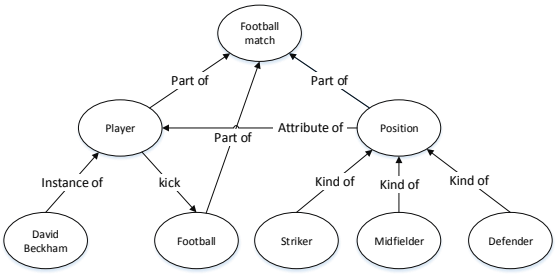


Fig.2: A simple football ontology

Therefore, to sum up, [13]ontology is a kind of formal description of the relationships between concepts in a specific field.

3 Extraction domain concepts

3.1 Corpus preprocessing

First of all, we need to make a preprocessing to the domain corpus, which includes word segmentation, part of speech tagging and removing of stop words and unuseful symbols. The stop words have a very high frequency in the corpus but are extremely low in meaning, and they will also cause interference to the extraction of domain concepts, so we should remove them. When word segmentation and part of speech tagging are made, the verb and noun in the corpus are extracted as candidate concepts, in order to be prepared for extracting domain concepts from the candidate concept set.

3.2 Algorithm of extracting domain concepts

Ontology is a model to describe relationships between concepts, therefore, it is necessary to extract relevant domain concepts to construct ontology. The traditional methods to extract domain concepts are based on statistics, However, these methods can only be used to extract the words with a high frequency in a text file, some low frequency words which play a very important role in the construction of a domain ontology may be ignored, hence, using statistical methods to extract domain concepts to construct domain ontology is not authoritative.

Statistical methods can not be ignored, but can not just rely on statistics, thus we should start to speculate from the semantic level of concepts, and then mix it with statistical methods together to extract the concepts. First of all, this paper introduced the concept screening method based on statistics.

ⅰ TF-IDF(Term Frequency–Inverse Document Frequency)

TF-IDF a common statistical method to assess the importance of a word to a collection of documents. TF is named as the word frequency, which can be used to measure a word's ability to describe the content of the document; IDF is known as reverse document frequency, which is used to measure a word's ability to distinguish documents. The TF-IDF algorithm is based on the assumption that: The feature words in the documents should be those words that appear frequently in the domain documents and appear with a low frequency in other documents of whole document collection. It takes the ratio of the number of feature words in a document and the number of all documents containing the feature word as the weight of the word. For the word t_i in a particular file, the importance of it can be expressed as follow:

$$TF_{ij} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

In formula, the numerator is the times of the word appears in the file d_i, the denominator refers to the sum of the number of all the words appear in the file d_i.

IDF(inverse document frequency)

$$IDF_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|}$$

In formula, the numerator is the total number of documents in a corpus, the denominator represents the number of files containing the word t_i, (the number of file $n_{i,j} \neq 0$). If the corpus does not contain the word, the denominator will be 0, so generally use the denominator $1+|\{ j:t_i \in d_j \}|$.

The high word frequency in a particular file and the low file frequency of the word in the entire document collection can generate a TF-IDF with high weight, so TF-IDF can preserve the feature

concepts of the domain field.

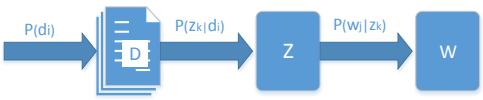
ii Plsa algorithm

Compared with MI, TF-IDF and other algorithms, plsa algorithm is much more advanced, it solved the problem of synonyms and ambiguous words and trained implicit classes by using the expectation maximization algorithm (EM) [2], plsa also made an expansion in semantic aspect in the context of a solid statistical basis, so I chose the plsa to extract the field concepts in the experiment. Before introducing the PLSA algorithm, we must first understand the model topic， which is a modeling method for a implicit theme of a text. A theme is a concept or one aspect that shows as a series of the relevant words which can represent the theme. Describing mathematically is that: a theme is the conditional probability distribution of the words in the vocabulary. The more closely related to the theme, the greater the conditional probability is. [1]Giving the following definition: d, w and z represent document, word and implicit theme respectively：

$$p(w|d) = \sum_z p(w|z)p(z|d)$$

p(w | d) represents the probability that the word w appears in the document d, as for training corpus, make word segmentation for text， the ratio of the frequency of the word and the frequency of all words in the documents can be calculated, for the unknown data, model is used to calculate the probability value. P(w | z) represents the frequency of a word appearing on the premise of a given theme, which describes the correlation between the word and the theme. P(z | d) refers to the probability of each topic appearing in documents, So the theme model is to make advantage of a large number of known word document information to get p(z | d) and P(w | z).

