GMap: Results for OAEI 2015

Weizhuo Li and Qilin Sun

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

Abstract. GMap is an alternative probabilistic scheme for ontology matching, 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. The noisy-or model is utilized to encode the probabilistic matching rules, which describe the influences among entity pairs across ontologies. In this paper, we briefly introduce GMap and its results of four tracks (i.e.,Benchmark, Conference, Anatomy and Ontology Alignment for Query Answering) on OAEI 2015.

1 Presentation of the system

1.1 State, purpose, general statement

The state of the art approaches have utilized probabilistic graphical models [5] for ontology matching such as OMEN [7], iMatch [1] and CODI [8]. However, few of them can keep inference tractable and ensure no loss in inference accuracy. In this paper, we propose an alternative probabilistic scheme, called GMap, combining the sum-product network (SPN) and the noisy-or model [6]. Except for the tractable inference, these two graphical models have some inherent advantages for ontology matching. For SPN, even if the knowledge such as individuals or disjointness axioms is missing, SPN can also calculate their contributions by the maximum a posterior (MAP) inference. For the noisy-or model, it is a reasonable approximation for incorporating probabilistic matching rules to describe the influences among entity pairs.

Figure 1 shows the sketch of GMap. Given two ontologies O_1 and O_2, we calculate the lexical similarity based on edit-distance, external lexicons and TFIDF [3] with the max strategy. Then, we employ SPN to encode the similarities based on individuals and disjointness axioms and calculate the contribution through MAP inference. 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 there is no additional matches identified, the matching is terminated.

1.2 Specific techniques used

The similarities based on individuals and disjointness axioms In open world assumption, individuals or disjointness axioms are missing at times. Therefore, we define

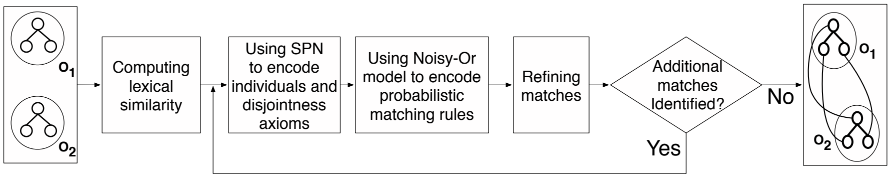
O1
O2
Computing lexical similarity
Using SPN to encode individuals and disjointness axioms
Using Noisy-Or model to encode probabilistic matching rules
Refining matches
Additional matches Identified?
No
Yes
O1
O2

Fig. 1: Matching process in GMap

a special assignment—”$Unknown$” of the similarities based on these individuals and disjointness axioms.

For 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 disjointness axioms, we utilize these axioms and subsumption relations within ontologies and define some rules to determine assignments of similarity. 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 as well as their descendants. The similarity also have three assignments. Assignment 1 (or 0) means the pair mismatches (or overlaps). If all the rules are not satisfied, the assignment is $Unknown$.

Using SPN to encode the simialrities based on individuals and disjointness axioms Sum-Product Network is a directed acyclic graph with weighted edges, where variables are leaves and internal nodes are sums and products [9]. As shown in Figure 2, we designed a sum-product network S to encode above similarities and calculate the contributions. All the leaves, called indicators, are binary-value. M represents the contribution of individuals and disjointness axioms and indicators M_1, M_2, M_3 comprise the assignments of it. $M_1 = 1$ (or $M_2 = 1$) means that the contribution is positive (or negative). If $M_3 = 1$, the contribution is $Unknown$. Similarly, Indicators D_0, D_1, I_1, I_2, I_3 correspond to assignments of the similarities based on individuals and disjointness axioms. The concrete assignment metrics are listed in Table 1–2 and the assignment metric of M is similar to the metric of similarity D.

