A Statistical Amalgamation Approach for Ontologies

Peng Liu[1], Chuang Xu[1], Xiaoxuan Wang[2], Xiaoying Wang[1], Gonghua Xu[3]

[1]Command Automation Institute, PLA University of Science & Technology, Nanjing, China
[2]Department of Surveying and Geo-informatics, Tongji University, Shanghai, China
[3]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, China
gloud@126.com, xuchuang1993@163.com, wxx1012@163.com, monica.wxy@163.com, xxgghh800808@126.com

***Abstract*—As ontology is subjective and varies in different domains, the amount of ontologies turns out to be huge but with poor compatibility. Mainstream method for ontology integration is mostly achieved by establishing mappings between ontologies. In this essay, the author put forward another way of ontology merging. After statistic machine learning on concept relations, the frequency of different ontologies appeared in concept relations reveals certainty factor and help to build a large-scale concept relations network including the statistic information and domain categories, so that the conceptions conveyed by different ontologies can be fused together and the merging concept space turns to be relatively objective. And the experiments results also help to demonstrate the feasibility of the ontology merging.**

***Index Terms*—ontology;statistic;sample skewness;machine learning**

I. PREFACE

Ontology is famous for Gruber's definition as "Ontology is a defined specification of conceptual model"[1].The importance of Ontology is demonstrated in many aspects and getting generally accepted. Nowadays, Ontology is widely used in semantic web, information intelligent retrieval System, and digital library,etc. As knowledge varies in different domains, and the establishment of ontology is subjective and distributive, researchers in different fields usually build unique ontologies according to their own requirements. So the heterogeneity of ontologies is ubiquitous, even in the same field, there are many different ontologies as well as some mixed ontologies which are hard to know the exact field related. It causes the problem that there is a huge number of ontologies with poor compatibility. Currently there are over 10000 ontologies which have been indexed by Swoogle.

Ontology mismatch is the direct cause of ontology heterogeneity. Thus the explication of these mismatching elements is the basic way to solve the problem of ontology heterogeneity. These mismatching situations can be divided into two levels: the mismatch on the linguistics level and the mismatch on the model level. Kitakami[2] and Visser[3] called these two levels as non-semantic and semantic. Visser and other people subdivided semantic into two categories: conceptualization mismatch and explication mismatch[3]. The ontology heterogeneity problems which focus on linguistics level, for example: syntax error and Logical error etc., can be solved by conversion easily. Other ontology heterogeneity problems which focus on concept level are more complicated and hard to solve.

Ontology heterogeneity causes serious problems. On the one hand, it affects the information sharing and its interoperability. For example, when applying Semantic Web[4] Service, the interactions between different application systems are very common and frequent, and then the ontology heterogeneity can be a big obstruction when application systems interact on information. On the other hand, local ontology has strong territoriality and subjectivity, which means, no matter how large its quantity is, it cannot depict the whole world with an objective and complete description. In this regard, it would deviate from objectivity when we use it as application guidance. In order to solve the problem of ontology heterogeneity, we urgently need to find out a good method of ontological knowledge sharing and ontology integrating. Nowadays the related researches are more active in foreign countries, but the related concepts of ontological knowledge sharing and ontology integrating are in chaos, various methods and patterns, lacking an acknowledged definition. The related concepts include ontology integration, ontology merging, ontology alignment, ontology mapping etc., all of which can be generalized as ontology reuse(see[5][6][7]). Ontology reuse means that the existing ontology gains knowledge and reuse. Now, ontology reuse has already been widely used among Geospatial Information Systems[4][5], bioinformatics[8] and other fields.

II. RESEARCH STATUS ON ONTOLOGY REUSE

During these years, researchers all over the world have done so much to eliminate the heterogeneity mainly by the method of seeking and establishing one to one mapping. The theory study mainly focuses on integrated models and the researchers also develop many tools of ontology integrating and ontology mapping.

Pinto believes that ontology reuse has two forms[9]: ontology integration and ontology merging, and he emphasizes that they are distinguished by subject domains; while Nov and Stumme claim that the two are equal[10]. Sowa thinks ontology integration makes interaction within different ontologies to achieve

doi:10.4304/jnw.7.2.243-248

ontology alignment or merging by developing and handling mapping process. University Oldenburg put forward the concept that ontology can be classified into mapping, connection and integration[11] according to the integrating levels. Ontology merging is to develop an ontology which contains all source ontologies, while ontology connecting describes mapping collections of different ontologies[12]. Although these concepts are not yet settled, and with varied translations, the core meanings all point at ontology reuse[13][14].

