Ontology learning and mapping in semantic web based on formal concept analysis technology

1 Yi Wang, 2 JieHong Luo
*1,Corresponding Author, 2 Department of information engineering, Guangdong Textile Polytechnic, Foshan Guangdong China, wangyi02023026@sina.com

Abstract

Ontology learning is a sub-task of information extraction, the goal is automatic or semi-automatically from a given corpus or data source to extract concepts and relations. Ontology mapping is to calculate the two main elements of the similarity between ontology, the similarity between elements and ontology definition, ontology and ontology-related rules or constraints. Ontology and formal concept analysis are derived from the philosophy, concept lattice formed by the relationship between the level concept, connotation and extension of the concept of form. This paper presents to ontology learning and mapping in semantic web based on formal concept analysis. The experimental results show that this method can effectively improve the performance of the concept and semantic information in semantic web.

Keywords: Ontology learning, Formal concept analysis, Ontology mapping, Concept lattice

1. Introduction

Ontology learning is a sub-task of information extraction, the goal is automatic or semi-automatically from a given corpus or data source to extract concepts and relations, the formation of the body. Ontology learning is an interdisciplinary task, typically, the term extraction process from the beginning the task, and usually includes several language processing (eg word, POS tagging, etc.); Then, through statistical methods to extract rules or relationships; final the concepts and relationships together to form a body. Ontology mapping is to calculate the two main elements of the similarity between ontology, the similarity between elements and ontology definition, ontology and ontology-related rules or constraints, so the mapping method is also divided into principal and interest with the body structure (defined) related methods, and instance-dependent method and the method associated with the field restrictions.

The mathematical basis of FCA proposed by Birkhoff, Birkhoff lattice structure and to clarify the partial order between the reciprocal relationships. He pointed out that a box set of objects and attributes can be set between binary relations to each building, the grid can be obtained through insight into the relationship between the original data structure. Wille first proposed to construct a system based on binary relation corresponding concept lattice (Galois lattice) ideas, also known as formal concept analysis, is to format each node represents a formal concept, which represents the corresponding extension of the concept of a group of objects, meaning was this group of objects with common characteristics (attributes) [1]. Concept lattice and Hasse diagram corresponding to the image reveals the generalization between concepts of the relationship with patients, reflecting a conceptual hierarchy (Concept Hierachy), to achieve the visualization of data, ideal for the database described in the knowledge discovery, data analysis and rules to become an effective tool to extract.

The basic semantic idea is to extend the current, making the network with all the information is semantic, the computer can understand and handle, and a computer to facilitate interaction and cooperation between. Ontology model using concept lattice and the body, said similarity measure is very effective, Fan proposed a formal concept analysis using ontology similarity measure method to extract the body in addition to the concept of equivalence between the outside and further extracted a hierarchical relationship between the ontology concepts, but because the diversity of body expression, using the extracted lattice structure types are still two kinds of relationships is incomplete.

This article proposes to ontology learning and mapping in semantic web by formal concept analysis to make up for these ontology learning deficiencies. So the formal concept analysis method can help constructing ontology and mapping, and provides unified, mature method guidance for the construction
of the ontology. First, the use of semantic properties similar to table a unified representation of ontology concepts and properties similar methods; then use the algorithm to form the background to do further processing to keep the information more complete; the final meaning of the concept used to measure the similarity between ontology concepts similarity, and the structure of concept lattice obtained by a body different types of relationships between concepts.

2. The research on ontology learning and ontology mapping technology

As time-consuming manual methods, allowing the body to build an arduous task. Therefore, how to use the body of knowledge acquisition reduces the cost of building a meaningful research. Study abroad in the direction of the very active, the related technology called ontology learning techniques (Ontology Learning), its goal is to use machine learning and statistical techniques such as automatic or semi-automatic from the existing data resources to get the desired body[2]. First, the dictionary can rely on ontology learning and knowledge base for ontology construction, following a brief introduction of these two methods.

(1) dictionary-based ontology learning

In view of the traditional dictionary definition of each word as a synonym, root, prototype and other relations, using the construction method is that the relationship between words and vocabulary words — upper and lower strata of the concept of the word to determine the relationship. To dictionary-based method for the construction of the basis of other construction methods, however, constructed in this way is usually a general description of the body, and is not associated with specific areas of the body, it must be combined with other methods as well as the participation of experts from the fields to form meaningful body structure, therefore methods can be used independently; in addition, the construction method is not only limited the scope of the size of the dictionary itself, and the formation of sub-areas of different areas, there can not adapt to environmental changes caused by omission of information required.

