ISSN: 2442-6571

# 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@stmikbumigora.ac.id *; [2] arisn@ugm.ac.id*
** corresponding author*

ARTICLE INFO

*Article history:*
Received December 15, 2016
Revised March 17, 2017
Accepted May 12, 2017

*Keywords:*
Schema Matching
Linguistic Approach
Natural Language Processing
Similarity Measure

ABSTRACT

Research related schema matching has been conducted since last decade. Few approach related schema matching has been conducted with various methods such as neuron network, feature selection, constrain based, instance based, linguistic, and so on. Some field used schema matching as basic model such as e-commerce, e-business and data warehousing. Implementation of linguistic approach itself has been used a long time with various problem such as to calculated entity similarity values in two or more schemas. The purpose of this paper was to provide an overview of previous studies related to the implementation of the linguistic approach in the schema matching and finding gap for the development of existing methods. Futhermore, this paper focused on measurement of similarity in linguistic approach in schema matching.

## 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 increcement 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

DOI: http://dx.doi.org/10.26555/ijain.v3i1.75

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.

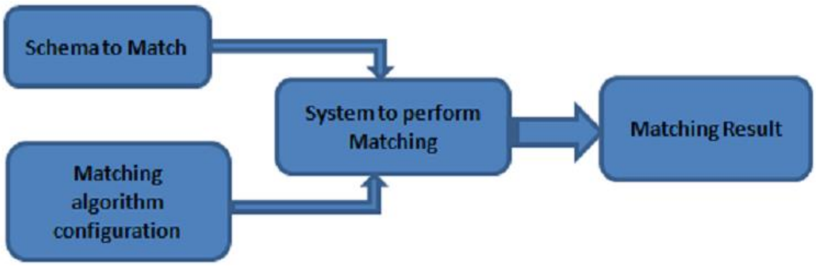

Fig. 1. Schema matching process

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]

$$N\ sim\ (T1, T2) = \frac{\sum \mathrm{t1 \text{€} T1[max\ t2 \text{€} T2sim(t1,t2)]} + \sum \mathrm{t2 \text{€} T2[max\ t1 \text{€} T1sim(t2,t1)]}}{\mathrm{|T1| + |T2|}} \tag{1}$$

### 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 $T_{1i}$ and $T_{2i}$ is a token element of $m_1$ and $m_2$ then the calculation of similarity of the names of $m_1$ and $m_2$ as follows [9]

$$\mathrm{ns\ (m1, m2)} = \frac{\sum_{\mathrm{i\ \text{£}\ TokenType}} \mathrm{w_i\ x\ ns(T_{1i}, T_{2i})}}{\sum_{\mathrm{i \text{€} TokenType}} \mathrm{w_i x\ |T_{1i}| + |T_{2i}|}} \quad \text{where} \sum w_i = 1 \tag{2}$$

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

$$\mathrm{lsim}\ (m_1, m_2)\ =\ \mathrm{ns}(m_1, m_2)\ \mathrm{x}\ max_{ci\ \text{€}\ C1, c2\ \text{€} C2}\ ns\ (c_1, c_2) \tag{3}$$

where $C_1$ and $C_2$ are sets from $m_1$ and $m_2$ 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:

$$sim(s, t) = \frac{\mathrm{max}(|s|, |t|) - editDistance(s, t)}{\mathrm{max(|s|, |t|)}} \tag{4}$$

### 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 tri(s) = {dis, ist, sta, tan, anc, nce}, and the 3-gram set of t is tri(t) ={ins, nst, sta, tan, anc, nce}. The intersection of the 3-gram is tris(s) ∩ tri(t). The similarity of 3-gram is measured to be:

$$sim(s, t) = \frac{2X|tri(s) \cap \mathrm{tri(t)}}{|tri(s)| + |tri(t)|} \tag{5}$$

### 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

$$sim\ (s,t) = \frac{1}{3} X \left(\frac{m}{|s|}\right) + \left(\frac{m}{|t|}\right) - \left(\frac{m-n}{m}\right) \tag{6}$$

where $m$ is the number of matched characters and n is the number of transpositions.

