Journal of Southeast University (English Edition) Vol. 23, No. 3, pp. 385 – 388 Sept. 2007 ISSN 1003—7985

Kullback-Leibler distance based concepts mapping between web ontologies

Wu Suyan Guo Qiao

(School of Information Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: A Kullback-Leibler(KL) distance based algorithm is presented to find the matches between concepts from different ontologies. First, each concept is represented as a specific probability distribution which is estimated from its own instances. Then, the similarity of two concepts from different ontologies is measured by the KL distance between the corresponding distributions. Finally, the concept-mapping relationship between different ontologies is obtained. Compared with other traditional instance-based algorithms, the computing complexity of the proposed algorithm is largely reduced. Moreover, because it proposes different estimation and smoothing methods of the concept distribution for different data types, it is suitable for various concepts mapping with different data types. The experimental results on real-world ontology mapping illustrate the effectiveness of the proposed algorithm.

Key words: semantic web; ontology mapping; Kullback-Leibler distance

With the growing access to heterogeneous and independent data repositories, the treatment of differences in the structure and semantics of the data stored in those repositories plays a major role in information systems[1]. Ontology mapping is an effective method to realize the interoperation of heterogeneous ontologies[2].

Determining the semantic similarity of concepts from different ontologies is the core of ontology mapping [3]. Clearly, many different definitions of similarity are possible, each being appropriate for certain situations. Examples include Cupid[4], COMA[5], S-Match[6], GLUE[7-8]. GLUE uses machine learning techniques based on instances to find equivalent concepts between two ontologies. The joint probability distribution of a concept is chosen to compute the similarity of all kinds of semantic relationships among different ontology concepts. According to the overlapping of sample spaces, the computation formula of the "equal" relation similarity is

$$\mathrm{sim}(A, B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{p(A, B)}{P(A, \bar{B}) + p(A, B) + p(\bar{A}, B)} \tag{1}$$

The computation of all kinds of semantic relationships among ontology concepts can be transformed to the computation of the four probability values: $p(A, B)$, $p(A, \bar{B})$, $p(\bar{A}, B)$, $p(\bar{A}, \bar{B})$. The greatest challenge of GLUE is that of computing the joint distribution of any two given concepts A and B.

GLUE addresses this problem using machine learning techniques as follows: It uses the frequencies of words in the content and the name of concept A's instances to learn a classifier C_A and then classifies instances of concept B by the classifier C_A. In such a way, a classifier C_B can be learned and then classifies instances of concept A. GLUE uses all the results to calculate the four probability values: $p(A, B)$, $p(A, \bar{B})$, $p(\bar{A}, B)$ and $p(\bar{A}, \bar{B})$.

GLUE has the disadvantages of high complexity and is only suitable for computing the instances in which data type is text. In this paper, the Kullback-Leibler (KL) distance is proposed to compute the similarity between the distributions of two concept instances. Compared with traditional algorithms, the computing complexity of our algorithm is largely reduced and it is suitable for various data types besides text.

1 KL Distance Based Concept Mapping

For the remainder of this paper, let O_1 and O_2 be two ontologies, A is the set of concepts in ontology O_1, marked $A = \{A_i \mid i = 1, \ldots, |A|\}$; B is the set of concepts in ontology O_2, marked $B = \{B_i \mid i = 1, \ldots, |B|\}$, where $|A|$ and $|B|$ are the concept numbers of A and B, respectively. X_i and Y_i are the instance sets of A_i and B_i, respectively, marked $X_i = \{x_{ij} \mid j = 1, \ldots, |X_i|\}$, $Y_i = \{y_{ij} \mid j = 1, \ldots, |Y_i|\}$, where $|X_i|$ and $|Y_i|$ are the instance numbers of A_i and B_i, respectively.

Given enough instances, if A_i and B_i are similar or equal, then the distributions of X_i and Y_i should be close or equal. The Kullback-Leibler distance is a widely-used measure in statistics which depicts similarity or "closeness" between two probability distributions[9]. It has been applied to natural language pro-

Received 2007-05-18.

Foundation item: Foundation of Next Generation Internet of China.

Biographies: Wu Suyan (1977—), female, graduate; Guo Qiao (corresponding author), female, professor, guoqiao@bit.edu.cn.

cessing, machine learning, and statistical physics[10]. In this paper, the KL distance is used to measure the similarity between the distributions of two concept instances.

1.1 KL distance

In this section, we consider the theoretical and computational properties of the relative entropy measuring the "similarity" between distributions.

