Review implementation of linguistic approach in schema matching

Galih Hendro Martono a,1,*, Azhari SN b,2

a Doctoral Program of Computer Science, Universitas Gadjah Mada, Indonesia
b Department of Computer Science & Electronics, Universitas Gadjah Mada, Indonesia
1 galih.hendro@smkbumigora.ac.id; 2 arisn@ugm.ac.id
* corresponding author

I. Introduction

The purpose of schema matching is to identify correspondence between two or more schemas [1]–[9]. Along with the development of information technology, data integration-related issues become more complex as increment of data quantity, development of technologies of current website, and implementation of schema matching in a variety aspect of human life. Indirectly, the problem itself made us look furthermore for solution related to data integration. One way to overcome the problem of data integration is schema matching. Some research suggests that schema matching can be applied to domains such as data integration, e-business, semantic web, e-commerce, data warehouse and semantic query processing [7], [10], [11].

There was many research related to schema matching, as research conducted by [11]–[18] about schema matching. Other research conducted by [19] combined two approach existing in schema matching namely constraint-based and instance-based to get better result. In the research known that the result has fairly good. This can be seen from precision values is 71.43%, recall is 75%, and F-Measure is 81.48% [19]. Errors results on this paper occurs in three case, including use of an id attribute with data type as auto increment; using codes that are defined in the same way but different meaning; and if encountered in common instance with the same definitions on the attributes but different meaning. To evaluated implementation of schema matching some surveys and evaluation has been conducted by [1], [2], [7], [10], [20]–[29]. [7] made taxonomy from schema matching approaches. One approach mentioned in the publication was linguistic approach. Some research related linguistic approach has been conducted prior such as [23], [30]–[45]. Some research in linguistic focused to calculation of similarity between of two or more schemas. In calculating the entity similarity in a database used help from dictionary and thesaurus. The use of dictionary and thesaurus to help searching words that have common in the word (synonym) or word that have same pronunciation but have different meaning (homonym). [31] was used multi-strategies to calculated similarity of element. This approach different from other approach because in the former research all variable information are defined as features in a single similarity function but in multi-strategies all variable information are defined base on different types of information, and a composite method was
used to combine the results of different similarities. To help in schema matching processes or calculating similarity of an entity in a schema, there are few tools can used such as COMA [8], [46], [47], COMA++ [1], [48], [49], RiMOM [31], [50], and SMART [4].

This paper conducted review of implementation of linguistic approach in schema matching processes. Linguistic approach doing element based on means of an element. Challenge in implementation of linguistic approach is difficulty in implementation of linguistic approach because difficulty of deciphering the meaning of a word that matches. Linguistic approach is one of material study from natural language processing. Natural language processing is branch of science that specifically examine interaction between computer and natural language of human. Natural language processing can be considered to be branch of artificial intelligence and its study material intersect with linguistic computational. In practice natural language processing works by taking into account knowledge of the language itself, both in terms of the words used, how words are combined to produce a sentence, the meaning of a word, the function of a word in a sentence and so on.

II. Research Method

This publication discussed about implementation linguistic approach in schema matching and focused on measure similarity to finding string in the schema matching process. In implementing the schema matching with the linguistic approach, datasets and algorithms used will be an input in the process of schema matching. Each element will be matched against the naming of elements based on the algorithm used and the system will process the input according to the algorithm used. Schema matching process can be seen in Fig. 1.

![Schema matching process](image)

Linguistic approach did match the element name of a database by using stemming, tokenization, string matching, and information retrieval techniques [2]. To do the matching words used dictionary and thesaurus. There are differences between dictionary and thesaurus. The thesaurus shows the relationship of a term with other terms, while the dictionary defines a term or word. Thesaurus can be used in information processing and information retrieval tool and can be used to discover the meaning of a word and can also be used to find the structure of vocabulary, such as use: ..., Use for: ..... and so on, or for example, the library can be national libraries, college libraries, school libraries, etc. One of tools that can be used to assist in finding a word synonyms and abbreviations can use WordNet.