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

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

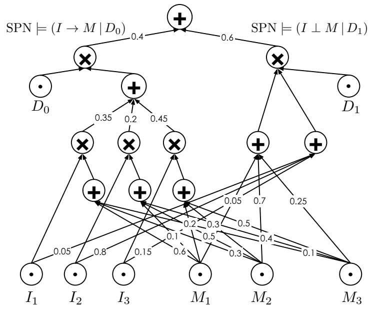

Fig. 2: The designed SPN : When $D_0 = 1, D_1 = 0$, it means that the distribution of M depends on the distribution of I; When $D_0 = 0, D_1 = 1$, the distributions of M and I are independent.

With the MAP inference in SPN [9], we can obtain the indicators' value of contribution M. The MAP inference has three steps. Firstly, replace sum nodes with max nodes. Secondly, with the bottom-up method, each max node can get a maximum weighted value. Finally, the downward pass starts from the root node and recursively selects the highest-value child of each max node, then the indicators' value of M are obtained. Moreover, even if individuals or disjointness axioms are missing at times, We can also calculate the contribution M by MAP inference. Assumed $I = 1$, $D = Unknown$ for one pair, then we can obtain $I_1 = 1, I_2 = 0, I_3 = 0, D_0 = 1, D_1 = 1$ with defined similarities and assignment metrics of SPN. As contribution M is not given, so we need to set $M_1 = 1, M_2 = 1, M_3 = 1$. After MAP inference, we observe $M_1 = 1$ which means that the contribution is positive. Moreover, it is able to infer $D_0 = 1$, which means the pair overlaps.

As the network S is complete and decomposable, the inference in S can be computed in time linear in the number of edges [4]. So MAP inference is tractable.

Combining the lexical similarity and the contribution calculated by SPN Considering the range of lexical similarity, we define a scaling factor α to limit the contribution of lexical similarity. It can help us to analyze the sources from different contributions. The SPN-based similarity (S_0) is defined in Eqs 1, which is calculated according to the indicators' value of M and D.

$$S_0(x_1, x_2) = \begin{cases} 0 & M_2 = 1, D_1 = 1 \\ \alpha * lexSim(x_1, x_2) + \lambda & M_1 = 1, D_0 = 1 \\ \alpha * lexSim(x_1, x_2) - \lambda & M_2 = 1, D_0 = 1 \\ \alpha * lexSim(x_1, x_2) & M_3 = 1, D_0 = 1 \end{cases} \tag{1}$$

where λ is a contribution factor that represents the contribution based on disjointness axioms and individuals. If contribution is positive (negative) and pair overlaps, the SPN-based similarity is equal to the scaled lexical similarity adding (subtracting) λ. If the contribution is $Unknown$ and pair overlaps, the SPN-based similarity is equal to the scaled lexical similarity. If the pair mismatches, then the inferred contribution is negative and the SPN-based similarity is equal to 0.

Using Noisy-Or model to encode probabilistic matching rules As listed in Table 3, we utilize probabilistic matching rules to describe the influences among the related pairs across ontologies.

Table 3: The probabilistic matching rules between entity 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

Considering the matching probability of one pair, we observe that the condition of each rule has two value (i.e., T or F) and all the matching rules are independent of each other approximately. Moreover, all of them benefit to improving the matching probability of this pair. Therefore, we utilize the noisy-or model [5] to encode them.

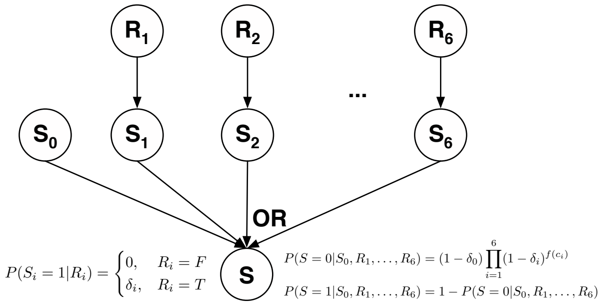
R_1
R_2
R_6
...
S_0
S_1
S_2
S_6
OR
S
$P(S_i = 1|R_i) = \begin{cases} 0, & R_i = F \\ \delta_i, & R_i = T \end{cases}$
$P(S = 0|S_0, R_1, \ldots, R_6) = (1 - \delta_0)\prod_{i=1}^{6}(1 - \delta_i)^{f(c_i)}$
$P(S = 1|S_0, R_1, \ldots, R_6) = 1 - P(S = 0|S_0, R_1, \ldots, R_6)$

Fig. 3: The network structure of noisy-or model designed in GMap

Figure 3 shows the designed noisy-or model applied in concept pairs and the extension to property pairs is straight-forward, 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 which is a leak probability [5]. We can easily calculate the matching probability of each pair, $P(S = 1|S_0, R_1, ..., R_6)$, according to the formulas listed in this figure, where 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 [5], so GMap can keep inference tractable in the whole matching process.