Definition 1 Let X, Y be two discrete random variables with a value space χ and a probability mass function $p(x) = P\{X = x\}, x \in \chi, q(y) = P\{Y = y\}, y \in \chi$. The relative entropy or the KL distance between two probabilities mass functions $p(x)$ and $q(y)$ is defined as

$$D(p \| q) = \sum_{x \in P} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)} \quad (2)$$

In the above definition, we use the convention that $0\log(0/0) = 0$ and the convention (based on continuity arguments) that $0\log(0/q) = 0$ and $p\log(p/0) = \infty$. Thus there is any symbol $x \in \chi$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p \| q) = \infty$.

The KL divergence is a standard information theoretic "measure" of the dissimilarity between two probabilities mass functions. It is not a metric in the technical sense, since it is not symmetric and does not obey the triangle inequality.

It is not difficult to prove that the KL measure has the following properties:

① $D(p \| q) \geqslant 0$;

② $D(p \| q) = 0 \Leftrightarrow p = q$.

While there are many theoretical reasons justifying the use of the KL divergence, there is a problem with employing it in practice. Recall that for distributions p and q, $D(p \| q)$ is infinite if there is some $y' \in y$ such that $p(y') = 0$ but $q(y')$ is nonzero. If we know p and q exactly, then this is sensible, since the value y' allows us to distinguish between p and q with absolute confidence. However, often it is the case that we only have estimates $\hat{p}$ and $\hat{q}$ for p and q. If we are not careful with our estimates, then we may erroneously set $\hat{q}(y)$ to zero for some y for which $q(y) > 0$, with the effect that $D(\hat{p} \| \hat{q})$ can be infinite when $D(p \| q)$ is not. There are several ways around this problem. One is to use smoothed estimates, as described in section 1.2.

1.2 Estimating concept distribution

Each concept is associated with a data type. This paper classifies data types as number, enumeration and text. Usually, no assumptions are made about the relationships among these concepts with different data types.

• Number To estimate the distribution probability of a concept with a number data type, the multi-interval quantization is used. $X = \{x_i \mid i = 1, \ldots, |X|\}$ denotes the instance set of a concept. Let $a = \min(x_i)$, $b = \max(x_i)$. We partition the close interval $[a, b]$ into N equal-width subintervals denoted as Δ_k, $k = 1, 2, \ldots, N$, and then the frequency f_k is the instance count in Δ_k. The probability of an instance in subinterval Δ_k is defined as

$$p(x_{\Delta k}) = f_k \Big/ \sum_{i=1}^{N} f_i \quad (3)$$

In order to solve the problem of $p\log(p/0) = \infty$ in section 1.1, we must have a preprocessing step to overcome $q = 0$. This paper uses smoothed estimates, since they smooth over zeroes in distributions[11]. This technique is the interpolation method that produces an estimate frequency by linearly interpolating for the zeroes in distributions.

If the frequency of instances is zero in subinterval Δ_k, the estimate frequency f_k is defined as

$$f_k = \frac{f_{k-1} + f_j}{j - k + 1} \quad (4)$$

where f_{k-1} is the frequency in Δ_{k-1} and f_j is the first nonzero frequency of a subinterval following the subinterval Δ_k.

This method guarantees the probability of instances to be nonzero in each subinterval, then it is successful in solving the problem that the KL distance is infinite.

According to Eq. (3), we generate the modified instances distribution vector $\boldsymbol{P}(X') = \{p(x'_{\Delta_1}), p(x'_{\Delta_2}), \ldots, p(x'_{\Delta_N})\}$.

• Enumeration For enumeration, the interval number N is equal to the size of the enumerative data type. Then figure out the frequency of instances and sort them in an ascending order in intervals. According to Eq. (3), we can generate the probability distribution of the concepts with enumerated data types. The smoothed estimates method of enumerative data types is the same as that of the text data type, described as in the following context.

• Text For text data type, we used a 1-gram language model for content-based text. This algorithm expresses all the instances of a concept as content lists[12]:

$$\mathrm{CL}_c = \{(w_1, f_1), (w_2, f_2), \ldots, (w_n, f_n)\} \quad (5)$$

where w_i is a word that appears in concept c instances and f_i is the frequency of w_i.

Select the N frequency f_i with maximal values to generate the instances distribution vector of concept c: $\boldsymbol{P}_c(W) = \{p_c(w_1), p_c(w_2), \ldots, p_c(w_N)\}$, $p_c(w_i)$ can be figured out by Eq. (3).

Sometimes a word w_i in concept A_i does not appear in concept B_i. In order to overcome the drawback that the KL distance is infinite, the estimate frequency

$f'_{B_i}(w_i)$ of w_i in B_i is defined as

$$f'_{B_i}(w_i) = \alpha \min f_{B_i}(w_j) \quad (6)$$

where $0 < \alpha \leqslant 1$ is an adjustable parameter.