Schema matching process with linguistic approach conducted with seeing the similarity of element naming in database exist. Calculation of name similarity can be done with tokenization and calculating word similarity value [34], [44]. Measurement of similarity between two or more elements is done to overcome problem of abbreviations, synonyms, hypernym and more on naming an element. In calculating the word similarity value need to consider several things such as synonyms (e.g. cars have in common with vehicle), hypernym (books can mean publishing or book), and similarity in pronunciation [51]. There are three stages in performing matching with linguistic approach, namely normalization, categorization, and comparison [9].

A. Normalization

Normalization in schema matching can be done with:

- **Tokenization.** Tokenization is a process of sentence splitting based on its composed word. Each word called as token or term. For examples POLines -> {PO, Lines}

- **Generalization.** Generalization of word used when word contain acronym. For examples {PO, Line} -> {Purchase, Order, Lines}
• Elimination. Elimination can be done by eliminated affix, preposition and conjunction so used are based word.
• Tagging. Tagging of word which has the same meaning or the likelihood of association such as price, cost, and value can be associated with the concept of money.

B. Categorization
Categorization is done by grouping the elements that have the same word association. The purpose of the categorization is to reduce the elements to be compared so that later, words or elements that have the same association are grouped into one category, and this category will be compared to see the similarities.

C. Element similarity value
Measurement of similarity values of token (T1 and T2) is done by using the formula [52]

\[ N \text{sim}(T1,T2) = \frac{\sum_{t1\in T1} [\text{max} t2\in T2 \text{sim}(t1,t2)] + \sum_{t2\in T2} [\text{max} t1\in T1 \text{sim}(t1,t2)]}{|T1|+|T2|} \] (1)

D. Comparison
Comparisons were made to compute the similarity value of the category before. Linguistic similarity calculation is done based on the similarity of the elements and calculate the average weight of tokens. If \(T1\) and \(T2\) is a token element of \(m1\) and \(m2\) then the calculation of similarity of the names of \(m1\) and \(m2\) as follows [9]

\[ \text{ns} (m1,m2) = \frac{\sum_{i\in \text{TokenType}} w_i \times \text{ns}(T1i,T2i)}{\sum_{i\in \text{TokenType}} w_i |T1i+|T2i|} \] where \(\sum w_i = 1\) (2)

Linguistic similarity (lsim) calculate by doing scaling from maximum of similarity value from two categories [9].

\[ \text{lsim} (m_1, m_2) = \text{ns}(m_1, m_2) \times \max_{c_1 \in C1, c_2 \in C2} \text{ns} (c_1, c_2) \] (3)

where \(C1\) and \(C2\) are sets from \(m1\) and \(m2\) belong, respectively. The result of this phase is a table of linguistic similarity coefficients between elements in the two schemas. The similarity is assumed to be zero for schema elements that do not belongs to any compatible categories.

In other research [53] used Lavenstein (edit-distance), 3-gram, and jaro-distance to compute the similarity between two sets. The detailed measure similarity described in the following.

E. Lavenstein (Edit-distance)
The measure similarity of two word (s and t) is measured by the following equation:

\[ \text{sim}(s,t) = \frac{\max(|s|,|t|) - \text{editDistance}(s,t)}{\max(|s|,|t|)} \] (4)

F. 3-grams
This algorithm compute similarity of word by separates words into two part namely s and t. Each part (s and t) is three sequential characters respectively, for example, the string s = distance and string t = instance will have the 3-gram sets of s is tris(s) = \{dis, ist, sta, tan, anc, nce\}, and the 3-gram set of t is tris(t) = \{ins, nst, sta, tan, anc, nce\}. The intersection of the 3-gram is tris(s) \cap tris(t). The similarity of 3-gram is measured to be:

\[ \text{sim}(s,t) = \frac{2|\text{tris}(s) \cap \text{tris}(t)|}{|\text{tris}(s)| + |\text{tris}(t)|} \] (5)

G. Jaro-distance
The number string will be matched separates in two string (s and t) and then will be found and counted first. Then, the number of transposing the matched characters in s to the place of t is counted. The similarity of transposition is computed as
\begin{equation}
\text{sim}(s,t) = \frac{1}{3} \times \left( \frac{m}{|s|} \right) + \left( \frac{m}{|t|} \right) - \left( \frac{m-n}{m} \right)
\end{equation}

where \( m \) is the number of matched characters and \( n \) is the number of transpositions.

