Ontology Alignment using Combined Similarity Method and Matching Method

Didih Rizki Chandranegara, Riyanarto Sarno
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
diedieh02@gmail.com, riyanarto@if.its.ac.id

***Abstract*—Ontology Alignment or Ontology Matching is used to identify entities contained in two ontologies or more, that have similar or matching. Ontology Alignment can determine a relationship between two ontologies or more. There has been several tools that can be used for Ontology Alignment. Ontology Alignment is widely use in finding similar entities in bibliography created by Bibtex and IFLA (International Federation of Library Associations and Institutions). Bibliography is writing in detail about the identity of an article such as books and journals. It is very important to know the differences and similarities between Bibliography format. It aims when changing one format Bibliography to another format should not change the overall. In this research, we combine several similarity methods and Brute Force String matching method to produce a better Ontology Alignment result. The results showed that the proposed method outperfom the previous method. And the results of Ontology Alignment showed that between Ontology BibTex and IFLA has a relationship. This is evidenced by several entities of both Ontology are matching.**

Keywords—Ontology Alignment; Matching; Jaccard Distance; JaroWinkler Distance Edit Distance; Monge Elkan; Level2 Monge Elkan; Brute Force.

I. INTRODUCTION

Ontology Alignment or Ontology Matching is a process to identify the relationship between the entities contained in the ontology [1]. Ontology Alignment is used to handle heterogeneity between two ontologies or more, so that between the ontology have a relationship with each other [2] [3] [4]. The heterogeneity can be the naming of entity, one entity use simple name while the other use detail name. For Ontology Alignment has been used in several tools such as SAMBO, Ri-MOM, and FALCON-AO [5].

In the research conducted by Jiang, Lowd, and Dou [6], used similariy on the relationship between two ontologies to find a link between the two ontologies. This research used a probabilistic framework to integrate the ontology scheme, and a knowledge-based strategy to compare the ontology by name and value contained in data properties in each ontology.

Essay and Abed in [7] combined 5 similarity methods which are Jaro Winkler Distance, Jaccard Distance, Levenshtein Distance, Dice Coefficient, and TriGram. This method need weight value of execution time. After the similarity values are obtained then the matching was perfomed.

Previous researches [8] [9] [10] conducted studies of ontology matching by using weighting. The results showed that the ontology matching with weights better than just using string matching keyword. While, Sun, Ma and Wang [11] proposed a combination of several String Similarity methods for the Ontology Matching.

In this research we proposed a method which was developed from research [7] using a combination of similarity methods i.e. Jaccard Distance, Jaro Winkler Distance, Edit Distance, Monge Elkan, and Level2 Monge Elkan, as well as for matching method used is Brute Force which is described in section 2. The reason for using 5 similarity methods is because has been widely used by many people and the application of used this methods has been found in some researchs. The sequence of similarity method is not affect the results of similarity value. While, for the use of Brute Force String Matching Method is because the entity of Bibtex Ontology and IFLA does not have a long string, so using this method are considered appropriate and may provide a good performance in execution.

Ontology used an ontology of Bibtex and IFLA (International Federation of Library Associations and Institutions). The purpose of this study was to identify the entities from two different Ontology (Bibtex and IFLA) that have similar or matching. In addition, this study also aims to determine whether using the combined method similarity which proposed can increase the value of the similarity. There has been no related research to determine the relationship between Bibtex Ontology and IFLA Ontology, so knowing the relationship between Bibtex Ontology and IFLA Ontology can make additional contributions for this research. For the limitation of the research is Ontology Alignment/ Matching is performed on Class, DataProperties and Object Properties on both Ontology, when relations between the entities are ignored. For the results of the

978-1-5090-1648-8/16/$31.00 ©2016 IEEE

experiments contained in Sections 3 and discussion of the experimental results are discussed in section 4. And then will proceed to conclusion.

II. METHODS

A. Jaccard Distance

This method is used to determine the ratio between the intersection and the union on two strings [7]. Here is a formula to determine Jaccard Distance [7] [11]:

$$Jaccard(s1, s2) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Where s1 and s2 is the first string and second string symbolized by A and B. This method is usually used to find the similarity between two strings.