Fernández-Breis and Martínez-Béjar bring up ontology integrating collaboration framework[15], whose algorithm is based on feature separation and identification between two ontologies. The development of global ontology requires the experts to handle the users with the aid of the system according to the concept, category and related terms input within the system. OISs[16] is a formalized framework brought up by Calvanese and others for ontology integration system. In the framework, ontology is written by description logic and the mapping of ontologies is expressed via proper system based on inquires and concepts of ontology can be mapped into views. Madhavan brings in a framework for ontology mapping, which can work between models written in different languages, and also suitable for ontologies without enough information. OntoMapO is a framework of accessing and integrating top-level ontology. It is actually a service of ontology mapping, which requires the ontology to be in the same form to achieve mapping by a relatively simple meta-ontology. And Kent brings in IFF for ontology structure. The framework supporting ontology sharing is based on the information flow theory by Barise and Seligman. Furthermore, the essay[17] studies ontology integration through the grammar of ontology language, points out the potential semantic mismatches may caused during ontology integration. From the point of architecture of integration, the essay[18] focuses on the 3 ways of integration of sources and checks the advantages and shortcomings. And the essay[19] introduces the concept of ontology library association and shows the ways of ontology algebra.

In the field of ontology research, machine learning is one of the vital technologies. Machine learning studies how computers can simulate or achieve human's learning behavior to acquire new knowledge and skills and rearrange the knowledge for self-improvement. Machine learning is mostly applied on the auto-construction of ontology as well as ontology integrating and mapping. Also there are some tools which have lead in machine learning. For example, Glue[20] is one of the typical tools. The multi-learning modules guided by multi-strategy learning conclude many applications of machine learning, from Naive Bayes and nearest-neighbor pattern classification to entity identification and retrieval and so on. Every module has been specially trained for different information to improve the classification accuracy and then all modules are connected to make prediction.

Existing ways to solve the heterogeneity of ontology are mostly to seeking for one to one mapping between ontologies and setting up mapping. These are mainly focused on the items and structure, so they can only see the simple relations and the results of mapping are not as good to be with wide suitability. Some researches pay attention to cases hierarchy and try to achieve mapping with the combination of machine learning and hand-classified sharing samples. But the accuracy of machine learning and the effective of sharing samples are susceptible. Some others tried to mix many methods, but the effective and the results are difficult points.

III. ONTOLOGY FUSION BASED ON STATISTICAL MACHINE LEARNING

We put forward a new idea to solve the ontology heterogeneity problem, aiming at achieving ontology reuse. This is basic on the statistical methods for machine learning. By automatically learning, it enables the ontology from different fields which meet certain specifications rapidly. Its working process is similar to the game Feeding Frenzy, the big fish (Matrix, we call it integrating concept space) continues to "eat" and "digest" the smaller fish (Daughter, that is, integrated ontology) and then gradually become bigger (as Matrix will gradually enrich its knowledge). We call this process as Ontology Fusion. Its implication includes the following three points:

- The problem of ontology fusion against ontology heterogeneity, is a process of ontology reusing and knowledge sharing.
- The purpose of ontology fusion is to integrate different ontology, which knowledge contained in, providing users with a description close to the objective concept, thereafter to achieve the ontology's reusability and interoperability in the aspects of data access.
- Ontology fusion is a process on the basis of statistical machine learning, and the matrix is of special structure with integrating concept space.

The integrating concept space is like a large concept of relationship network. It remarks every concept and conceptual relationship which has appeared in daughters. The integrating concept space can be expanding, it is empty at the beginning, but expands as the integrating concept and relationship are increasing. The sameness with ontology is they can be indicated in the diagram as several vertices (conception) and several edges which connect different vertices (relationship between conceptions); the difference is, every edge in the integration concept space, must record the statistical information on concept relation intensity. The statistical information on every edge includes the occurrence number and intensity this concept relation has appeared in every daughter, and the occurrence number and intensity it appeared in specific domain.