(2) Knowledge-based ontology learning

To knowledge-based construction methods must construct knowledge in related fields, built mainly by the knowledge base consisting of some of the basic rules of composition and simple examples. When the user to find relevant information, the first type the keyword, and then using knowledge within the rules to filter out a lot of irrelevant information, and using similar Example to do comparison, pumping select the most relevant results, and finally using the knowledge database of rules to re-construct the relevant body and can give summary and results. This method is the difference with other methods: Knowledge of the existing body of rules can be considered a show, but only reuse these existing rules within the knowledge base related to the combination of small body.

According to the structure of the source data level, but also ontology learning techniques can be classified as follows: structured data based on ontology learning; unstructured data based on ontology learning; semi-structured data based on ontology learning. The following were introduced.

(1) Structured data based on ontology learning

Structured data mainly include object-oriented relational database or data in the database. Now they use a relational database application to organize and store data. But the relational model has a fatal weakness, that it can not be expressed using a table model, the semantics of complex objects. Therefore, it is not suitable for complex data types are numerous and semantic modeling of information systems field. Body pulled from the database, the general practice: the use of database reverse engineering, or mapping technology to the relational model into an intermediate model, and then the body of the middle model conversion costs [3]. For example: the relational model into a conceptual model, the conceptual model is actually the extended entity - relationship model of formalized representation, then by the user to revise the conceptual model to generate the final body, as shown by equation1.

\[
\hat{S} = g(Y_1[n_1], \ldots, Y_k[n_k], \ldots, Y_{n_k}[n_{n_k}])
\]

(1)

(2) Unstructured data based on ontology learning

Unstructured data is data that there is no fixed structure, plain text is the Web in the presence of a class of unstructured data. Currently, the unstructured data based on ontology learning technology research focused on access from the plain text in the body. Plain text according to certain syntactic expression of specific semantics, so the reader can be based on some background knowledge to understand the meaning. The absence of a certain structure, to make the machine can automatically
understand text and extract the required knowledge; we must use natural language processing (NLP) technology to its pre-treatment as shown by equation 2, and then use statistics, machine learning and other means to acquire knowledge from.

\[
\text{corr}(S_i, S_j) = \rho_{ij} = K_{\sigma}(d_{ij}) = \frac{E[S_i, S_j]}{\sigma^2_S}
\]  

For the concept of access: the concept of common statistical method is to calculate the concentration of the frequency in the text, if the frequency is greater than a specified threshold, it is in the domain ontology as a concept; to obtain the relationship between concepts: based on the model, the concept of clustering, association rule mining method. Model-based approach need to determine the sequence of words in the text matches a pattern, if the match, you can identify the corresponding relations. Conceptual clustering method is to use the concept of semantic distance between, on the concept of hierarchical clustering; clustering result is the classification of relationships between concepts. Association rule mining method used to obtain non-classified relations between concepts, the basic idea is that if two concepts often appear in the same document (or paragraph, sentence) as is shown by figure 1, then this relationship must exist between the two concepts.

\[
\text{Figure 1. Ontology learning flow chart}
\]

(3) Semi-structured data based on ontology learning
A large number of XML format and HTML format pages, and they follow the document type definition (XML Schema or DTD) and so has the implicit structure of semi-structured data is data. Ontology learning is to use some of the mapping rules to derive the body. In addition, the machine-readable dictionary is a special kind of semi-structured data, often using linguistic analysis, semantic analysis and pattern matching methods to obtain the domain-specific concepts and concept relationships [4]. Specifically: In 2002, Papatheodorou et al proposed an XML or RDF format from the documentation to get the concept of the relationship between the classification method; 2003, Volz et al proposed an XML Schema and DTD based on the ontology learning.

\[
\text{corr}(S_i, S_j) = \rho_{ij} = K_{\sigma}(d_{ij}) = \frac{E[S_i, S_j]}{\sigma^2_S}
\]  

The so-called ontology mapping, is generated in the body has been established on the link, so that both sides can use a common interface, the same thing with a common understanding. Mapping rules can be described as conditions, functions, logic, or a collection of tables and relationships and so on. Simple approach is to use one to one mapping: that is to compare two ontology O1, O2 in the entity, if found in the O1 body of an entity e1 i and O2 in the body of the entity e2 j implied the same meaning, it can be considered O1 and O2 in the body has established a map. Mare Ehrig gives a formal ontology mapping function.

Semantic Mapping: Its mission is based on the similarity matrix to generate the appropriate mapping rules. Mapping rules do set the concept of how to convert the source ontology to the target ontology. Mapping generally 1:1, n and m: n three. Also m: n mapping can be used m-1: n mapping to combine said. Mapping rules should include examples of the conditions to be mapped and the corresponding transfer function.
The paper has proposed a variety of conceptual similarity computing model. They can be broadly divided into two groups: continuous metric space model of set theory matching model. The latter can be further divided into: geometric (Geometric) model, conversion (Transformational) model, in particular Feature. Geometric model is calculated in the n-dimensional space physical distance between the feature vectors to determine the degree of similarity between the concepts. Conversion mode Type is converted to an entity by another entity conversion steps needed to determine the degree of similarity between the concepts.