## III. Results and Discussion

### A. *Results*

This section will discussions about 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 1. The results of evaluation

| **Methods vs Scheme** | | **Accuracy** | | |
|---|---|---|---|---|
| | | ***Preision*** | ***Recall*** | ***F-score*** |
| SYM | University | **1.0000** | **0.6150** | **0.7620** |
| | Person | 0.6000 | 0.6000 | 0.6000 |
| | Student | **1.0000** | **1.0000** | **1.0000** |
| | PO | **1.0000** | **1.0000** | **1.0000** |
| | ***Average*** | **0.9000** | **0.80375** | **0.8405** |
| Cupid | University | 0.8750 | 0.5380 | 0.6670 |
| | Person | 0.7500 | 0.6000 | 0.6670 |
| | Student | 0.5000 | 0.6000 | 0.5454 |
| | PO | **1.0000** | **1.0000** | **1.0000** |
| | ***Average*** | 0.7813 | 0.6845 | 0.7199 |
| COMA++ | University | 0.8750 | 0.538 | 0.6667 |
| | Person | **1.0000** | **1.0000** | **1.0000** |
| | Student | 0.5000 | 0.6000 | 0.5454 |
| | PO | 0.6000 | 0.8570 | 0.7058 |
| | ***Average*** | 0.7438 | 0.7488 | 0.7295 |
| Similarity Folooding | University | **1.0000** | 0.4615 | 0.6315 |
| | Person | 0.0000 | 0.0000 | 0.0000 |
| | Student | 0.8000 | 0.8000 | 0.8000 |
| | PO | **1.0000** | 0.1428 | 0.2500 |
| | ***Average*** | 0.7000 | 0.3518 | 0.4204 |

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 lavenstein (edit-distance), Jaro, NGram, SoftTFIDF using benchmarks data set and on conference data set. The experimental result shown in Fig 2 – Fig 3.

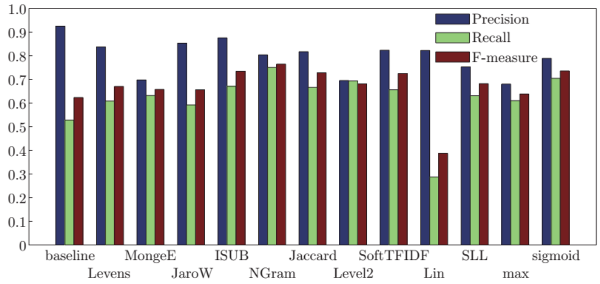

Fig. 2. Performance of sring similarity metrics on benchmarks data set

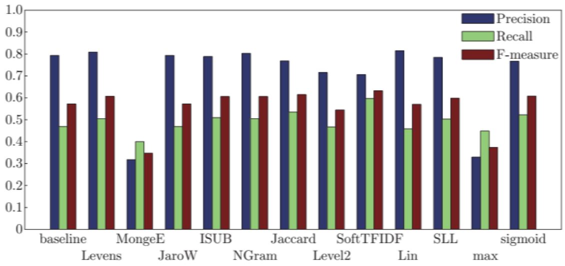

Fig. 3. Performance of string similarity metrics on conference data set

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.

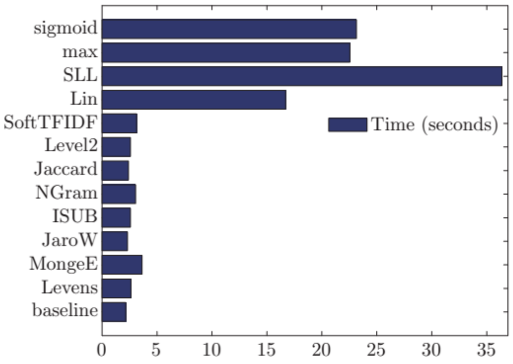

Fig. 4. Computation time on benchmarks data set

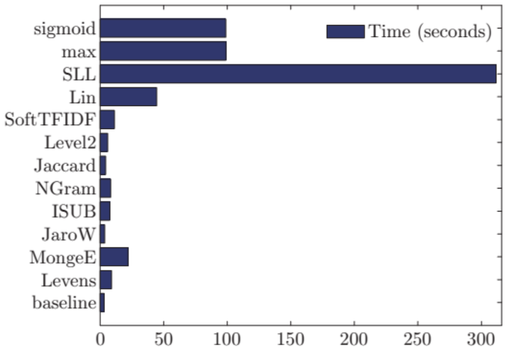

Fig. 5. Computation time on conference data set

### *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:

$$\text{Precission (P)} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

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

$$Recall = \frac{\text{TP}}{\text{M}} \quad (7)$$

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

$$F - Measure\ (F) = \frac{2 \text{ x P x R}}{\text{P+R}} \quad (8)$$

### *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, holonyum, 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 homonymies). 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 entitites 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

[1] H. H. Do and E. Rahm, "Matching large schemas: Approaches and evaluation," *Inf. Syst.*, vol. 32, no. 6, pp. 857–885, 2007.

