Combining Sum-Product Network and Noisy-Or Model for Ontology Matching

Weizhuo Li

Institute of Mathematics,Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing, P. R. China
liweizhuo@amss.ac.cn

Abstract. Ontology matching is the key challenge to achieve semantic interoperability in building the Semantic Web. We present an alternative probabilistic scheme, called GMap, which combines the sum-product network and the noisy-or model. More precisely, we employ the sum-product network to encode the similarities based on individuals and disjointness axioms across ontologies and calculate the contributions by the maximum a posterior inference. The noisy-or model is used to encode the probabilistic matching rules, which are independent of each other as well as the value calculated by the sum-product network. Experiments show that GMap is competitive with many OAEI top-ranked systems. Futhermore, GMap, benefited from these two graphical models, can keep inference tractable in the whole matching process.

1 Introduction

Ontology matching is the process of finding relationships or correspondences between entities of different ontologies[5]. Many efforts have been conducted to automate the discovery in this process, e.g., incorporating more elaborate approaches including scaling strategies[3, 6], ontology repair techniques to ensure the alignment coherence[8], employing machine learning techniques[4], using external resources to increase the available knowledge for matching[2] and utilizing probabilistic graphical models to describe the related entities[1, 10, 11].

In this paper, we propose an alternative probabilistic schema, called GMap, based on two special graphical models—sum-product network(SPN) and noisy-or model. SPN is a directed acyclic graph with variables as leaves, sums and products as internal nodes, and weighted edges[12]. As it can keep inference tractable and describe the context-specific independence[12], we employ it to encode the similarities based on individuals and disjointness axioms and calculate the contributions by the maximum a posterior inference. Noisy-or model is a special kind of Bayesian Network[9]. When the factors are independent of each other, it is more suitable than other graphical models, specially in the inference efficiency[9]. Hence, we utilize it to encode the probabilistic matching rules. Thanks to the tractable inference of these special graphical models, GMap can keep inference tractable in the whole matching process. To evaluate GMap, we adopt the data sets from OAEI ontology matching campaign. Experimental results indicate that GMap is competitive with many OAEI top-ranked systems.

2 Methods

In this section, we briefly introduce our approach. Given two ontologies O_1 and O_2, we calculate the lexical similarity based on edit-distance, external lexicons and TFIDF[5]. Then, we employ SPN to encode the similarities based on individuals and disjointness axioms and calculate the contributions. After that, we utilize the noisy-or model to encode the probabilistic matching rules and the value calculated by SPN. With one-to-one constraint and crisscross strategy in the refine module, GMap obtains initial matches. The whole matching procedure is iterative. If it does not produce new matches, the matching is terminated.

2.1 Using SPN to encode individuals and disjointness axioms

In open world assumption, individuals or disjointness axioms are missing at times. Therefore, we define a special assignment—"*Unknown*" for the similarities based on these individuals and disjointness axioms.

For the similarity based on individuals, we employ the string equivalent to judge the equality of them. When we calculate the similarity of concepts based on individuals across ontologies, we regard individuals of each concept as a set and use Ochiai coefficient[1] to measure the value. We use a boundary t to divide the value into three assignments(i.e., 1, 0 and *Unknown*). Assignment 1(or 0) means that the pair matches(or mismatches). If the value ranges between 0 and t or the individuals of one concept are missing, the assignment is *Unknown*.

For the similarity based on disjointness axioms, we utilize these axioms and subsumption relations within ontologies and define some rules to determine its value. For example, x_1, y_1 and x_2 are concepts that come from O_1 and O_2. If x_1 matches x_2 and x_1 is disjoint with y_1, then y_1 is disjoint with x_2. The similarity also have three assignments. Assignment 1(or 0) means the pair mismatches(or overlaps). Otherwise, the similarity based on disjointness axioms is *Unknown*.

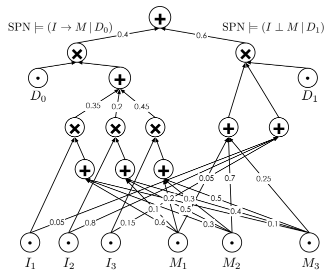

Fig. 1: The designed sum-product network

As shown in Figure 1, we designed a sum-product network S to encode above similarities and calculate the contributions, where M represents the contributions and leaves M_1, M_2, M_3 are indicators that comprise the assignments of M. All the indicators are binary-value. $M_1 = 1$(or $M_2 = 1$) means that the contributions are positive(or negative). If $M_3 = 1$, the contributions

[1] https://en.wikipedia.org/wiki/Cosine_similarity

are *Unknown*. Leaves I_1, I_2, I_3, D_0, D_1 are also binary-value indicators that correspond to the assignments of similarities based on individuals(I) and disjointness axioms(D). The concrete assignment metrics are listed in Table 1–2.