Plsa is a kind of topic model, with a certain probability to select the theme z corresponding to the d after a given document d, Hofmann proposed the PLSA model based on probability statistics in SIGIR'99, and the EM algorithm is used to study the parameters of the model. [3,5]The probability model graph of PLSA is as follow:



In the graph above, the D represents documents, the Z is a implicit theme, the W is a observed word, $P(d_i)$ represents probability of the word appearing in document di, $P(z_k | d_i)$ represents the probability of the word under the condition of giving theme z_k appearing in document di, $P(w_j | z_k)$ represents the probability of appearing the word w_j under the condition of giving theme z_k, and each theme obeys the Multinomial distribution for all words, not only so, each document also obeys the Multinomial distribution for all topics. The entire document's generation process steps are as follows:

(1) Select the document d_i with the probability of $P(d_i)$;
(2) Select the theme z_k with the probability of $P(z_k | d_i)$;
(3) Generate a word with the probability of $P(w_j | z_k)$.

What we can observe is data pairs (d_i, w_j), z_k is implicit variable, and union distribution of (d_i, w_j) is as follow:

$$p(d_i, w_j) = p(d_i)p(w_i|d_j),\ p(w_j|d_i) = \sum_k p(w_j|z_k)p(z_k|d_i)$$

$P(z_k | d_i)$ and $P(w_j | z_k)$ correspond to the two groups of Multinomial distribution respectively, then we

need to estimate the parameters of those two groups' distribution, that is given P(z, d) and obtain P(z, d) and P(w, z).According to P(d, w) to structure likelihood function, then maximize it, likelihood function is as follow:

$$L = \sum_{i}^{N}\sum_{j}^{M} n(d_i, w_j)\log p(d_i, w_j)$$
$$= \sum_{i}^{N} n(d_i)\left[\log p(d_i) + \sum_{j=1}^{N}\frac{n(d_i, w_j)}{n(d_i)}\log\sum_{k=1}^{K} p(w_j|z_k)p(z_k|d_i)\right]$$

We can learn from the likelihood function that the independent variables are P(z, d) and P(w, z), however, there is the additive relation in function so that it will be extremely difficult to take the derivative of the likelihood function, consequently, the [4]EM algorithm is applied. Its basic ideas are:

1. E step: calculate a posteriori probability out under the condition of implicit variables have given current estimated parameters.
2. M step: Maximize Complete data's expectation of logarithmic likelihood function, now use implicit variables' posteriori probability to calculate according to E step we can get new parameter values.
3. Calculating the two steps iteratively until convergence.

In E step, calculating implicit variables' posteriori probability under the condition of current parameter values are calculated by using Bias formula, that is:

$$p(z_k|d_i, w_j) = \frac{p(w_j|z_k)p(z_k|d_i)}{\sum_{l=1}^{K} p(w_j|z_l)p(z_l|d_i)}$$

In this step, assumed all $P(z_k \mid d_i)$ and $P(w_j \mid z_k)$ are known, because the initial value is random assignment, so the parameter values in later iterative process obtained from the M step of previous round.

In M step,maximize Complete data's expectation of logarithmic likelihood function, in plsa algorithm, Incomplete data is (d_i, w_j), implicit variable is theme z_k, and complete data is triple (d_i, w_j, z_k), it's expectation is that:

$$E[L] = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\sum_{K=1}^{K} P(z_k|d_i, w_j)\log\left[p(w_j|z_k)p(z_k|d_i)\right]$$

$(z_k \mid d_i, w_j)$ is known, it is the estimated value of the previous E step, then maximize expectation, that is to calculate function extremum above. The method is Lagrange multiplier, objective function is E[L] in plsa algorithm, constraint conditions are as follows:

$$\sum_{j=1}^{M} p(w_j|z_k) = 1$$
$$\sum_{k=1}^{K} p(z_k|d_i) = 1$$

Lagrange function can be written at this time:

$$H = E[L] + \sum_{k=1}^{K}\tau K(1 - \sum_{j=1}^{M} p(w_j|z_k)) + \sum_{i=1}^{N}\rho i(1 - \sum_{k=1}^{K} p(z_k|d_i))$$

First of all, obtaining partial derivative of the function, then by uniting equations we can get parameter values of E step. Finally enter E step again and use new parameter values at the same time to calculate a posteriori probability out under the condition of implicit variables have given current estimated parameters . Calculating through the way of constant iteration until eventually meet the termination conditions.