[2] P. A. Bernstein, J. Madhavan, and E. Rahm, "Generic Schema Matching: Ten Years Later," *Proc. VLDB Endow.*, vol. 4, no. 11, pp. 695–701, 2011.

[3] B. He, K. C. Chang, and J. Han, "Discovering Complex Matchings across Web Query Interfaces : A Correlation Mining Approach," *Sigkdd*, pp. 148–157, 2004.

[4] T. Okawara, J. Tanaka, A. Morishima, and S. Sugimoto, "A Support Tool for XML Schema Matching and Its Implementation," in *Data Engineering*, 2005, pp. 1–4.

[5] L. Ratinov and E. Gudes, "Abbreviation expansion in schema matching and Web integration," in *Proceedings - IEEE/WIC/ACM International Conference on Web Intelligence, WI 2004*, 2004, pp. 485–490.

[6] B. He and K. C. C. Chang, "Statistical schema matching across web query interfaces," *Proc. 2003 ACM SIGMOD Int. Conf. Manag. data*, no. 1, p. 228, 2003.

[7] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," *VLDB J.*, vol. 10, no. 4, pp. 334–350, 2001.

[8] C. Clifton, E. Hausman, and A. Rosenthal, "Experience with a Combined Approach to Attribute Matching Across Heterogeneous Databases," *Proc. 7th IFIP Conf. Database Semant.*, pp. 428–453, 1997.

[9] J. Madhavan, P. A. Bernstein, and E. Rahm, "Generic schema matching using Cupid," in *Proc of 27th International Conference on Very Large Data Bases*, 2001, pp. 49–58.

[10] P. Shvaiko, "A Survey of Schema-based Matching Approaches," *J. Data Semant.*, vol. 3730, pp. 146–171, 2005.

[11] J. Berlin and A. Motro, "Database schema matching using machine learning with feature selection," *Adv. Inf. Syst. Eng.*, pp. 452–466, 2002.

[12] B. Kim, N. Ho, D. Lee, and S. J. Hyun, "A clustering based schema matching scheme for improving matching correctness of web service interfaces," *Proc. - 2011 IEEE Int. Conf. Serv. Comput. SCC 2011*, pp. 488–495, 2011.

[13] X. Zhong, Y. Fu, Q. Liu, X. Lin, and Z. Cui, "A holistic approach on deep web schema matching," *2007 Int. Conf. Converg. Inf. Technol. ICCIT 2007*, pp. 169–174, 2007.

[14] P. Sinha, R. K. Raj, and C. J. Romanowski, "A Holistic Approach to Schema Matching," *2009 WRI World Congr. Comput. Sci. Inf. Eng.*, pp. 116–120, 2009.

[15] B. Villányi and P. Martinek, "Towards a novel approach of structural schema matching," pp. 103–107, 2012.

[16] E. Rahm, "Towards Large-Scale Schema and Ontology Matching," *Schema Matching Mapp.*, pp. 3–27, 2011.

[17] T. Milo and S. Zohar, "Using schema matching to simplify heterogeneous data translation," *Vldb*, pp. 1–21, 1998.

[18] H. Elmeleegy, M. Ouzzani, and A. Elmagarmid, “Usage-based schema matching,” pp. 20–29, 2008.

[19] E. Sutanta, R. Wardoyo, K. Mustofa, and E. Winarko, “A hybrid model schema matching using constraint-based and instance-based,” *Int. J. Electr. Comput. Eng.*, vol. 6, no. 3, pp. 1048–1058, 2016.

[20] N. Noy, “Semantic integration: a survey of ontology-based approaches,” *SIGMOD Rec.*, vol. 33, no. 4, pp. 65–70, 2004.

[21] M. Peluang and P. Model, “Kajian Model dan Prototipe Schema Matching.”

