Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **4**, Number **3**, August **2019**

An Automatic Biomedical Ontology Meta-matching Technique

Xingsi Xue*

College of Information Science and Engineering
Intelligent Information Processing Research Center
Fujian Key Lab for Automotive Electronics and Electric Drive
Fujian Provincial Key Laboratory of Big Data Mining and Applications
Fujian University of Technology
No.3 Xueyuan Road, University Town, Minhou, Fuzhou City, Fujian Province, 350118, China
*Corresponding Author:jack8375@gmail.com

Haiyan Yang

College of Information Science and Engineering
Fujian University of Technology
No.33 Xuefu South Road, University Town, Minhou, Fuzhou City, Fujian Province, 350118, China
yanghy@fjut.edu.cn

Jie Zhang

School of Computer Science and Engineering
Yulin Normal University
No.299 Education Middle Road, Yulin City, Guanxi Province, 537000, China
jgxyzjzj@126.com

Jing Zhang

School of Computing
Ulster University
Jordanstown, Northern Ireland, BT370QB U.K.
j.zhang4@ulster.ac.uk

Dongxu Chen

Fujian Medical University Union Hospital
No.29 Xinquan Road, Gulou, Fuzhou City, Fujian Province, 350001, China
dxchen@fjmu.edu.cn

Received Jan 2019, Revision April 2019

Abstract. *Biomedical ontology matching aims at determining the heterogeneous biomedical concepts, and bridging the semantic gap between heterogeneous biomedical ontologies. The foundation of a biomedical ontology matching technique is the Biomedical Concept Similarity Measure (BCSM), which calculates the similarity value between two biomedical concepts. Since various BCSMs have different advantages, usually several BCSMs are aggregated together to improve the result's confidence. How to tune the aggregating weights to ensure the quality of the alignment is called biomedical ontology meta-matching problem, which is a challenge in the ontology matching domain. Currently, researchers mainly focus on how to tune the aggregating weights for various similarity measures to improve the quality of the ontology alignments. However, the ignorance of the effects brought about by different biomedical concept mapping's preference on different similarity measures significantly reduces the alignment's quality. To overcome this drawback, in this work, we formally define the biomedical ontology meta-matching problem, and then present an Ordered Weighted Average (OWA) based approach to automatically aggregate various biomedical concept similarity measures. In our method, the aggregating weights determined for each concept mapping is associated with the ordered position of the similarity value instead of a particular concept similarity measure. The experiment utilizes the large biomed track provided by Ontology Alignment Evaluation Initiative (OAEI) to test our proposal's performance, and the experimental results show the effectiveness of our proposal.*

Keywords: Biomedical ontology matching, Biomedical concept similarity measure, Ordered Weighted Average

---

1. **Introduction.** Biomedical ontology is a state-of-the-art technique for solving the biomedical concept heterogeneity problem. However, different biomedical ontology engineers might describe the same biomedical concepts in different ways, yielding the biomedical ontology heterogenous problem. Biomedical ontology matching can determine the heterogeneous biomedical concepts, and bridge the semantic gap between heterogeneous biomedical ontologies. Biomedical Concept Similarity Measure (BCSM) is the kernel technique of a biomedical ontology matching technique [1], which calculates the similarity value between two biomedical concepts. Since various BCSMs have different advantages, usually several BCSMs are aggregated together to distinguish the heterogeneous biomedical concepts. How to tune the aggregating weights to ensure the quality of the alignment, i.e. biomedical ontology meta-matching problem, is now a challenge in the ontology matching domain [2].

Currently, researchers mainly focus on how to determine the optimal weights for aggregating various BCSMs. However, the ignorance of the effects brought about by different biomedical concept mapping's preference on different BCSMs significantly reduces the alignment's quality [3]. For example, it is better to use the linguistic measure instead of syntactic measure to distinguish two terms "Myocardium" and "Cardiac Muscle Tissue". To ensure the alignment's quality, in this work, we formally construct a mathematical model for the biomedical ontology meta-matching problem, and then we present an Ordered Weighted Average (OWA) [4] based approach to automatically aggregate various BCSMs. In our approach, the aggregating weight of each concept mapping is determined by the sorting position of the similarity value instead of a particular BCSM.