4 Acquisition of classification relationship

Ontology describes concepts and the relationships between concepts in a specific field, after extracting a field concept set by the PLSA algorithm, the next step is to cluster these concepts in order to get classification relationships between concepts, the second part of the paper has made an overview of the basic relationships between concepts, the clustering of concepts is mainly to get classification relationship, that is to say the words that are semantically similar are clustered into one class by clustering algorithm. The core of clustering is to measure the distance between elements, by calculating the distance between a element and a center of the classes to determine whether the element should be aggregated into the same one class. Common clustering algorithms have K-means and hierarchical clustering, however, because the hierarchical clustering algorithm is more complex and can not achieve better effect, so I selected K-means algorithm[6] in this experiment, but the K-means algorithm needs to set a size of the classes at the time of initialization, so by constantly debugging, it is found that when the value of K is adjusted to 5, the effect is relatively good.

Algorithm thought: [7]calculating the distance between different samples to determine the close relationship between them, If the relationship is close, and put them in the same one class.

(1) First of all, choosing a K value, it means to how many types the data is divided. The selection of K value is very important, and it has a great influence on the result., then select the initial clustering point. The clustering point of this experiment is to select randomly in the data, by using the way of calculating average value repeatedly to avoid converging to a local minimum value.

(2) Finally, calculating the distance between all points of the data set and the clustering points, and then add them to the class that is closest to their clustering points. Next calculating the average value of each cluster and regard the point as a new clustering point. Repeat these two steps until the convergence, in this way we can get the final result.

Algorithm's pseudo code
def loadDataSet(fname)
Read data set from file
def distEclud(vecA , VecB)
Calculating Euclidean distance between two concepts by the vector
def randCent(dataset , k)
generating randomly initial clustering points, and select randomly points within the range of data points in code
def kMeans(dataset , k ,distMeas = distEclud , createCent = randCent)
Define k-means algorithm, input data set and K value
Show(dataset , k ,centroid, clusterAssment)

Fig.3 shows the algorithm's clustering effect, a diamond in the figure is a random cluster point, each color represents one class, by constantly adjusting the value of K, and when we adjusted the value of K to 5, the result was relatively satisfactory. Of course, the result still has a lot of room to improve, later research we will continue to improve the result of clustering.

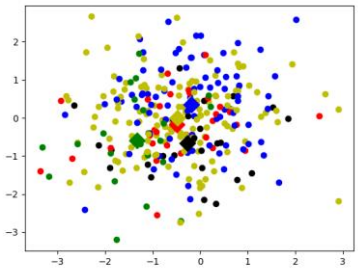


Fig.3: The clustering result of K-means algorithm

Apart from the existence of taxonomic relationships, there are non taxonomic relationships in field concepts, the extraction of non taxonomic relationships between concepts is mainly based on the association rules algorithm in this experiment, by calculating the confidence and support between two concepts to determine whether they exist non classification relationship, Then, in the context of the concept, the relationship between the two concepts can be described by statistical features. And then, in the context of the emergence of the concept pair, using statistical features to obtain the relationship between the concept pair. However, the accuracy by using this method to extract the non classification relationship is relatively low, and it also needs to be artificially modified.

5 Ontology construction

The purpose of this experiment is to construct a football ontology, through a corpus of football news we extracted a number of concepts in the field of football and also obtained the relationship between concepts by using clustering algorithm. With the classes and relationships between a concept pair we can build a football ontology, we used ontology editing tool named protégé developed by Stanford University in this experiment, it is not only able to edit ontology, but owns the function of Ontology visualization. As shown in Fig.4, through the OWLViz of protégé's plug-in unit we can clearly see that all concepts in domain field are clustered to five categories, and it shows the relationship between a concept pair.

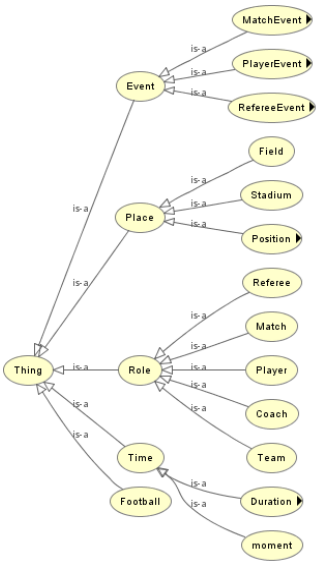


Fig.4: The football ontology's main classes and the Classification relationships

What's more, through Fig.5 we can more directly and clearly see the whole football ontology's all concepts and relationships between a concept pair. Ontology mainly reflects the classes and relationship

between concepts, in addition to the concepts, a solid line shows the classification relationship between concepts, while a dotted line represents the non taxonomic relation between concepts.