B. JaroWinkler Distance

This method is used to measure the similarity between two strings. Where the greatest similarity value indicates that the two strings are similar [7]. Here is a formula to measure the similarity value of Jaro Winkler [7] [11] :

$$d_w = d_j + (\delta_p(1 - d_j)) \quad (2)$$

With δ_p is the value of prefix length between two strings. While d_j is the value of Jaro which obtained from :

$$d_j = \begin{cases} 0, if\ m=0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_1|} + \frac{m-t}{m}\right) \end{cases} \quad (3)$$

Where m is the same number of characters between two strings. While s_1 and s_2 is the length of the string 1 and string 2. Value of t is get from the amount of transposition between two strings.

C. Edit Distance

Edit distance is the method used to convert a string into another string [3]. In this method there are process of Copy, Insert, Substitute and Delete [12]. After the value of the Edit Distance is obtained, then the next process did normalized distance shown in the following formula :

$$sim(s_1, s_2) = 1 - \frac{Edit\ Distance}{\max\_length(s_1, s_2)} \quad (4)$$

Where s_1 and s_2 is a first string and a second string. And *max_length* is the maximum string length between the first string and second string.

D. Monge Elkan

Monge Elkan is a variation of the Smith-Waterman, where to do the matching between two strings with character : {d t} {g j} {l r} {m n} {b p v} { a e i o u} {, .}. Here is scoring the Monge Elkan method [11] [13]:

1. If the two strings are the same character, then will be given a score of +5.
2. If the two strings are different, then be compared to the character set : {d t} {g 8 W-1 {m 4 UJ P 4 { a e a o u} {, .}, then if found appropriate characters with character sets, then will be given a score of +3.

Here is the formula used to find the value of Monge Elkan :

$$MongeElkan = \frac{total\_score}{\min_{legth}(s_1, s_2) * 5} \quad (5)$$

E. Level2 MongeElkan

This method is the development of methods Monge Elkan [11] [14]. For the formula used is as follows :

$$Level2ME = \frac{\max_{score}(MongeElkan)}{length(s_1)} \quad (6)$$

F. Brute Force String Matching

Brute Force is a string matching method that has a fast performance [15]. In addition to having fast performance, a brute force method is a simple algorithm that can be used for a search string [16]. So this method is suitable for use on a String Matching.

In this method there are two inputs, namely; pattern and text. Pattern is a string that will be searched in the text. String searches carried out from the beginning of the character pattern to match the appropriate characters with text [16] [17]. Illustration of the process of string matching this method can be seen in Fig. 1.