MAFRA using semantic bridge (semantic bridge) to represent the mapping between ontology’s, MAFRA main procedures include: (1) conversion (standardized), the source ontology and target ontology into the same format; (2) similarity computation, computing concepts and the similarity between attributes, the calculation method of multi-strategy learning similarity calculation; (3) to generate semantic bridge, according to the semantic similarity calculation results bridge the semantic bridge ontology using semantic bridge (SB0), said; (4 ) the implementation phase, according to the semantics of bridge building in the source ontology instances into target ontology instances; (5) post-processing stage, the results of the implementation phase of the returned feedback[5]. PROMPT is a semi-automatic ontology mapping and merging tools based on syntactic element names and mapping methods have proposed ontology mapping, and provides a graphical user interface for users to modify and confirm the interpretation of the results ontology mapping; difference is in the QOM , the need to calculate the elements of the search space is iterative, dynamic programming method that uses iterative search space.

\[
D(M) = \sigma^4 - \frac{\sigma^4}{M (\sigma^2 + \sigma^2)} (2 \sum_{i=1}^{M} \rho(s, i) - 1) + \frac{\sigma^4}{M^2 (\sigma^2 + \sigma^2)} \sum_{i=1}^{M} \sum_{j=1}^{M} \rho(i, j)
\]  

First, domain knowledge from the learner to learn relevant knowledge, the learning process is the source ontology, target ontology and the correlation between domain knowledge to complete the mapping. This knowledge includes not only areas of the field restrictions, business rules, including the user had previously mapped the relevant ontology, ontology instances and ontology mappings between.

Input: clustering ontology k and contains a data set of n objects
Output: k ontology, so that the minimum cost ontology value
Steps: i, randomly select ontology k objects as initial cluster centers.
ii, IF nolayer [j] is not the last element in the array TNEN.
iii, according to the average cluster objects, each object is assigned to the most similar cluster.
iv, update the ontology concept average, that is calculated for each object in the object cluster average.
v, calculate the cost ontology E.

3. Formal concept analysis and concept lattices technology

Based on the lattice contribution to the theory Brikhoff, Germany Willle [6] in 1982 as a professor of mathematics first introduced the concept lattice (concept lattice), laid the theoretical basis of formal concept analysis, the concept of the philosophy of mathematics described, to achieve the concept of a formal description of the method. Concept lattice theory is the theory of formal concept analysis of
core data structures, knowledge discovery, and is considered a powerful mathematical tool for data analysis. Concept lattice is a formal description and analysis tool, due to its excellent features, the current software engineering, data mining, information retrieval and other successful applications in many fields.

Concept lattice reflects the concept lattice nodes connotation and extension of the unity of concept lattice hierarchy is a partial order between nodes relationship, that the concept of generalization and instantiation. There are a lot of research work explores the concept lattice from the grid nodes and the hierarchy to extract rules. Experimental studies have shown a variety of rules based on concept lattice extraction system is effective, formal concept analysis is very suitable for that rule-based knowledge. In knowledge discovery, the author of many cells collected from the extraction rules on the issue, but also by the direct use of lattice matching and classification node instances.

Let $K = (G, M, I)$ is a formal context, $B \leq (K) = ((G, M, I),  is the formal context of concept lattice of $K$, then $B$ is a complete lattice, for $(G, M, I)$ of any non-empty set, the least upper bound $\text{sup}(B)$ and the greatest lower bound $\text{inf}(B)$ were it.

$$\bigvee_{i \in I} (X_i, g(X_i)) = (g(f(\bigcup_{i \in I} X_i)), \bigcap_{i \in I} g(X_i))$$

$$\bigwedge_{i \in I} (X_i, g(X_i)) = (\bigcap_{i \in I} X_i, f(\bigvee_{i \in I} g(X_i)))$$

(5)

Concept lattice can be graphically represented as a labeled graph (labelled line diagram). Generated graph as follows: If $C_1 < C_2$, $C_3$, and no element grid makes $C_1 < C_3 < C_2$, then there is an edge from $C_1$ to $C_2$. Line diagram, we use the concept of a black circle indicates the form, use the line-up that sub-concept - the concept of parent relationship. For an object, if $C$ is a concept that contains the smallest object, the object's name is attached to the corresponding circle on $C$. For a feature, if $C$ is a concept that contains the characteristics of the largest, the feature's name is attached to the corresponding circle on $C$. Concept lattice of a label is often used as a communication line graph mode, which makes the concept of a given data structure of the background became clear and easy to understand, enabling visualization of concept lattices show as figure3.