Table 1: Metric for Similarity D

Assignments	Indicators
$D = 1$	$D_0 = 0, D_1 = 1$
$D = 0$	$D_0 = 1, D_1 = 0$
$D = Unknown$	$D_0 = 1, D_1 = 1$

Table 2: Metric for Similarity I

Assignments	Indicators
$I = 1$	$I_1 = 1, I_2 = 0, I_3 = 0$
$I = 0$	$I_1 = 0, I_2 = 1, I_3 = 0$
$I = Unknown$	$I_1 = 0, I_2 = 0, I_3 = 1$

With the maximum a posterior(MAP) inference in SPN[12], we can obtain the contributions M. As the network S is complete and decomposable, the inference in S can be computed in time linear in the number of edges[7].

2.2 Using Noisy-Or model to encode probabilistic matching rules

We utilize probabilistic matching rules to describe the influences among the related pairs across ontologies and some of rules are listed in Table 3.

Table 3: The probabilistic matching rules among the related pairs

ID	Category	Probabilistic matching rules
R_1	class	two classes probably match if their fathers match
R_2	class	two classes probably match if their children match
R_3	class	two classes probably match if their siblings match
R_4	class	two classes about domain probably match if related objectproperties match and range of these property match
R_5	class	two classes about range probably match if related objectproperties match and domain of these properties match
R_6	class	two classes about domain probably match if related dataproperties match and value of these properties match

When we focus on calculating the matching probability of one pair, the matching rules are independent of each other as well as the value calculated by SPN. Therefore, we utilize the noisy-or model to encode them.

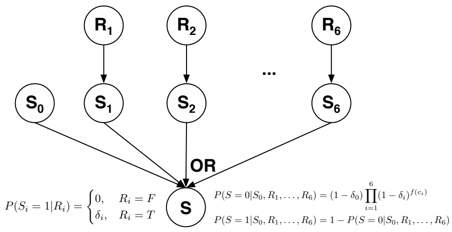

Fig. 2: The network structure of noisy-or model designed in GMap

Figure 2 shows the designed network, where R_i corresponds to the ith rule and S_i is the conditional probability depended on the condition of R_i. S_0 represents the SPN-based similarity that is a leak probability[9]. The matching probability of one pair, $P(S = 1|S_0, R_1, ..., R_6)$, is calculated according to the formulas in the lower-right corner. c_i is the count of satisfied R_i and sigmoid function $f(c_i)$ is used to limit the upper bound of contribution of R_i. As the inference in the noisy-or model can be computed in time linear in size of nodes[9], GMap can keep inference tractable in the whole matching process.

3 Evaluation

To evaluate our approach[2], we adopt three tracks(i.e., Benchmark, Conference and Anatomy) from OAEI ontology matching campaign in 2014[3].

3.1 Comparing against the OAEI top-ranked systems

Table 4 shows a comparison of the matching quality of GMap and other OAEI top-ranked systems, which indicates that GMap is competitive with these promising existent systems. For Anatomy track, GMap does not concentrate on language techniques and it emphasizes one-to-one constraint. Both of them may cause a low alignment quality. In addition, all the top-ranked systems employ alignment debugging techniques, which is helpful to improve the quality of alignment. However, we do not employ these techniques in the current version.