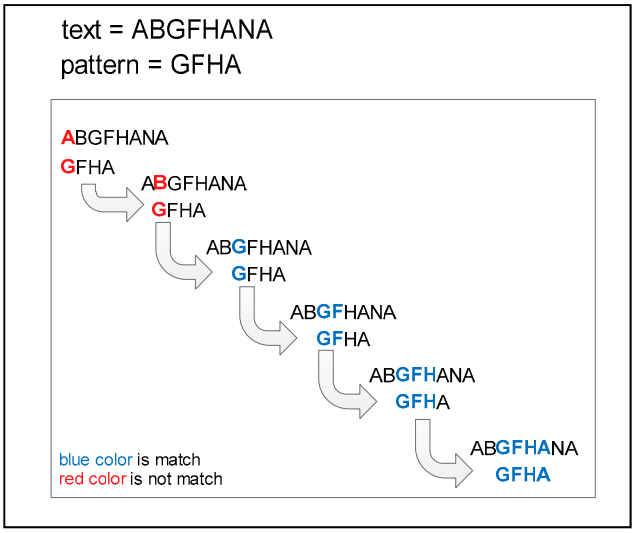

Fig. 1. Ilustration of Brute Force String Matching

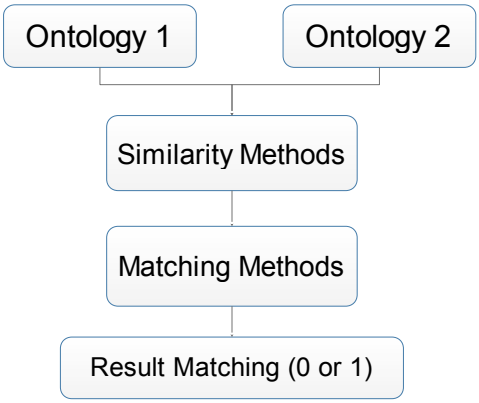

Fig. 2. Ontology Alignment Process

G. Ontology Alignment Process

This process is used to identify the relationship between two or more Ontology [2]. If many entities between Ontology are aligned, then the relationship between the Ontology getting stronger. Here are the steps of Alignment Process (It has been presented in Fig. 2) :

1. Insert the two Ontology as input
2. After that do the similarity of the entities contained in the Ontology 1 and Ontology 2.
3. Having obtained the value of similarity between the entities, then selected which have a value greater than the threshold (In this research used a threshold = 0.5)
4. The results of similarity that has been sorted, and then do the matching process using string matching (one of them is Brute Force String Matching)
5. If the value of the result of matching is equal to 1, then the entity Ontology 1 and Ontology 2 have a relationship. Meanwhile, if the result of matching indicates a value equal to 0 then the entity of Ontology 1 and Ontology 2 do not have a relationship.

H. Proposed Method

The proposed method in this study is a combination of similarity and matching method. On the concept of similarity adopt methods of research conducted Essayeh and Abed [7], where for the used method is use a combination of several methods of similarity and using weights. But in this study, some similarity method used [7] has been replaced with some other similarity method. The changed method from research [7] are Levenshtein Distance, Dice Coefficient, and trigram replaced by Edit Distance, Monge Elkan, and Level 2 Monge Elkan. The purpose of this replacement is to increase the value of the similarity of the previous methods. As for the matching method used is using Brute Force string matching. The proposed method in this research have been presented in Fig. 3.

The proposed method is using weights as intermediaries to merge similarity method. The weights of each method is obtained from the execution time for the process of similarity divided by the total execution time of all methods of similarity. For more valid results, the execution of each method performed 5 times, this is to get the average execution time. It is the formula to get the weights [7] :

$$W_i = \frac{average(execution_{time}method)_i}{SUM(execution_{time}all\ method)} \quad (7)$$

Where i = 1 to 5, sequentially i=1 for Jaccard Distance Method, i=2 for Jaro Winkler Distance Method, i=3 for Edit Distance Method, i=4 for Monge Elkan Method, and i=5 for Level2 Monge Elkan method. If all the weights are summed, it will produce a value equal to 1.

The results of the similarity symbolized by $S_{local(i)}$, where the value of i is also the result of each method similarity, while for the result of the combination is symbolized S_{global}. Here is the formula used to find the value S_{global} [7]:

$$S_{global} = \sum_{i=1}^{5} W_i * S_{local(i)} \quad (8)$$

Result of S_{global} were compared with a threshold (threshold = 0.5). If S_{global} > threshold, then will proceed to the matching process. And if not, then it will not proceed to the matching process.

For the matching process used is Brute Force Method, used value 0 or 1. If the result show the value is 1, then entity between two Ontology is matching. If the result show the value is 0, then the entity between the two Ontology is not matching.