Statistical information for each edge can be expressed as:$\{(S,V),[(s_i,v_i)]i=1...n\}$.S represents the occurrence number this concept relation has appeared in all the daughters,V represents the relationship intensity factor calculated by machine learning on the basis of S,each (s_i,v_i) and the global situation.s_i and v_i stand for the

occurrence number and intensity the relationship appeared in the domain i.[(s_i,v_i)]i=1...n stand for the situation of each (s_i,v_i) from domain 1 to domain n.Therefore, Some concept relationships are thick,but others are fine in the integration concept space ,and moreover, we can enter into any relationship to observe the thickness of it in different domains.

Calculation of the intensity factor V based on penalty parameter

Let $F_1,F_2\ldots F_n$ be N domains,$N_1,N_2\ldots N_n$ be N samples,C represents penalty parameter,S_{ij} stands for the occurrence number the relationship i appeared in the domain j， V_{ij} stands for the intensity factor of the relationship i in the domain j.Set penalty parameter of the domain possessing the least samples be1， Let F_j be the domain possessing the least samples,and then, the value of parameter C should make the following formula hold,in which

$$C_i=\begin{cases} C_1,\ X_1 \in F_1 \\ C_2,\ X_2 \in F_2 \\ \ldots \\ C_n,\ X_n \in F_n \end{cases}$$

suitable for

$$C_i=\frac{N_j}{N_i};$$

The formula indicates the penalty in various fields is inversely proportional to the number of samples, and then, the intensity factor is to be the following:

$$V_{ij} = C_i \times S_{ij}$$

M stands for integrating concept space, and A stands for ontology to be integrated. The algorithm that the integrating concept space as the matrix integrates a daughter (to be integrated ontology) is as follows.

Algorithm.

Input: M,A.

Output: the concept relationship intensity factor V corrected by machine learning on the basis of the global situation and the change of matrix.

1. $\forall$ concept C $\in$ A;
2. if C $\notin$ M then
   creat C $\subset$ M
   end if;
3. $\forall$ relationship R $\in$ C(the other end of relationship R is concept D);
4. if D $\notin$ M then
   creat D $\subset$ M
   end if;
5. register the statistical information of relationship R $\subset$ M;
6. if the relationship related to C= $\varnothing$ then
   return step 1;
   else
   return step 3;
   end if.

IV. PERFORMANCE ANALYSIS

Five ontologies to fusion are as follows: Ont1, Ont2, Ont3, Ont4 and Ont5.