[22] R. Blake, “ScholarWorks at UMass Boston A Survey of Schema Matching Research,” 2007.

[23] X. L. Sun and E. Rose, “Automated Schema Matching Techniques: An Exploratory Study Heterogeneity Problems Interoperability Concerns Semantic heterogeneity Semantic interoperability Structural heterogeneity Structural interoperability,” *Res. Lett. Inf. Math. Sci*, vol. 4, pp. 113–136, 2003.

[24] Z. Bellahsene, A. Bonifati, F. Duchateau, and Y. Velegrakis, “On Evaluating Schema Matching and Mapping,” *Schema Matching Mapp.*, pp. 253–291, 2011.

[25] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez, “Ontology matching: A literature review,” *Expert Syst. Appl.*, vol. 42, no. 2, pp. 949–971, 2015.

[26] M. Granitzer, V. Sabol, K. W. Onn, D. Lukose, and K. Tochtermann, “Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques,” *Futur. Internet*, vol. 2, no. 3, pp. 238–258, 2010.

[27] J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. Trojahn, “Ontology Alignment Evaluation Initiative : Six Years of Experience,” *J. Data Semant. XV, LNCS 6720*, vol. 6720, pp. 158–192, 2011.

[28] A. Doan and A. Y. Halevy, “Semantic integration research in the database community: A brief survey,” *AI Mag.*, vol. 26, no. 1, p. 83, 2005.

[29] H. Wache *et al.*, “Ontology-Based Information Integration: A Survey of Existing Approaches,” *Int. Jt. Conf. Artif. Intell. Work. Ontol. Inf. Shar.*, pp. 108–117, 2001.

[30] L. Palopoli, D. Saccà, G. Terracina, and D. Ursino, “Uniform techniques for deriving similarities of objects and subschemes in heterogeneous databases,” *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 2, pp. 271–294, 2003.

[31] J. Li, J. Tang, Y. Li, and Q. Luo, “RiMOM: A dynamic multistrategy ontology alignment framework,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 8, pp. 1218–1232, 2009.

[32] J. Euzenat, D. Loup, and M. Touzani, “[OLA] Ontology alignment with OLA.”

[33] H. He, W. Meng, C. Yu, and Z. Wu, “Automatic integration of Web search interfaces with WISE-integrator,” *VLDB J.*, vol. 13, no. 3, pp. 256–273, 2004.

[34] J. Lu, S. Wang, and J. Wang, “An Experiment on the Matching and Reuse of XML Schemas,” *Web Eng. Proc. 5th Int. Conf. ICWE 2005, Sydney, Aust. July 27-29, 2005*, pp. 273–284, 2005.

[35] P. Mitra and G. Wiederhold, “Resolving Terminological Heterogeneity In Ontologies Declaratively,” *Proc. Work. Ontol. Semant. Interoperability 15th Eur. Conf. Artif. Intell.*, pp. 45–50, 2002.

[36] W. E. Djeddi and M. T. Khadir, “A Dynamic Multistrategy Ontology Alignment Framework Based on Semantic Relationships using WordNet.”

[37] H. He, W. Meng, C.t.yu, and Z.wu, “Wise-integrator: an automatic integrator of web search interfaces for e-commerce,” *VLDB*, pp. 357–368, 2003.

[38] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, “Semantic integration of heterogeneous information sources,” *Data Knowl. Eng.*, vol. 36, no. 3, pp. 215–249, 2001.

[39] R. Steinberger, B. Pouliquen, and J. Hagman, “Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC,” *Comput. Linguist. Intell. Text Process. Proc. CICLing 2002*, vol. LNCS (2276, pp. 415–424, 2002.

[40] F. Boudin, J. Y. Nie, and M. Dawes, “Using a medical thesaurus to predict query difficulty,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7224 LNCS, pp. 480–484, 2012.

[41] T. Sabbah, A. Selamat, M. Ashraf, and T. Herawan, “Effect of thesaurus size on schema matching quality,” *Knowledge-Based Syst.*, vol. 71, pp. 211–226, 2014.

[42] L. A. P. Leme, D. F. Brauner, K. K. Breitman, M. A. Casanova, and A. Gazola, “Matching object catalogues,” *Innov. Syst. Softw. Eng.*, vol. 4, no. 4, pp. 315–328, 2008.