III. Results and Discussion

A. Results

This section will discuss implementation of the measure similarity that was published in [53], [54]. [53] presented a hybrid schema matching approach based on Cupid scheme to find the similarity of generic schema and generate match result. This approach called SYM. The proposed SYM approach includes two phases. In the linguistic similarity matching phase, a new linguistic matching method was proposed to find the similarity between two element names in schemas. The structural similarity matching phase calculate the structure similarity between two sets of nodes. The approach is evaluated by testing on several benchmarks of real schemas and comparing with other methods such as Cupid, COMA++, and Similarity Flooding. The results of evaluation this approach are shown in Table 1 [53].

<table>
<thead>
<tr>
<th>Methods vs Scheme</th>
<th>Accuracy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Preision</td>
<td>Recall</td>
<td>F-score</td>
</tr>
<tr>
<td>SYM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>1.0000</td>
<td>0.6150</td>
<td>0.7620</td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>0.6000</td>
<td>0.6000</td>
<td>0.6000</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.9000</td>
<td>0.80375</td>
<td>0.8405</td>
<td></td>
</tr>
<tr>
<td>Cupid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>0.8750</td>
<td>0.5380</td>
<td>0.6670</td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>0.7500</td>
<td>0.6000</td>
<td>0.6670</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>0.5000</td>
<td>0.6000</td>
<td>0.5454</td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.7813</td>
<td>0.6845</td>
<td>0.7199</td>
<td></td>
</tr>
<tr>
<td>COMA++</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>0.8750</td>
<td>0.538</td>
<td>0.6667</td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>0.5000</td>
<td>0.6000</td>
<td>0.5454</td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td>0.6000</td>
<td>0.8570</td>
<td>0.7058</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.7438</td>
<td>0.7488</td>
<td>0.7295</td>
<td></td>
</tr>
<tr>
<td>Similarity Flooding</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>1.0000</td>
<td>0.4615</td>
<td>0.6315</td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>0.8000</td>
<td>0.8000</td>
<td>0.8000</td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td>1.0000</td>
<td>0.1428</td>
<td>0.2500</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.7000</td>
<td>0.3518</td>
<td>0.4204</td>
<td></td>
</tr>
</tbody>
</table>

Form the results of table 1 found that SYM method has the top accuracy in three of the domains and average (on University Schema, Student Schema, and PO Schema). Reference [54] had evaluated a wide range of string similarity metrics such as levenshtein (edit-distance), Jaro, NGram, SoftTFIDF using benchmarks data set and on conference data set. The experimental result shown in Fig 2 – Fig 3.

Galih Hendro Martono and Azhari SN (Review implementation of linguistic approach in schema matching)
Fig. 2 and Fig. 3 reveal a wide disparity among the performance of string similarity metrics. On benchmarks data set, all of the algorithms have the top performance in term of F-measure. Only Lin measure get the worst since its recall is quite low. On conference data set, the SoftTFIDF, Sigmoid, and Jaccard get the top F-measure. Max and Monge-Elkan have the worst F-measure. Fig. 4 and Fig. 5 shown computation time on benchmarks data set and on conference data set. Its figure out the hybrid methods spend too much time compared to non-hybrid methods.
B. Evaluate