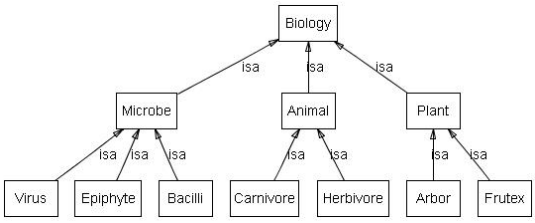

Figure 1. Ont1

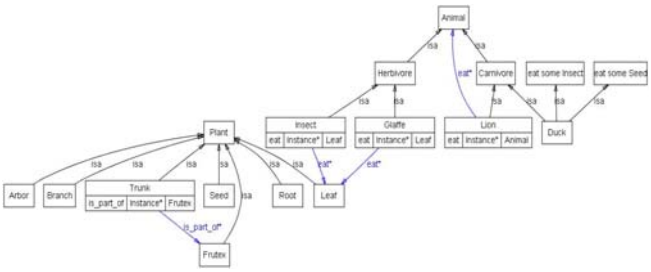

Figure 2. Ont2

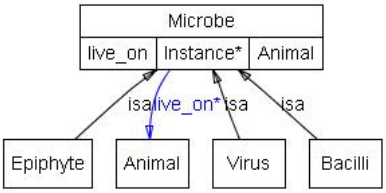

Figure 3. Ont3

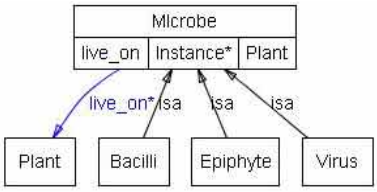

Figure 4. Ont4

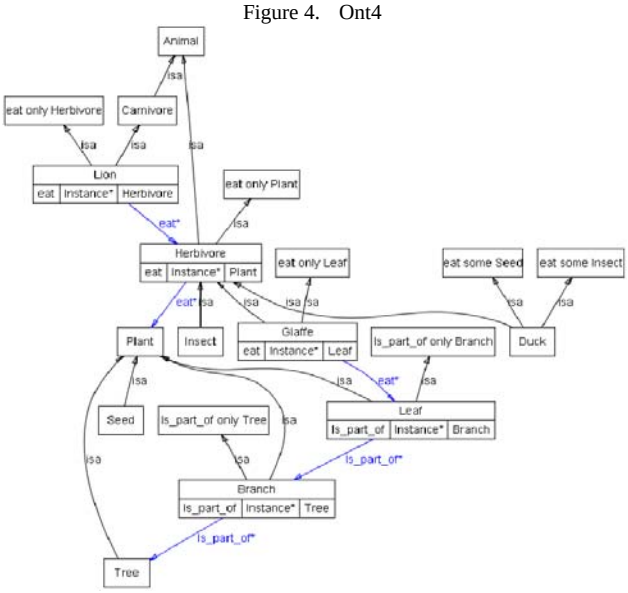

Figure 5. Ont5

The amalgamation is shown in the following.

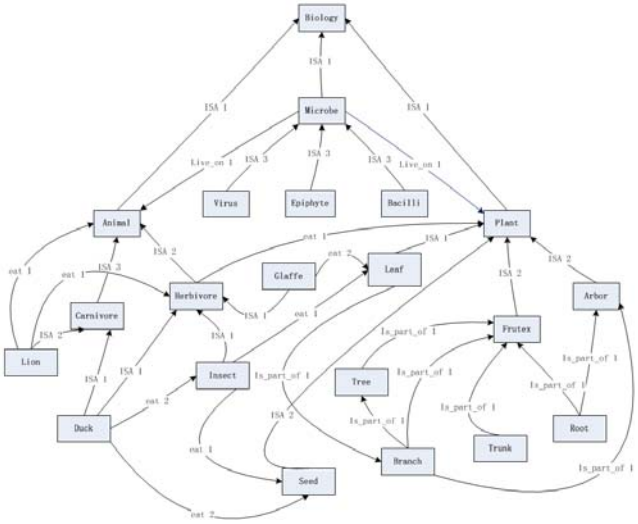

Figure 6. amalgamation of ontologies

An example of the query results is shown in Table 1.

Now we elaborate on the method for solving ontology reliability differences caused by the disproportion of ontology number.

Let's suppose that there are two concepts involved with four domains and they show four relationships. i=1,2,3,4 stands for four different domains. The number of ontology in corresponding domain is 150,2,80,20. R1, R2, R3,R4 stands for four different relationships between two concepts ,described in Table 2.

TABLE II. EXAMPLE OF QUERY RESULTS

i / relation	1	2	3	4
R1	8	0	0	0
R2	70	0	80	0
R3	0	2	0	17
R4	3	0	0	2

The Statistical information in Table 2 shows that the intensity of relationship R2 in domain 1 is 70,and the intensity of relationship R3 in domain 2 is only 2. Clearly, It is a statistical distortion typically caused by sample gap. The Statistical information can be adjusted according to calculational method of intensity factor and algorithm of ontology fusion. After adjustment the intensity information is shown in Table 3.

TABLE I. STATISTICAL INFORMATION BETWEEN TWO CONCEPTS

Class / Ont	Virus	Lion	Duck
Ont1	Subclass of Microbe	Null	Null
Ont2	Null	Subclass of Carnivore eat Animal	Subclass of Carnivore eat some Insect eat some Seed
Ont3	Subclass of Microbe Live_on Animal	Null	Null
Ont4	Subclass of Microbe Live_on Plant	Null	Null
Ont5	Null	Subclass of Carnivore eat Herbivore	Subclass of Herbivore eat some Insect eat some Seed
Amalgamation of Ont	Subclass of Microbe---3 Live_on Animal -------1 Live_on Plant --------1	Subclass of Carnivore--2 eat Animal----------1 eat Herbivore----------1	Subclass of Carnivore--1 Subclass of Herbivore-1 eat some Insect--------1 eat some Seed----------1
outcome	Subclass of Microbe Live_on Animal or Plant	Subclass of Carnivore eat Animal	Subclass of Carnivore& Herbivore eat some Insect&Seed