1.3 Adaptations made for the evaluation

There are two kinds of parameters that need be set. one mainly comes from networks and it is set manually based on some considerations [2]. The others are adapted by I3CON data set[2] such as scaling factor (α), contribution factor (λ) in Eqs 1 and threshold (θ). Nevertheless, we do not make any specific adaptation for OAEI 2015 evaluation campaign and all parameters are the same for different tracks.

1.4 Link to the system and parameters file

The latest version of GMap can be seen on https://github.com/liweizhuo001/GMap1.1.

1.5 Link to the set of provided alignments

The results of GMap can be seen on https://github.com/liweizhuo001/GMap1.1.

2 Results

In this section, we present the results of GMap achieved on OAEI 2015. Our system mainly focuses on Benchmark, Anatomy, Conference. Adding to that, we also present the results of the test Ontology Alignment for Query Answering which not follow the classical ontology alignment evaluation on the SEALS platform.

2.1 Benchmark

The goal of Benchmark is to evaluate the matching systems in scenarios where the input ontologies lack important information. Table 4 summarizes the average results[3] of it.

Table 4: Results for Benchmark track

Test	Precision	Recall	F-Measure
biblio	0.93	0.53	0.68
energy	0.32	0.02	0.11

GMap had a good performance in biblio, ranking third in F-measure, because it makes use of the string resource such as identifiers, labels and comments. Specially in ontologies 201–210 of biblio, as the mapping concepts have the same group of individuals but different names, SPN can play a role in improving the alignment quality of GMap.

[2] http://www.atl.external.lmco.com/projects/ontology/i3con.html

[3] The new test set about energy exists some troubles.

2.2 Anatomy

The Anatomy track consists of finding an alignment between the Adult Mouse Anatomy (2744 classes) and a part of the NCI Thesaurus (3304 classes) describing the human anatomy. The results are shown in Table 5.

Table 5: Results for Anatomy track

Matcher	Runtime (s)	Size	Precision	F-Measure	Recall	Recall+	Coherent
AML	40	1477	0.956	0.944	0.931	0.82	√
XMAP	50	1414	0.928	0.896	0.865	0.647	√
LogMapBio	895	1549	0.882	0.891	0.901	0.738	√
LogMap	24	1397	0.918	0.88	0.846	0.593	√
GMap	2362	1344	0.916	0.861	0.812	0.534	-

GMap ranked fifth in Anatomy track. We analyze that GMap does not concentrate on language techniques such as the abbreviations and emphasizes one-to-one constraint. Both of them may cause a low recall. In addition, these top-ranked systems employ alignment debugging techniques, which is helpful to improve alignment quality. However, we do not employ these techniques in the current version.

2.3 Conference

Conference track contains sixteen ontologies from the conference organization domain. There are two versions of reference alignment. The original reference alignment is labeled as RA1, and the new reference alignment, generated as a transitive closure computed on the original reference alignment, is labeled as RA2. Table 6 shows the results of our system in this track.

Table 6: Results for Conference track

	Precision	Recall	F-Measure
RA1	0.66	0.65	0.65
RA2	0.63	0.59	0.61

For Conference track, GMap ranked sixth of the 14 participants, which outperforms others in recall except AML but its precision is lower than them. There are mainly two reasons. One is the lexical similarity which combines the similarities based on edit-distance, external lexicons and TFIDF with the max strategy. The other is the noisy-or model which is hard to describe the negative effect on pairs matching [5]. Both of them would retain some false positive matches after matching finished. Specially in property pairs, even though their domains and ranges mismatch, GMap can not describe this negative impact. Therefore, employing alignment debugging techniques are comparatively ideal method solutions to deal with this problem.