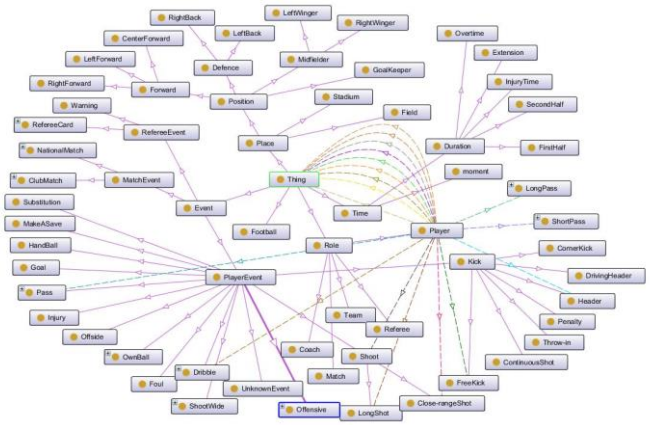


Fig.5: The football ontology's visualization

6 Conclusion

In this paper, we introduced the generation process of domain ontology based on Semantic Web, and introduced in detail the acquisition of domain concepts by the method of combining statistical features with PLSA algorithm, then the domain concepts are clustered by the K-means algorithm, finally, the ontology was visualized by using the ontology editing tool protégé . We can apply the domain ontology to semantic retrieval, search engines can understand the semantics of searching keywords to match concepts and then provide more relevant results, in this way, search engines' efficiency of retrieval can be improved a lot to a certain extent. However, how to automatically and accurately obtain non classification relationships between the concepts has still much room for improvement, we will put more effort into this aspect in futural research.

References

1. T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of SIGIR'99,1999.
2. S.Deerwester, S. T. Dumais, G. W. Furnas, Landauer. T. K., and R. Harshman. Indexing by lantent semantic analysis. Journal of the American Society for Information Science, 41, 1990.
3. Thomas Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning 42, no. 1 (January 1, 2001): 177-196.
4. Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.
5. Qiaozhu Mei, A Note on EM Algorithm for Probabilistic Latent Semantic Analysis 2008.
6. P.S. Bradley and U. Fayyad, "Refining Initial Points for K-means Clustering," Proc. 15th Int'l Conf. Machine Learning, pp. 91-99, 1998.
7. K. Alsabti, S. Ranka, and V. Singh, "An Efficient k-means Clustering Algorithm," Proc. First Workshop High Performance Data Mining, Mar. 1998.
8. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – the story so far. International Journal on Semantic Web and Information Systems, 5(3):1–22, 2009.
9. Xiajiong Shen , Yan Xu, Junyang, Yu Ke Zhang, "Intelligent Search Engine Based on Formal Concept Analysis" ,IEEE International Conference on Granular Computing, pp 669, 2-4 Nov, 2007.

10. Grigoris Antoniou and Frank van Harmelen. A Semantic Web Primer (Cooperative Information Systems). The MIT Press, April 2004.
11. Nancy Ide and David Woolner. Historical Ontologies, chapter Words and Intelligence II: Essays in Honor of Yorick Wilks, pages 137–152. Springer, 2007.
12. Maximilian Kalus. Semantic networks and historical knowledge management: Introducing new methods of computerbased research. Ann Arbor, MI: MPublishing, University of Michigan Library, 2007.
13. Gruninger, M. and Fox, M.S. (1995). Methodology for the Design and Evaluation of Ontologies. In: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal.
14. Neches, R., Fikes, R. E., Finin, T., Gruber, T. R., Senator, T., & Swartout, W. R. (1991). Enabling technology for knowledge sharing. AI Magazine, 12(3):36--56.
15. Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5:199--220.
16. S. Staab and R. Studer (eds.), Handbook on Ontologies, International Handbooks on Information Systems, DOI 10.1007/978-3-540-92673-3, Springer-Verlag Berlin Heidelberg 2009.
17. [Arpirez-Vega et al., 1998] J. Arpirez-Vega, A. GomezPerez, A. Lozano-Tello, and H. Sofia Pinto. (ONTO)2 Agent: An Ontology-Based WWW Broker to Select Ontologies. In Proceedings of ECAI98's Workshop on Application of Ontologies and Problem Solving Methods, pages 16–24, 1998.