TABLE III. INTENSITY INFORMATION AFTER ADJUSTMENT

i / relation	1	2	3	4
R1	0.107	0	0	0
R2	0.933	0	2	0
R3	0	2	0	1.7
R4	3	0	0	0.2

The reliability of relation R1, R2, R3,R4 in the concept space is adjusted as following.

$$R1(s_i,v_i)_{i=1}=0.107,\quad R2(s_i,v_i)=\begin{cases}0.933, & i=1\\ 2, & i=3\end{cases}$$

$$R3(s_i,v_i)=\begin{cases}2, & i=2\\ 1.7, & i=4\end{cases},\quad R4(s_i,v_i)=\begin{cases}3, & i=1\\ 0.2, & i=4\end{cases}$$

It can be seen from the result data, after adjustment the statistical reliability in domain 2 increases significantly but the statistical reliability in domain 1 is just the reverse, and the reliability of other domains is also adjusted accordingly. The method is proved to be effective on solving ontology reliability differences caused by the disproportion of ontology number.

V. DISCUSSION

Features of ontology integration is firstly counting on relations of the concepts, after the integration of ontology from various sources, and finally comes to a huge concept space with stats and field information. This

method avoids conceptual triviality of heterogeneity mapping process and solves the problem by statistical method. There are several more merits: high automation level without frequent human involvement; huge amount of sourcing ontologies with more reliable results; widely suitable for inter-field integration as well as separate field information.

In this way, we can avoid the inconsistence of different ontologies on concept level generally faced by traditional ontology reuse. For the different meanings cause by different fields, our statistic study will classify it as different relations of concepts in the matrix or the same relation varies in different field. As for different ontology modules, statistic study will help to synthesis the information of granules. For those ontologies which employ various words to express same meaning, statistic study will help to build equivalent relations between the words. As for semantic conflicts, we'd better arrange consistency treatment using traditional ontology mapping first and then continue statistic study. Comparatively speaking, the consistency treatment is much easier on linguistic level than conceptual level. Because the inconsistency on linguistic level is definite, while on conceptual level, the inconsistency is much more sophisticated due to the huge volume and varied forms of concepts. In the following essay, ontology only refers to those comply with norms of OWL Lite.

The main challenge which this paper put forward on the machine learning process is skewed data of the dataset. Because the distribution of the learning samples (the ontology integration), are skewed or disequilibrium, that is, the number of samples in different areas may have difference on magnitude order. When there is data deviation, samples cannot reflect the data distribution of the entire space accurately. The corresponding measures can be divided into two aspects: On the one hand, the global optimization can be adjusted. Although the data skew can result in difference of data intensity in relationship network, the more intense parts still have sparse regions and vice versa. In the calculation of the relationship strength, the local relative frequency has a larger effect coefficient than the absolute frequency. The key is to find the right parameter in mathematical model through the machine learning to balance those various factors; On the other hand, as the domain of fields which involved in learning of ontology are getting wider and the quantity is getting greater, the degree of the skewness will be lower. In fact, the skewness of the sample set also exist among search results of Google and other search engines which uses PageRank evaluation method. Through link exchange, some Web sites can get a higher PageRank value and raise their rankings. The PageRank will be distorted because of the skewness of the sample set when a Search Engine Spiders accesses these close relative websites, however, by adjustment of the algorithm, these interference behavior can be screened from under normal circumstances. This shows the skewness of dataset has the possibility to get an adequate solution.

In the application of ontology fusion achievement, the integrating concept space has registered numerous information about conceptual relationship intensity and its domain information, so it is hard to express it with the standard OWL language. In this case, direct application of OWL becomes impossible, therefore special treatment is needed. There are mainly two ways: One is to develop a type of specialized application on integrating concept space. This application can take full advantage of the prolific conceptual relationship information and the information of relationship intensity between different fields on integration concept. Thus it has stronger semantic understandability than traditional multiplexing ontology. Another one is to develop a kind of conversion software which can transform integrating concept space into ontology. Using this software, we can select and pick up ontology that fits OWL or RDF rules. Not only can it pick-up the large ontology which already include all the concepts, but also the local ontology which only have a part of concepts as well as domain ontology of certain fields. The converted ontology is unavoidable to lose some of the rich statistics which mother has originally. But even so, its accuracy and integrity will still exceed the traditional ontology reuse in most of the cases. This is because the conversion operation itself can use statistics to make a more rational choice, and since the ontology integration is completely automatic, it can provide much more samples of ontology learning than the traditional methods (typically, human-machine interactive methods).