1.3 Concept mapping

The KL distance is a measure of the distance between two distributions. It will reach the minimum when the two distributions being compared are maximally similar.

To select the concept B_i in ontology O_2, which is the "closest" to a concept A_0 in ontology O_1, the computing equation is shown as follows:

$$B^* = \arg \min_i D(p(A_0) \| p(B_i)) \quad (7)$$

In practice, concept A_0 might have no match with any concept in ontology O_2. So we specify an appropriate threshold max_KL. The KL distance any acceptable match must be less than max_KL.

2 Experimental Results and Analysis

In this section, the experiments to determine concept similarity using two real world ontologies are introduced. Two movie ontologies are from two different BBS organizations of Beijing Institute of Technology and constructed independently. Fig. 1 shows parts of the two ontologies. Arrows in the figure denote the relationships among concepts.

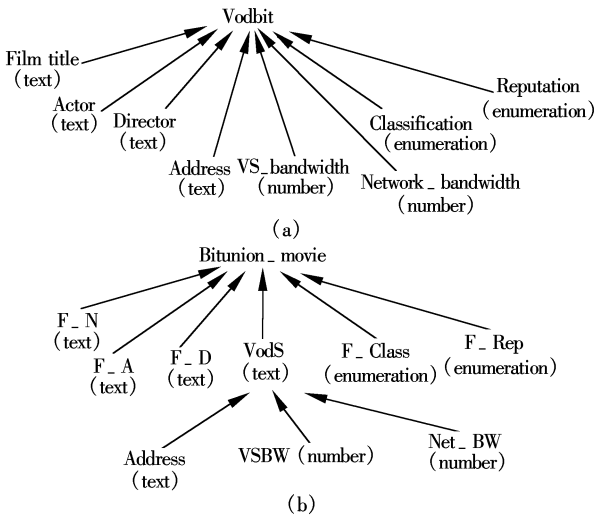

Fig. 1 Part of two movie ontologies. (a) Part of movie 1 ontology; (b) Part of movie 2 ontology

We extracted 500 instances for each of the ontologies at random, and performed some trivial data cleaning such as removing the instances with too big or too small values. We also removed the size of instances less than 130 bytes because they tend to be empty or vacuous and thus do not contribute to the matching process.

First, we will illuminate the match between concepts with number data type. Two pairs of concepts (VS_bandwidth and VSBW, Network_bandwidth and Net_BW) are used. In this part, the count of subintervals is $N = 10$. Fig. 2 shows their distributions and Tab. 1 shows the KL distance of these distributions.

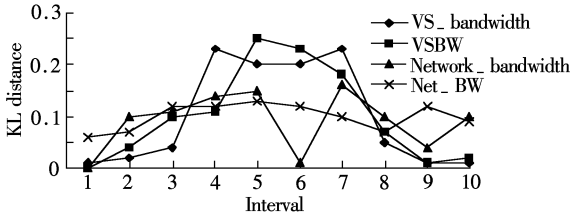

Fig. 2 Distributions of concepts with number data type

Tab. 1 The KL distance of concepts with number data type

Movie 1	Movie 2	
	VSBW	Net_BW
VS_bandwidth	0.104 1	0.379 0
Network_bandwidth	0.265 0	0.117 6

From Fig. 2, it is obvious that the distribution of VS_bandwidth is close to that of VSBW and the distribution of Network_bandwidth is close to that of Net_BW, which is reflected correctly in Tab. 2. This shows that the KL distance is very effective.

Secondly, we illuminate the match among concepts with enumeration data types. Fig. 3 is their instance probability distribution and Tab. 2 is the computed KL distance of these distributions. The mapped pairs in concepts are marked by blacking their KL distance.

Tab. 2 The KL distance of concepts with enumeration data type

Movie 1	Movie 2	
	F_Class	F_Rep
Classification	0.022 36	0.212
Reputation	0.286	0.004 7

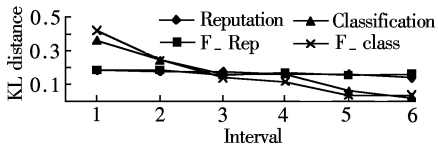

Fig. 3 Distributions of concepts with enumeration data type

From Fig. 3, it is obvious that the distribution of the classification is close to that of F_Class and the distribution of the reputation is close to that of F_Rep, which is reflected correctly in Tab. 2. This shows the KL distance is very effective.

Thirdly, we illuminate the match among concepts with text data types. The detailed matching of concepts with text data types is shown in Tab. 3. Each column in the figure shows the KL distance which is computed based on Eq. (7).