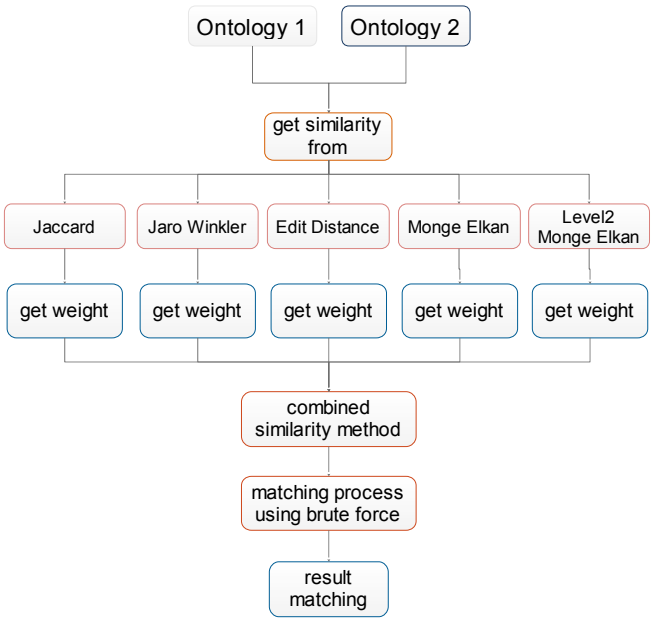

Fig. 3. Proposed Method

TABLE I. EXECUTION TIME

Methods	Execution Times (*ms*)		
	Class	***Object Properties***	***Data Properties***
Jaccard	0.022	0.018	0.0103
Jaro Winkler	0.004	0.003	0.002
Edit Distance	0.004	0.003	0.002
Monge Elkan	0.035	0.039	0.023
Level2 Monge Elkan	0.044	0.054	0.009

TABLE II. RESULTS MATCHING ENTITY OF PROPOSED METHOD

Entity	Bibtex	IFLA
Class	Agent	Agent
	Artifact	Artifact
	Corporate_Body	Corporate_Body
	Person	Person
	Work	Work
Object Properties	hasPart	hasPart
	place_publication	Place
	responsibility	responsibility
Data Properties	audience	Audience
	dimensions	dimensions
	edition	Edition
	frequency	Frequency
	language	Language
	name	Name
	number	Number
	penname	Name
	series_title	Title
	series_title	Series
	title	Title

III. RESULTS

Experiments were performed using the Java programming language. For the data used is ontology Bibtex and IFLA obtained from the data Ontology Department Of Computer Science, University of Toronto [18]. Ontology alignment process is the entity of BibTex and IFLA Ontology, where such entities including Class, Data Properties, and Object Properties. On BibTex Ontology there are 43 Class, 22 Object Properties, and 24 Data Properties. And on IFLA Ontology there are 12 Class, 36 Object Properties, and 63 Data Properties.

In table 1 shows the execution time of the method which is execution time on Class, Object Properties, and Data Properties. Thus, to make the process of similarity Class, Object Properties and Data Properties will be done with a different execution time.

For the results of the proposed method has been presented in Table 2. Where, in the table there is the name of entities which matching between BibTex and IFLA Ontology.

IV. DISCUSSION

From the results of experiments conducted indicate that the entity Class between BibTex dan IFLA Ontology there are 5 Class which matched from 43 class BibTex and 12 class IFLA. Whereas for Object Properties from experiment shows that there are 3 object properties which matched from 22 object properties BibTex and 36 object properties IFLA. And the last result shows there are 11 data properties which matched from 24 data properties BibTex and 63 data properties IFLA.

The first discussion was to compare the results of the similarity of the proposed method with previous method [7]. Results from previous methods [7] has been presented in Table 3 and the results of the proposed method has been presented in Table 4.

In previous research similarity method [7] using a combination of Jaro Winkler Distance, Distance Jaccard, Levenshtein Distance, Dice Coefficient, and TriGrams. And the result of the previous method showed the average value of similarity under 0.2. While the proposed method shows similarity values above 0.2. In the proposed method there is an increase in the value of the similarity of the previous methods and to the matching process can provide a good matching results.

Furthermore, the first discussion is also using WordNet as a comparison to show that the value of similarity in the proposed methods is increase (has been presented in Table 5). WordNet method used are WUP, LIN, and PATH. Where, results from WordNet methods will be combined with the following formula:

$$Wordnet\ Method = \frac{(WUP+LIN+PATH)}{3} \quad (9)$$

When compared to the previous method with a combination of WUP, LIN and PATH in Wordnet, does not seem that there is the same similarity values or approach to the value of the combined method in WordNet. However, the proposed method there are two words (Writer with Provider and Writer and Product) that show similarity value is almost equal to the combined WUP, LIN and PATH. This suggests that the similarity in the proposed methods show an increase in value of similarity than the previous method.

The following discussion is to compare the proposed methods with results that combine several methods. Experiments were performed using four combinations of similarity methods and matching method. Where the results of the experiment also uses weight. With threshold is 0.5.