REFERENCES

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

[2] H Kitakami, Y Mori, and M Arikawa,"An intelligent system for integrating autonomous nomenclature databases in semantic heterogeneity,"Database and Expert System Applications, DEXA' 96, number 1134 in Lecture Notes in Computer Science, pp. 187 − 196, Zurich, Switzerland,1996.

[3] P R S Visser, D M Jones, and T J M Bench-Capon,"Shave MJR.An Analysis of Ontology Mismatches. Heterogeneity versus Interoperability,"Spring Symposium on Ontological Engineering (AAAI 1997).

[4] T B Lee, H Hendler, and O Lassila,"The semantic Web,"Scientific American , 284 (5) , pp.34-43,2001.

[5] S Z Fan and S Z Li,"International Conference on Interoperability for Enterprise Software and Applications,"China,IESA '09, April, 2009. Beijing.

[6] C G Bernardo, H Ian, K Yevgeny, and S Ulrike,"Modular Reuse of Ontologies: Theory and Practice,"Journal of Artificial Intelligence Research, 2008,31: pp.273-318.

[7] A Alsayed, S Eike, and S. A Gunter,"Sequence-based Ontology Matching Approach," Proceedings of The 18th European Conference on Artificial Intelligence , Workshop on Contexts and Ontologies, July 21, 2008. Patras, Greece.

[8] B Bmc, J X Wei, H D Man, and M Barbara,"BioMed Central Proceedings Open Biomedical Ontology-based Medline exploration," BMC Bioinformatics 2009, 10(Suppl 5):S6.

[9] H S Pinto, A G Perez, and J P Martins,“Some Issues on Ontology Integration,” Proceedings of IJCAI99’s Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends, 1999, 7.1−7.12.

[10] G Stumme and A Maedche,“FCA-Merge: bottom-up merging of ontologies,”Proceedings of IJCAI,2001,Seattle,USA,2001,pp.225-230.

[11] A Sven, H Liane, and H Axel,“Identifying ontology integration methods and their app- licability in the context of product classification and knowledge integration tasks,”Report No.WI-OL-TR20122005. Oldenburg: Department of Business Information Systems, University of Oldenburg, Germany, 2005.

[12] N F Noy and M Musen,“PROMPT: Algorithm and tool for automated ontology merging and alignment,”Proceedings of the 17th National Conference on Artificial Intelligence (AAAI2000), Austin Texas, USA, 2000.

[13] Y Kalfoglou and M Schorelmmer,“Ontology Mapping: The State of the Art,”The Knowledge Engineering Review,2003,18(1), pp.1-31.

[14] H S Pinto and J P Martins,“A Methodology for Ontology Integration,” Proceedings of the International Conference on Knowledge Capture, Technical papers, ACM Press,2001, pp.131-138.

[15] J T Fernández-Breis and R Martínez-Béjar,“A Cooperative Framework for Integrating Ontologies,”International Journal of Human-Computer Studies, 2002,56(6) , pp.662-717.

[16] D Calvanese and G D Giacomo,“Lenzerini M. Ontology of integration and integration of ontologies,”Proceedings of the 2001 Description Logic Workshop (DL 2001), 2001, pp.10-19.

[17] M Klein,“Combining and Relating Ontologies: An Analysis of Problems and Solutions,” Workshop on ontologies and information sharing, IJCAI, Seattle, WA, 2001, pp.53-62.

[18] H Wache and T Vgele,“Ontology-based Integration of Information-a Survey of Existing Approaches,” Proceedings of the Workshop Ontologies and Information Sharing, 2001.

[19] M Prasenjit,“A Graph-oriented Model for Articulation of Ontology Interdependencies,”Proceedings of the International Conference on Extending Database Technology (EDBT),Springer-Verlag,2000, pp.86-100.

[20] A Doan, J Madhavan, P Domingos and A Y Halevy,“Ontology Matching: A Machine Learning Approach,”Handbook on Ontologies in Information Systems, 2004.