![Figure 3. Concept lattice building diagram](image)

(6)

Concept lattice in the application process, you first need to solve the problem of grid construction. As the concept lattice of completeness, even for the appropriate size of the data, but also will have a large frame structure, its structure is undoubtedly a very time-consuming process. Therefore, the efficient generation of concept lattice algorithm becomes more urgent. From the constructor's point of
view, according to concept lattice construction algorithm can be divided into three categories: batch algorithms, incremental construction algorithm and parallel algorithm.

4. Application of formal concept analysis in Ontology learning and mapping

In practice, FCA and ontology as two formal methods the difference was not, they shall have emphasized: the main idea of the importance of asking the same, have stressed the need for formal specification model [8]. Difference is that the ontology 's goal is to provide a consensus to support knowledge-intensive applications, while the FCA is not a model of reality, the concept lattice is given on the basis of the data, the analysis of domain knowledge and structure, are man-made product; FCA is always dependent on a given set of objects, while the ontology is not given in the case of the data can also be set up in the FCA in the extension of the concept and content are equally important in terms of both concept and content of the ontology part is stressed.

On the one hand, FCA is a technology for ontology engineering. A concept lattice structure of the given data to extract the conceptual level, as the basis for the body applications, for manual or semi-automatic ontology, the ontology has been merged; and visualization of the body, the body of the navigation (navigation) and analysis tasks. In theory, this combination of methods from the Ontology of formal methods (including description logic, concept maps, etc.) can be used to assist the completion of FCA. According to ISO704 definition, the formal representation of the body mainly in the presentation layer and the FCA is focused on the conceptual level, so the FCA for the formal representation of the body to provide the necessary supplement, as is shown by equation7.

$$\{(A, B) \mapsto \left(\left(A \cap \hat{G}_1, B \cap \hat{M}_1\right), \left(A \cap \hat{G}_2, B \cap \hat{M}_2\right)\right)\}$$

The ontology learning process roughly divided into the following steps:

1. Firstly use natural language processing technology (NLP) to retreat the collected plain text, get words set of the text.
2. Using the method of probability and statistics, generally use the TF-IDF method get the key concept words which can represent the text, the specific method is: calculate the frequency of the concept vocabulary in text, if this frequency appears more than specified threshold, will be as the concept of the domain ontology. And then form binary relation table of words, files based on the funded vocabulary concept, combined with the corresponding text set.
3. Because of the dual relationship table exists odd values, in order to build lattice easily, must be converted into a single value binary relation table. Then build concept lattice according to the building lattice method on the single value formal context, then the constructed concept lattice is not ultimately ontology.
4. Finally, explore how to combine concept lattice and ontology learning, that is, from concept lattice converted into the ontology.

On the other hand, the application of ontology is to improve FCA applications. FCA does not contain any structure in the property that if the properties of FCA as a concept in the ontology, you can build relationships between attributes and interdependencies in order to improve the quality of FCA applications. Such as using the ontology help FCA to handle large database problems.

$$\text{RG} \subseteq G \times G: \{(g_1, g_2) | g_1, g_2 \in G \text{ Weight}(g_1) = \text{Weight}(g_2)\}$$

$$\left(M_1 \cup M_2 \right) \times M_3 = \left(M_1 \times M_2 \right) \cup \left(M_2 \times M_3 \right)$$

When the two ontology node elements have some common characteristics, then when they are defined similar. The degree of similarity can indicate similarity. Formally, the similarity calculation has met the ontology as shown by equation9.

$$\text{Sim}((E_1, I_1), (E_2, I_2)) = \left|\frac{E_1 \cap E_2}{r}\right| \times (1-w) + \left[\frac{1-w}{m} \text{max}_{p \in p(I_1, I_2)} \left(\sum_{a, b \in I} \text{as}(a, b)\right)\right] \times w$$

This can be converted into a single value in equation 8 form the background, and in the later calculation of the correlation between the concepts, only concerned about the specific values. Due to space limitations, here omitted converted single-value form the background, the background values for a single algorithm can be used to create grid Godwin, get Figure 4.
5. Conclusions

In this study, the introduction of formal concept analysis techniques to ontology of learning and mapping, you can get some help from the field of abstract concepts and relationships between concepts, and said ontology out in the concept lattice, with lattice to represent the concept of a tree is easier than understand and can be used as a guide to build a ontology, but it only provides a guide to methods, the final choice of the ontology still at the developer's needs. This paper presents to ontology learning and mapping in semantic web based on formal concept analysis. Moreover, domain ontology is constructed by a case, which has perfect hierarchical structure of concept and semantic information to retrieval. So that search efficiency is improved. Another body still requires the use of the initial natural language understanding to obtain a representative vocabulary properties and encounter complex areas, the establishment of the concept lattice corresponding to more complex, useful information will be inundated. Based on the above content, it could be in the following areas for further research.

6. References