Generally speaking, the matching quality greatly depends on the data distribution characteristics of a data source, which is also illuminated by our experimentation data in Tab. 3. It is well-known that words ap-

pearing in the film title are more random than in the actor name, which results in the probability distribution of film title being more ruleless than of the actor name. So the KL(filetitle‖F_N) is greater than the KL (actor‖F_A).

Tab. 3 The KL distance of concepts with text data type

Movie 1	Movie 2		
	F_N	F_A	F_D
Filmtitle	0.134 0	1.528 6	1.678 9
Actor	1.432 5	0.093 1	0.258 4
Director	1.563 2	0.156 4	0.015 6

3 Conclusion

Solutions that try to provide some automatic support for ontology matching have received steady attention in recent research. This paper introduces a KL distance method. It computes the KL distance between the distributions of two concept instances to fulfill concept matching. Compared with the additional algorithm based on instances, the computing complexity of our method is largely reduced and it can fit various data types. The experimental results with real-world ontology mapping illustrate the effectiveness of our algorithm.

Future work will focus on the method of combining our method with other matching algorithms which include approaches based on syntax, definition and hierarchy according to the ontology model. Aside from the above work, other future research involves extending our techniques to handle more sophisticated mappings among ontologies, such as one to more and more to one.

References

[1] Berners-Lee T, Hendler J, Lassila O. The semantic web [EB/OL]. (2001-05-17) [2006-03-15]. http://www-personal.si.umich.edu/~rfrost/courses/SI110/readings/In_Out_and_Beyond/Semantic_Web.pdf.

[2] Xu Youzhi, Shen Jie, Chen Zhimin. Ontology-based information retrieval of web services in virtual enterprise [C]// *Proceedings of the IEEE International Conference on Services Computing*. Shanghai, China, 2004: 441 - 444.

[3] Egenhofer M J, Rodriguez M A. Determining semantic similarity among entity classes from different ontologies [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2003, **15**(2): 442 - 456.

[4] Madhavan Jayant, Bernstein Philip A, Rahm Erhard. Generic schema matching with cupid [C]// *Proceedings of the 27th International Conference on Very Large Data Bases*. Roma: Morgan Kaufmann, 2001: 49 - 58.

[5] Do Hong-Hai, Rahm Erhard. COMA—a system for flexible combination of schema matching [C]// *Proceedings of the 28th VLDB Conference*. Hong Kong, China, 2002: 610 - 621.

[6] Fausto Giunchiglia, Pavel Shvaiko, Mikalai Yatskevich. S-Match: an algorithm and an implementation of semantic matching [C]// *Proceedings of the First European Semantic Web Symposium*. Trento, Italy, 2004: 61 - 75.

[7] Doan A, Madhaven J, Dhamankar R, et al. Learning to match ontologies on the semantic web[J]. *The International Journal on Very Large Data Bases*, 2003, **12**(4): 303 - 319.

[8] Doan A H, Madhavan J, Domingos P, et al. Ontologies matching: a machine learning approach [C]// *Handbook on Ontologies in Information Systems*. Springer-Verlag, 2003: 397 - 416.

[9] Cover T M, Thomas J A. *Elements of information theory* [M]. John Wiley and Sons, 1991.

[10] Cui Xiaodong, Alwan A. Efficient adaptation text design based on the Kullback-Leibler measure[C]// *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. (ICASSP' 02). Orlando, FL, USA, 2002: 613 - 616.

[11] Xu Zhiming, Wang Xiaolong, Guan Yi. The data smooth technology of *N*-gram language models [J]. *Application Research of Computers*, 1999, **16**(7): 37 - 39.

[12] Sun Shuang, Zhang Yong. Clustering method based on semantic similarity[J]. *Journal of Nanjing University of Aeronautics and Astronautics*, 2006, **38**(6): 712 - 716.

语义网中基于 Kullback-Leibler 距离的本体映射方法

吴素研　　郭　巧

(北京理工大学信息科学技术学院,北京 100081)

摘要:提出了一种基于 Kullback-Leibler(KL)距离的本体映射方法. 该方法将本体中每个概念抽象为一个概率分布,并通过相应的实例数据对其进行估计;对于不同本体的 2 个概念,通过计算相应概率分布之间的 KL 距离而求得其相似度. 进而求得本体间概念的映射关系. 该方法与传统的方法相比,极大地降低了计算的复杂度,并且此算法针对不同的数据类型提出了不同的概念分布的估计和平滑方法,所以能够适用于各种数值类型的概念映射. 通过试验,证明了此方法的有效性.

关键词:语义网;本体映射;Kullback-Leibler 距离

中图分类号:TP393