[43] S. Castano and V. De Antonellis, “A schema analysis and reconciliation tool environment for\nheterogeneous databases,” *Proceedings. IDEAS’99. Int. Database Eng. Appl. Symp. (Cat. No.PR00265)*, 1999.

[44] S. Castano, V. De Antonellis, and S. C. Di De Vimercati, “Global viewing of heterogeneous data sources,” *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 2, pp. 277–297, 2001.

[45] I. W. S. Wicaksana and R. A. Hakim, “Pendekatan Schema Matching dalam Bahasa Indonesia.”

[46] H.-H. Do and E. Rahm, “COMA: a system for flexible combination of schema matching approaches,” *Proc. 28th Int. Conf. Very Large Data Bases*, pp. 610–621, 2002.

[47] H.-H. Do, “Schema Matching and Mapping-based Data Integration,” *Dep. Comput. Sci.*, no. August, p. 222, 2006.

[48] D. Engmann and S. Massmann, “Instance Matching with COMA++,” *Citeseer*, pp. 144–156, 2004.

[49] D. Aumueller, H.-H. Do, S. Massmann, and E. Rahm, “COMA++ - Schema and ontology matching with COMA,” *Proc. 2005 ACM SIGMOD Int. Conf. Manag. data SIGMOD 05*, vol. pages, p. 906, 2005.

[50] C. Shao, L. M. Hu, J. Z. Li, Z. C. Wang, T. Chung, and J. B. Xia, “RiMOM-IM: A Novel Iterative Framework for Instance Matching,” *J. Comput. Sci. Technol.*, vol. 31, no. 1, pp. 185–197, 2016.

[51] E. Rahm, P. A. Bernstein, and U. Leipzig, “On Matching Schemas Automatically On Matching Schemas Automatically On Matching Schemas Automatically,” *Rep. Nr*, vol. 1, 2001.

[52] G. X. M. L. Schema, A. Algergawy, R. Nayak, and G. Saake, “QUT Digital Repository : XML Schema Element Similarity Measures :,” 2009.

[53] B. C. Chien and S. Y. He, “A hybrid approach for automatic schema matching,” *9th Int. Conf. Mach. Learn. Cybern.*, vol. 6, no. July, pp. 2881–2886, 2010.

[54] Y. U. of E. Sun, L. U. of E. Ma, and S. N. U. Wang, “A Comparative Evaluation of String Similarity Metrics for Ontology Alignment,” *J. Inf. Comput. Sci.*, vol. 12, no. 3, pp. 957–964, 2015.

[55] P. Bertolazzi, L. De Santis, and M. Scannapieco, “Automatic record matching in cooperative information systems,” *Proc. ICDT*, no. i, pp. 13–20, 2003.

[56] A. McCallum, K. Nigam, and L. H. Ungar, “Efficient clustering of high-dimensional data sets with application to reference matching,” *Proc. sixth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. KDD 00*, pp. 169–178, 2000.

[57] W. W. Cohen, “Integration of heterogeneous databases without common domains using queries based on textual similarity,” *Proc. 1998 ACM SIGMOD Int. Conf. Manag. data*, vol. 27, no. 2, pp. 201–212, 1998.

[58] I. Bhattacharya and L. Getoor, “Collective entity resolution in relational data,” *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 5–es, 2007.

[59] W. W. Cohen and J. Richman, “Learning to match and cluster large high-dimensional data sets for data integration,” *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 475–480, 2002.

[60] L. J. Nederstigt, S. S. Aanen, D. Vandić, and F. Frasincar, “An automatic approach for mapping product taxonomies in E-commerce systems,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7328 LNCS, pp. 334–349, 2012.

[61] L. Nederstigt, D. Vandic, and F. Frasincar, “An Automated Approach to Product Taxonomy Mapping in E-Commerce,” *Manag. Intell. Syst.*, pp. 1–10, 2012.

[62] B. Jeong, D. Lee, H. Cho, and J. Lee, “A novel method for measuring semantic similarity for XML schema matching,” *Expert Syst. Appl.*, vol. 34, no. 3, pp. 1651–1658, 2008.

[63] P. Bruza, “QUT Digital Repository : Combining Structure and Content Similarities for XML Document Clustering,” no. November, pp. 27–28, 2003.