TABLE III. PREVIOUS METHOD

	Author	Provider	Product	Creator
Writer	0.119	0.159	0.075	0.137
Publisher	0.125	0.161	0.091	0.091

TABLE IV. PROPOSED METHOD

	Author	Provider	Product	Creator
Writer	0.177	0.391	0.294	0.348
Publisher	0.284	0.178	0.17001	0.149

TABLE V. WORDNET METHOD (WUP, LIN & PATH)

	Author	Provider	Product	Creator
Writer	1	0.39096	0.2968	0.4628
Publisher	0.4083	0.40896	0.2647	0.4083

The following experiments are performed:

1. The first experiments using a combination of methods Jaccard, Jaro Winkler and Brute Force. The results of the first experiment no detectable entities.
2. In the second experiment used a combination of methods Jaccard, Jaro Winkler, Edit Distance and Brute Force. And the results are also no detectable entities. From both these experiments not detectable entity because the value of the resulting similarity is lower than the threshold.
3. In the third experiment using a combination Jaccard, Jaro Winkler, Edit Distance, Monge Elkan and Brute Force. The results obtained indicate that the entity Class is matched equally by the proposed method. While on the entities object properties there are only two entities matching. And the entities data properties there 8 entities which matching. The results of the third experiment can be seen in Table 6.
4. And the last experiment using a combination Jaccard, Jaro Winkler, Edit Distance, Level2 Monge Elkan and Brute Force. There is a growing entity matching, that is on the Data Properties, from 8 to 9 entities Data Properties which matching. The results of last experiments are shown in Table 7.

Based on the fourth experiment, it can be seen that combining some similarity method can increase the value of similarity. And the proposed method using the combined methods of similarity there is an increase in the ontology matching entities between BibTex and IFLA. In addition, the results of this experiment showed that between Bibtex ontology and IFLA ontology has relation between entities of Bibtex and IFLA. Thus, for writing bibliographies using Bibtex format and IFLA format there are similarities data between Bibtex and IFLA. And does not need to make full changes the content of bibliography, whether it's changing the bibliography Bibtex to IFLA or IFLA to Bibtex.

In another experiment, use of several different threshold that are 0.2, 0.25, 0.3, 0.35, 0.4, and 0.45. And based on the results of experiments conducted indicate that the entity of Class, Object Properties and Data Properties shows stable results. Where the total entities matching Class is 5, Object

TABLE VI. JACCARD, JARO WINKLER, EDIT DISTANCE, MONGE ELKAN & BRUTE FORCE

Entity	Bibtex	IFLA
Class	Agent	Agent
	Artifact	Artifact
	Corporate_Body	Corporate_Body
	Person	Person
	Work	Work
Object Properties	hasPart	hasPart
	responsibility	responsibility
Data Properties	audience	Audience
	dimensions	Dimensions
	Edition	Edition
	frequency	Frequency
	Language	Language
	name	Name
	number	Number
	title	Title

TABLE VII. JACCARD, JARO WINKLER, EDIT DISTANCE, LEVEL2 MONGE ELKAN & BRUTE FORCE

Entity	Bibtex	IFLA
Class	Agent	Agent
	Artifact	Artifact
	Corporate_Body	Corporate_Body
	Person	Person
	Work	Work
Object Properties	hasPart	hasPart
	responsibility	responsibility
Data Properties	audience	Audience
	dimensions	Dimensions
	Edition	Edition
	frequency	Frequency
	Language	Language
	name	Name
	number	Number
	penname	Name
	title	Title

Properties is 3 and Data Properties is 11. From these results indicate that the threshold value used in the proposed method, there are the same results with some threshold that has been in trial. And from these results indicate that the similarity of the proposed method showed stable results with some threshold that has been in trial. So use a threshold between 0.2 to 0.5 will not provide the addition or subtraction entities which matching between BibTex and IFLA ontologies.