The rest of the paper is organized as follows: Section 2 presents the mathematical model of the biomedical ontology matching problem; Section 3 describe in details various BCSMs; Section 4 presents the OWA-based biomedical ontology meta-matching process; Section 5 gives the experimental study; and finally, Section 6 draws the conclusion.

2. **Biomedical Ontology Matching.** A biomedical ontology O consists of a concept set, a property set and a axiom set [5], and a biomedical ontology alignment A can be defined as a concept mapping set, and each concept mapping is a 3-tuples $(c_1, c_2, simValue)$ where c_1 and c_2 are two biomedical concepts from two ontologies respectively, and $simValue \in [0, 1]$ is the similarity value between c_1 and c_2 . Basing on the principle that the larger scale of the concept mapping set and the higher mean similarity valuer is in an alignment, the higher quality of it is [6], we utilize the following equation to measure a biomedical ontology alignment A's quality:

$$f(A) = 0.5 \times MC(A) + 0.5 \times \frac{\sum_{i=1}^{|A|} simValue_i}{|A|} \tag{1}$$

where $|A|$ is A's cardinality, $MF()$ calculates A's MatchCoverage [7] based on A's concepts mappings number, $simValue_i$ is the ith concept mapping's similarity value. On this basis, the biomedical ontology meta-matching process is defined as a five-tuple $(O_1, O_2, A_{set}, W_{set}, F)$, where:

- O_1 and O_2 are two biomedical ontologies, A_{set} is a set of the biomedical ontology alignments corresponding to diverse BCSMs;
- W_{set} is a set of various aggregating weight sets;
- $F : W_{set} \to S \in [0, 1]$ evaluates the quality of a weight set $W \in W_{set}$:

$$F(W) = f(A), \quad A = \sum_{i=1}^{|A_{set}|} w_i A_i \quad with \quad w_i \in W \quad \text{and} \quad A_i \in A_{set} \tag{2}$$

3. **Biomedical Concept Similarity Measure.** BCSM takes as input two biomedical concepts and returns a value in [0,1] reflecting their similarity. In particular, two concepts are the same if their similarity value is 1, and completely different if 0. Generally, there are three broad categories of BCSM, i.e. syntactic measures, linguistic measures and taxonomy measures. In the next, we will introduce them one by one in details.

3.1. **Syntactic measure.** Syntactic measure works by calculating the edit distance of two biomedical concepts. In this work, we use two classic syntactic measures, i.e. Levenshtein distance [9] and Jaro distance [10]. Given the biomedical concepts c_1 and c_2 , the similarity values based on Levenshtein distance $sim_{Levenshtein}(c_1, c_2)$ and Jaro distance $sim_{Jaro}(c_1, c_2)$ are respectively defined as follows:

$$sim_{Levenshtein}(c_1, c_2) = max\Big(0, \frac{min(|c_1|, |c_2|) - dist(c_1, c_2))}{min(|c_1|, |c_2|)}\Big) \tag{3}$$

$$sim_{Jaro}(c_1, c_2) = \frac{1}{3}\Big(\frac{comm(c_1, c_2)}{|c_1|} + \frac{comm(c_1, c_2)}{|c_2|} + \frac{comm(c_1, c_2) - comm'(c_1, c_2)}{comm(c_1, c_2)}\Big) \tag{4}$$

where: $|c_1|$ and $|c_2|$ are the character number of c_1 and c_2, respectively, $dist(c_1, c_2)$ is c_1 and c_2's edit distance, $comm(c_1, c_2)$ is common characters' number of $|c_1|$ and $|c_2|$, $comm'(c_1, c_2)$ is the number of the common character pairs with different positions.