In this section will be discussed about evaluation from implementation of algorithms. Evaluation from algorithms is important to conduct to see performance the algorithm when applied. To evaluate performance of algorithm used can be done by saw value of precision, recall, dan F-Measure. Evaluation of precision, recall, dan F-Measure has been widely applied in field of computer science like for precision evaluation conducted by [55]–[57], recall by [56], and F-Measure by [56], [58], [59]. Value of precision, recall, and F-Measure has range value of 0-1. Precision is comparison between values identified True Positive (TP) with the number of values identified True and False (TP+FP) or can be formulated as:

\[
\text{Precision (P)} = \frac{TP}{TP+FP}
\]

Recall is value comparison identified true with expected value. Recall can be formulated as:

\[
\text{Recall} = \frac{TP}{M}
\]

F-Measure provide the level of accuracy in matching process based on algorithm used. Calculation of F-Measure calculated value of precision and recall.

\[
F - \text{Measure (F)} = \frac{2 \times P \times R}{P + R}
\]

C. Related Work

Implementation thesaurus in schema matching has been long used as in web document classification, summarization, index, and calculate the semantic similarity of documents written in the same or in the different language [41]. In e-commerce research related this problem has been done by [33], [60], [61]. In the paper conducted by [33], linguistic approach to seeking entity similarity value can be used to web search interface for example providing a unified access e-commerce search search engines selling similar products in allowing users to search and compare products from multiple sites. Similar, the approach had also done by [42]. They proposed an approach to match pairs of catalogues using the estimated mutual information (EMI) matrix to measure similarity and defined how to derive thesaurus. In the research [45] used linguistic approach for database integration with Indonesian language database using WordNet. Because this time the database WordNet not support Indonesian language, the researchers translate existing words by using dictionary English-Indonesian. This study illustrates the application of linguistic and tools WordNet for cases other languages. WordNet is an opensource application that contains a collection of database dictionary English, in contrast to a dictionary generally focused on words, WordNet focuses on the meaning of the word. The meaning of a word in WordNet is represented in the form of synset (synonym set). In addition, WordNet can also search for a relationship between meaning as hypernym, hyponymy and hypernymy, holonym, and so on. WordNet can help in finding a match in the schema matching words. For Indonesian WordNet was developed by the Information Retrieval Lab Faculty of Computer Science, University of Indonesia. Indonesian WordNet synset has 1203, 1659 unique words, and relations existing synset relations reached 2261.

In their research [30] used graph as supported tools to calculate similarity value from a word (synonym and homonyms). Value of a word similarity can be calculated using token. In the application of linguistic approach, one problem encountered is the similarity value calculation method for measuring similarity [30], [52], [62], [63]. In their publication [31] states method used for linguistic approach called edit distance based strategy and vector distance (VD) based strategy. Futhermore, [31] combining multiple strategies to results of different similarities. This technique is based on calculating similarities between entities of two schemas by various type of information, e.g entity names, taxonomy structures, constraint and entities instances. In his paper [41] studied the effect of thesaurus size on schema matching quality using different thesaurus. Beside that, their proposed a new method in calculating the similarity between vectors extracted from thesaurus database.
IV. Conclusion

In this paper, linguistic was utilized as one of schema matching method. Many experiments were conducted to study of implements linguistic based using dictionary and thesaurus on schema matching. Generally, researches related linguistic approach discussed element entity similarity in schema. One application can be used to schema matching based on linguistic is WordNet which is application for dictionary and thesaurus saved English database. To evaluate of method used by using the results of the precision, recall, and F measure values.

Several interesting issues in implementation linguistic approach in schema matching is implementation in XML document and OEM graph, and resolving case in heterogeneous data and large amount of data. Implementation of linguistic approach can be combined with other approaches such us artificial intelligence, artificial neuron network, data mining, and machine learning. Furthermore, implementation of linguistic approach in other schema case or general case can be conducted with testing this approach. For development related to calculation for similarity values in matching process can develop other methods which has been done by [30], [32]. Other issues from schema matching, generally is used of semi-automatic and automatic approach. From this paper is expected can developed other research related to implementation of natural language processing in schema matching using method exist or combine with other method (hybrid).

References


Galih Hendro Martono and Azhari SN (Review implementation of linguistic approach in schema matching)