V. CONCLUSION

The results of this research showed that the similarity of the proposed methods can increase the value of the similarity than the previous methods. From some experiments were performed using a combination of several methods, it has been shown that the proposed method gives results matching more entities. While, for some threshold used show that the proposed method showed stable results for entities which macthing. And the results of Ontology Alignment showed that between Ontology BibTex and IFLA has a relationship. This is evidenced by several entities of both Ontology are matching. And for writing bibliographies using Bibtex format or IFLA format, no need full change the contents of bibliography, because of the experimental results in this study show that there are similarities content between Bibtex format and IFLA format.

For next research, the similarity process can be adding another method. So, the matching process can give the better results than before. In addition, we can used relationship between the entities in the Bibtex ontology and IFLA ontology, which is useful to know the detail relation between Bibtex ontology and IFLA ontology.

REFERENCES

[1] E. Jim´enez-Ruiz, T. R. Payne, A. Solimando and V. Tamma, "Limiting Logical Violations in Ontology Alignment Through Negotiation," *Proc. KR,* vol. 16, 2016.

[2] M. H. Khan, S. Jan, I. Khan and I. A. Shah, "Evaluation of linguistic similarity measurement techniques for ontology alignment," in *International Conference on Emerging Technologies (ICET)*, 2015.

[3] X. Xue, J. Liu, P.-W. Tsai, X. Zhan and A. Ren, "Optimizing Ontology Alignment by using Compact Genetic Algorithm," in *11th International Conference on Computational Intelligence and Security (CIS)*, 2015.

[4] J. W. Son, H. G. Yoon and S. B. Park, "A Ontology Kernel-A Convolution Kernel for Ontology Alignment," *Journal Of Information Science and Engineering,* vol. 31, no. 2, pp. 415-432, 2015.

[5] R. Zhu, Y. Hu, K. Janowicz and G. McKenzie, "Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics," *Transactions in GIS,* 2016.

[6] S. Jiang, D. Lowd and D. Dou, "Ontology Matching with Knowledge Rules," *Database and Expert Systems Applications,* pp. 94-108, September 2015.

[7] A. Essayeh and M. Abed, "Towards ontology matching based system through terminological, structural and semantic level," *Procedia Computer Science,* vol. 60, pp. 403-412, 2015.

[8] W. Hayuhardhika, N. Putra, Sugiyanto, R. Sarno and M. Sidiq, "Weighted Ontology and weighted tree similarity algorithm for diagnosing Diabetes Mellitus," in *International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, Jakarta, 2013.

[9] A. Arwan, B. Priyambadha, R. Sarno, M. Sidiq and H. Kristianto, "Ontology and semantic matching for diabetic food recommendations," in *International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2013.

[10] R. Sarno, H. Ginardi, E. W. Pamungkas and D. Sunaryono, "Clustering of ERP business process fragments," *in International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 2013.

[11] Y. Sun, L. Ma and S. Wang, "A Comparative Evaluation of String Similarity Metrics for Ontology Alignment," *Journal of Information & Computational Science,* pp. 957-964, 2015.

[12] B. Mikhail, M. Raymond, C. William, R. Pradeep and F. Stephen, "Adaptive Name Matching in Information Integration," *IEEE Intelligent Systems,* pp. 16-23, 2003.

[13] A. E. Monge and C. P. Elkan, "The Field Matching Problem: Algorithms and Applications," 1996.

[14] W. W. Cohen, P. Ravikumar and E. S. Fienberg, "A Comparison of String Metrics for Matching Names and Records," in *Kdd workshop on data cleaning and object consolidation*, 2003.

[15] S. Ahn, H. Hong, H. Kim, J.-H. Ahn, D. Baek and S. Kang, "A hardware-efficent multi-character string matching architecture using brute-force algorithm," in *International SoC Design Conference (ISOCC)*, 2009.

[16] R. A. Abdeen, "An Algorithm for String Searching Based on Brute-Force Algorithm," *IJCSNS,* vol. 11, no. 7, p. 24, 2011.

[17] V. SaiKrishna, A. Rasool and N. Khare, "String Matching and its Applications in Diversified Fields," *International Journal of Computer Science Issues,* vol. 9, no. 1, pp. 219-226, 2012.

[18] "University Of Toronto," Computer Science, [Online]. Available: http://www.cs.toronto.edu/semanticweb/maponto/index.html. [Accessed 28 April 2016].