3.2. **Linguistic measure.** Linguistic measure utilize an electronic dictionary to calculate the similarity value of two biomedical concepts. In this work, we select The Unified Medical Language System (UMLS) [11] as the electronic knowledge base to calculate a synonymy-based distance. Given two biomedical concepts' names n_1 and n_2, the linguistic similarity value between them is calculated as follows:

$$sim_{Linguistic}(n_1, n_2) = \begin{cases} 1, & \text{if } n_1 \text{ and } n_2 \text{ are synonymous;} \\ 0.5, & \text{if } n_1 \text{ and } n_2 \text{ are hyponymous or hypernymous;} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

3.3. **Taxonomy-based measure.** Taxonomy-based measure makes use of the biomedical ontology's concept hierarchy to determine the distance between two biomedical concepts. In this work, we first construct a profile for each biomedical concept, which includes the information of its direct ascendant, descendants and siblings. Then, the taxonomy-based similarity value of two biomedical concepts c_1 and c_2 is calculated as follows:

$$sim_{Taxonomy}(c_1, c_2) = \frac{\sum_{i=1}^{|p_1|} \max\limits_{j=1\cdots|p_2|}(sim_e(p_{1i}, p_{2j})) + \sum_{j=1}^{|p_2|} \max\limits_{i=1\cdots|p_1|}(sim_e(p_{1i}, p_{2j}))}{f + g} \tag{6}$$

where p_1 and p_2 two profiles respectively corresponds to c_1 and c_2, $|p_1|$ and $|p_2|$ are respectively the cardinalities of p_1 and p_2. In particular, $sim_e(p_{1i}, p_{2j})$ measures two profile elements' similarity value through SMOA distance [12].

4. **Similarity Aggregation with Order Weighted Average.** It is a difficult task to determine the suitable w_i for aggregating various biomedical concept similarity measure since different biomedical concept mappings have different preferences on the concept similarity measure. In this work, we investigate the ordered weighted averaging (OWA) technique to automatically tune the aggregating weights. Given a set of similarity values $SimValue = (simValue_1, simValue_2, \cdots, simValue_n)$ on a biomedical concept mapping, where n is the number of BCSMs. After reordering the elements in $SimValue$ in descending order, we obtain $SimValue' = (simValue'_1, simValue'_2, \cdots, simValue'_n)$, and the final similarity value ca be calculated according the the following equation:

$$simValue_{final} = \sum_{i=1}^{n} w_i \cdot simValue'_i, \quad \sum_{i=1}^{n} w_i = 1, \quad w_i \in [0,1] \tag{7}$$

where w_i is the ith BCSM's aggregating weight. Here, a weight w_i is determined by the sorting position of the similarity value instead of a particular concept similarity measure. Then, the ith aggregating weight w_i is equal to $Q(\frac{i}{n}) - Q(\frac{(i-1)}{n})$ [13], where given two predefined thresholds $a, b \in [0,1]$, $Q(r) = 0, \frac{r-a}{b-a}, 1$ respectively when $r < a, a \leq r \leq b, r > b$.

Given n BCSMs $sm_1, sm_2, \cdots, sm_n$, a biomedical concept mapping (c_1, c_2), $sm_i(c_1, c_2)$ is is the similarity value between c_1 and c_2 determined by sm_i, which indicates the degree to which (c_1, c_2) satisfies sm_i. In this work, we utilize the *Alh* principle, i.e. satisfies at least half of BCSMs, to combine BCSMs, which can achieve the highest average alignment quality on all testing cases.

5. **Experimental Studies and Analysis.** In this work, we exploit the Large Biomed [1] track provided by the Ontology Alignment Evaluation Initiative (OAEI) [2]. Large Biomed track aims at matching three biomedical ontologies FMA (with 78,989 biomedical concepts), SNOMED CT (with 122,464 biomedical concepts) and NCI (with 66,724 biomedical concepts).

As shown in Table 1, our proposal's f-measures are better than OAEI's participants in all three tasks. In particular, the our proposal's precision values are in general high, which further indicates that our proposal is effective.

6. **Conclusion.** To improve the biomedical ontology alignment's quality, an OWA-based biomedical ontology meta-matching technique is proposed. Our approach is able to determine the aggregating weights for various biomedical concept similarity measures automatically. The experiment utilized the OAEI's large biomed track to test our proposal's performance, and the experimental results show the effectiveness of our proposal.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Nos. 61503082 and 61841603), the Natural Science Foundation of Fujian Province (Nos. 2016J05145, 2015J01652 and 2017J05098), the Program for New Century Excellent Talents in Fujian Province University (No. GY-Z18155), the Program for Outstanding Young Scientific Researcher in Fujian Province University (No. GY-Z160149), the Scientific Research Foundation of Fujian University of Technology (Nos. GY-Z17162 and GY-Z15007), the Guangxi Natural Science Foundation (No. 2018JJA170050) and the Education Department of Fujian Province Science and Technology Project (JZ160461).