2.4 Ontology Alignment for Query Answering (OA4QA)

The aims of OA4QA are investigating the effects of logical violations affecting computed alignments and evaluating the effectiveness of repair strategies employed by the matchers. In the OAEI 2015 the ontologies and reference alignment (RA1) are based on the conference track. RAR1 is a repaired version of RA1 different from RA2 in the conference track. The table 7 presents the results for the whole set of queries.

Table 7: Results for OA4QA track

Matcher	Answered queries	RA1			RAR1		
		P	R	F	P	R	F
GMap	9/18	0.324	0.389	0.343	0.303	0.389	0.330

Since GMap did not consider mapping repair techniques, it was only able to answer half of queries, which influenced the obtained precision and recall at last.

3 General comments

3.1 Comments on the results

GMap achieved qualified results in its first participation in OAEI, which is competitive with other systems in some tracks such as Benchmark, Conference, Anatomy. Both of the employed graphical models are able to improve the quality of alignment in terms of the defined lexical similarity [6]. Most improvements are attributed to the noisy-or model because it makes use of rich relations specified in ontologies such as in Anatomy track. If there are some individuals and disjointness axioms declared in ontologies, SPN will work such as biblio (201–210) in Benchmark track. More importantly, Combining SPN and the noisy-or model is able to increase precision and recall further.

However, some weaknesses still remain. For example, the alignment incoherence of GMap is unsolved, which influences the performance of GMap. In addition, it is important for us to consider the efficiency of GMap such as running time and memory usage for large-scale mapping problems.

3.2 Discussions on the way to improve the proposed system

GMap still has a lot of room for improvement. Employing alignment debugging techniques are able to solve the alignment incoherent and reduce some false positive matches in alignment such as the pair {Conference: has_members, edas: hasMember} in Conference track. In addition, seeking available data sets to learn parameters of the sum-product network and the noisy-or model is also one direction of our future works.

4 Conclusion

In this paper, we have presented GMap and its results of four tracks (i.e.,Benchmark, Conference, Anatomy and Ontology Alignment for Query Answering) on OAEI 2015. The results show that GMap is competitive with the top-ranked systems in some tracks by means of combining some special graphical models (i.e.,SPN, Noisy-or model). On the other hand, for those disadvantages exposed, we discuss the possible solutions. In the future, we would like to participate in more tracks and hope to efficiently solve the instance matching and large biomedical ontologies matching challenges.

Acknowledgments. This research was partly supported by the Natural Science Foundation of China (No. 61232015), the National Key Research and Development Program of China (Grant No. 2002CB312004), the Knowledge Innovation Program of the Chinese Academy of Sciences, Key Lab of Management, Decision and Information Systems of CAS, Institute of Computing Technology of CAS, and the Key Laboratory of Multimedia and Intelligent Software at Beijing University of Technology.

References

1. Albagli, S., Ben-Eliyahu-Zohary, R., Shimony, S.E.: Markov network based ontology matching. Journal of Computer and System Sciences **78**(1) (2012) 105–118
2. Ding, L., Finin, T.: Characterizing the semantic web on the web. In: The Semantic Web-ISWC 2006. Springer (2006) 242–257
3. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer Science & Business Media (2013)
4. Gens, R., Pedro, D.: Learning the structure of sum-product networks. In: Proceedings of The 30th International Conference on Machine Learning. (2013) 873–880
5. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
6. Li, W.: Combining sum-product network and noisy-or model for ontology matching
7. Mitra, P., Noy, N.F., Jaiswal, A.R.: Omen: A probabilistic ontology mapping tool. In: The Semantic Web–ISWC 2005. Springer (2005) 537–547
8. Niepert, M., Meilicke, C., Stuckenschmidt, H.: A probabilistic-logical framework for ontology matching. In: AAAI, Citeseer (2010)
9. Poon, H., Domingos, P.: Sum-product networks: A new deep architecture. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE (2011) 689–690