Table 4: The comparison of GMap with the OAEI top-ranked systems

	Benchmark(Biblio)			Conference			Anatomy		
System	P	R	F	P	R	F	P	R	F
AML	**0.92**	0.4	0.55	**0.85**	0.64	**0.73**	**0.956**	**0.932**	**0.944**
LogMap	0.39	0.4	0.39	0.8	0.59	0.68	0.918	0.846	0.881
XMAP	1	0.4	0.57	0.87	0.49	0.63	0.94	0.85	0.893
CODI	n/a	n/a	n/a	0.74	0.57	0.64	0.967	0.827	0.891
GMap	0.63	**0.57**	**0.60**	0.67	**0.66**	0.66	0.930	0.802	0.862

3.2 Evaluating the contributions of these two graphical models

We separate SPN and the noisy-or model from GMap and evaluate their contributions respectively. As listed in Table 5, SPN is suitable to the matching task that the linguistic levels across ontologies are different and both of ontologies use same individuals to describe the concepts such as Biblio(201–210) in Benchmark track. Thanks to the contributions of individuals and disjointness axioms, SPN can improve the precision of GMap. When the structure information is very rich across the ontologies, the noisy-or model is able to discover some hidden matches with the existing matches and improve the recall such as in Anatomy track. However, if the ontology does not contain above features such as in Conference track, the improvement is not evident. Nevertheless, thanks to the complementary of these two graphical models to some extent, combining the sum-product network and the noisy-or model can improve the alignment quality as a whole.

[2] The software and results are available at https://github.com/liweizhuo001/GMap.
[3] http://oaei.ontologymatching.org/2014/

Table 5: The contributions of the sum-product network and the noisy-or model

	Biblio(201-210)			Conference			Anatomy		
System	P	R	F	P	R	F	P	R	F
string equivalent	0.680	0.402	0.505	0.8	0.43	0.56	**0.997**	0.622	0.766
lexical similarity(ls)	0.767	0.682	0.722	0.666	0.657	0.661	0.929	0.752	0.831
ls+spn	0.776	0.685	0.728	0.667	0.657	0.661	0.930	0.752	0.832
ls+noisy-or	0.782	0.701	0.739	**0.667**	**0.660**	**0.663**	0.937	0.772	0.847
ls+spn+noisy-or	**0.794**	**0.703**	**0.746**	**0.667**	**0.660**	**0.663**	0.930	**0.803**	**0.862**

4 Conclusion and Future Work

We have presented GMap, which is suitable for the matching task that many individuals and disjointness axioms are declared or the structure information is very rich. However, it still has a lot of room for improvement. For example, language techniques is essential to improve the quality of initial matches. In addition, dealing with alignment incoherent is also one of our future works.

Acknowledgments. This work was supported by the Natural Science Foundation of China (No. 61232015). Many thanks to Songmao Zhang, Qilin Sun and Yuanyuan Wang for their helpful discussion on the design and implementation of the GMap.

References

1. Albagli, S., Ben-Eliyahu-Zohary, R., Shimony, S.E.: Markov network based ontology matching. Journal of Computer and System Sciences 78(1), 105–118 (2012)
2. Zhang, S., Bodenreider, O.: Experience in aligning anatomical ontologies. International journal on Semantic Web and information systems 3(2), 1–26 (2007)
3. Djeddi, W.E, Khadir, M.T.: XMAP: a novel structural approach for alignment of OWL-full ontologies. In: Proc. of Machine and Web Intelligence(ICMWI). pp. 368–373 (2010)
4. Doan, A.H, Madhavan, J., Dhamankar, R., et al.: Learning to match ontologies on the semantic web. The VLDB Journal 12(4), 303–319 (2003)
5. Euzenat, J., Shvaiko, P.: Ontology Matching(2nd Edition). Springer (2013)
6. Faria, D., Pesquita, C., Santos, E., et al.: The agreementmakerlight ontology matching system In: 2013 OTM Conferences. pp. 527–541 (2013)
7. Gens, R., Pedro, D.: Learning the structure of sum-product networks. In: Proc. of International Conference on Machine Learning(ICML). pp. 873–880 (2013)
8. Jimenez-Ruiz, E., Grau, B.C.: LogMap: Logic-based and scalable ontology matching. In: Proc. of International Semantic Web Conference(ISWC). pp.273–288 (2011)
9. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT press (2009)
10. Mitra, P., Noy, N.F., Jaiswal, A.R. OMEN: A probabilistic ontology mapping tool. In: Proc. of International Semantic Web Conference(ISWC). pp. 537–547 (2005)
11. Niepert, M., Noessner, J., Meilicke, C., Stuckenschmidt, H.: Probabilistic-logical web data integration. Reasoning Web. pp. 504–533 (2011)
12. Poon, H., Domingos, P.: Sum-product networks: A new deep architecture. In: Proc of International Conference on Computer Vision Workshops(ICCV Workshops). pp. 689–690 (2011)