REFERENCES

[1] T. T. A. Nguyen and S. Conrad, Ontology Matching using multiple similarity measures, Knowledge Discovery, *7th International Joint Conference on Knowledge Engineering and Knowledge Management*, vol.1, pp.603–611, Lisbon, Portugal, 2015.

[2] P. Shvaiko and J. Euzenat, Ontology matching: state of the art and future challenges, *IEEE Transactions on knowledge and data engineering*, vol.25, no.1, pp.158–176, 2013.

[3] X. Xue and J. Pan, A Compact Co-Evolutionary Algorithm for Sensor Ontology Meta-Matching, *Knowledge and Information Systems*, vol.56, no.2, pp.335–353, 2018.

[4] Q. Ji, P. Haase P and G. Qi, Combination of similarity measures in ontology matching using the owa operator, *Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice*, Springer, Berlin, Heidelberg, pp.281–295, 2011.

[5] G. Acampora, V. Loia and A. Vitiello, Enhancing ontology alignment through a memetic aggregation of similarity measures, *Information Sciences*, vol.250, pp.1-20, 2013.

[1] http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei

[2] http://oaei.ontologymatching.org/2018

TABLE 1. Comparison of our approach with OAEI's participants on Large Biomed track

Task1: whole FMA and NCI ontologies			
Systems	*recall*	*precision*	$f-measure$
AML	0.87	0.84	0.86
LogMap	0.81	0.86	0.83
LogMapBio	0.83	0.83	0.83
XMap2	0.74	0.88	0.80
FCAMapX	0.84	0.67	0.74
LogMapLt	0.82	0.68	0.74
DOME	0.67	0.80	0.73
Our approach	0.87	0.89	0.87
Task2: whole FMA and SNOMED ontologies			
Systems	*recall*	*precision*	$f-measure$
FCAMapX	0.76	0.82	0.79
AML	0.69	0.88	0.77
LogMapBio	0.65	0.83	0.73
LogMap	0.65	0.84	0.73
XMap2	0.61	0.72	0.66
LogMapLt	0.21	0.85	0.33
DOME	0.20	0.94	0.33
Our approach	0.79	0.86	0.82
Task3: whole SNOMED and NCI ontologies			
Systems	*recall*	*precision*	$f-measure$
AML	0.67	0.90	0.77
FCAMapX	0.68	0.80	0.73
LogMapBio	0.63	0.85	0.72
LogMap	0.60	0.87	0.71
LogMapLt	0.57	0.80	0.66
DOME	0.49	0.91	0.63
XMap2	0.58	0.64	0.61
Our approach	0.72	0.88	0.79

[6] J. Bock and J. Hettenhausen, Discrete particle swarm optimisation for ontology alignment, *Information Sciences*, vol.192, pp.152–173, 2012.

[7] X. Xue and Y. Wang, Optimizing Ontology Alignments through a Memetic Algorithm Using both MatchFmeasure and Unanimous Improvement Ratio, *Artificial Intelligence*, vol.223, pp.65–81, 2015.

[8] C. J. Van Rijsberge, *Information Retrieval*, University of Glasgow, Butterworth, London, 1975.

[9] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet physics doklady*, vol.10, no.8, pp.707–710, 1966.

[10] M. A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, vol.84, no.406, pp.414-420, 1989.

[11] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research*, vol.32(suppl_1), pp.D267–D270, 2004.

[12] G. Stoilos, G. Stamou and S. Kollias, A string metric for ontology alignment, *International Semantic Web Conference*, Springer, Berlin, Heidelberg, pp.624–637, 2005.

[13] R. R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, *IEEE Transactions on systems, Man, and Cybernetics*, vol.18, no.1, pp.183–190, 1988.