International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 3, June 2016, pp. 1048 ~ 1058
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i3.9790

A Hybrid Model Schema Matching Using Constraint-Based and Instance-Based

Edhy Sutanta[1], Retantyo Wardoyo[2], Khabib Mustofa[2], Edi Winarko[2]
[1]Doctoral Program of Computer Science at Department of Computer Sciences & Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2]Department of Computer Science & Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia

Article Info

Article history:

Received Dec 25, 2015
Revised Apr 5, 2016
Accepted Apr 20, 2016

Keyword:

Constraint-based
Heterogeneous database
Hybrid model
Instance-based
Schema matching

ABSTRACT

Schema matching is an important process in the Enterprise Information Integration (EII) which is at the level of the back end to solve the problems due to the schematic heterogeneity. This paper is a summary of preliminary result work of the model development stage as part of research on the development of models and prototype of hybrid schema matching that combines two methods, namely constraint-based and instance-based. The discussion includes a general description of the proposed models and the development of models, start from requirement analysis, data type conversion, matching mechanism, database support, constraints and instance extraction, matching and compute the similarity, preliminary result, user verification, verified result, dataset for testing, as well as the performance measurement. Based on result experiment on 36 datasets of heterogeneous RDBMS, it obtained the highest P value is 100.00% while the lowest is 71.43%; The highest R value is 100.00% while the lowest is 75.00%; and F-Measure highest value is 100.00% while the lowest is 81.48%. Unsuccessful matching on the model still happens, including use of an id attribute with data type as autoincrement; using codes that are defined in the same way but different meanings; and if encountered in common instance with the same definition but different meaning.

Corresponding Author:

Edhy Sutanta,
Department of Informatics Engineering, IST AKPRIND, Yogyakarta, 55222, Indonesia.
Doctoral Program of Computer Science, Department of Computer Sciences & Electronics, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia.
Email: edhy_sst@akprind.ac.id, edhy_sst@yahoo.com

1. INTRODUCTION

Schema matching is a matching process inter-schema to find similar relationship of pair of attributes s [1], or arrange mapping and matching schema in two application systems [2]. Schema matching is a solution of Enterprise Information Integration (EII) [3] which is done at back end level to solve the problems of schematic heterogeneity [4], that is a different naming (type, format, and precision) in the schema definitions [5]. Technically, schema matching is an integration process on heterogeneous database and will produce a generalization or specialization in the database [6]. Schema matching plays important role in applications that requires interoperability between systems with heterogeneous data sources [7]. Schema matching is a main problem on developing the relationship between elements in the two database schema [2],[8]-[11]. Schema matching was originally done manually on a specific application domain [12], so it is needed a new model that is more general and appropriate for the application and different schema languages [13]. The main problem of schema matching is often found not clear naming in the schema, difficulties in synonyms naming, and schema language differences so that the matching method may not provide 100%

right in the result [2]. The schema matching cannot be done automatically because the mapping of computing is usually corrected by the user to obtain the correct verified results [14]-[15].

Development model and software on schema matching are still open to find proper ways to combine existing methods [11],[16]. The use of combinational matchers [17]-[18], can be implemented in hybrid or in composite [16-17]. Hybrid model is also called intra-matcher parallelism [19] using some criteria concurrently matching [13],[20]-[21] to give results and better performance than using individual matcher [17]. Simple concept of hybrid matcher is to combine two different methods simultaneously processed, while the composite matcher combines two methods that are processed in a sequence. Schema matching using a hybrid matcher was applied in CLIO [22]-[26], CUPID [18], and SYM [27]. While the schema matching using a composite matcher found in SEMINT [21],[28]-[29], LSD [30], Cupid [18], COMA [14], COMA++ [15], COMA 3.0 [31]-[32], IMAP [33], PROTOPLASM [34]-[37], FALCON-AO [2],[38], and ASMOV [39]. Refers to [40]-[41], development of new schema matching models and prototype is still open especially on hybrid models. The next section describes the proposed a new hybrid model schema matching that was developed based on constraint-based and instance-based.

2. THE PROPOSED MODEL

The proposed model of hybrid schema matching is by combining two methods (constraint-based and instance-based) implemented simultaneously. Constraint-based and instance-based are methods categories according to [11],[16]), in which involves the DTM (data type matcher), CM (constraint matcher), and IDM (instance data matcher) methods (categories according to [42]). Generally, the proposed model is developed refering to the general model of data processing, consisting of 4 sections, namely input, process, output, and verification and evaluation as shown in Figure 1. The description of each section are as follows:

1. Input, receives input by `DBSource` (as a reference database) and `DBTarget` (database to be matched), the type of DBMS, extracting constraints, data type conversion, extracting instances, and checking the similarity inter attributes in `DBSource` and `DBTarget`.
2. Process, conducting matching process, which match each attribute in `DBSource` with each attribute in `DBTarget` and then calculate the value of similarity (SIM_{MN}) on each possible pair matched attributes, and determine a pair of attributes declare matched.
3. Output, show the similarity mapping pair of attributes pair of attributes s, that is pairs of attributes that has the $SIM_{MN}MAX$ and SIM_{MN}=1, namely a preliminary result.
4. Verification and Evaluation. Verification is the process to determine whether the preliminary results generated by the model are correct or still need to be manually corrected by the user. Thus, the process is supervised approach. Preliminary result has been verified by the user produces the verified result in the form of mapping pair of attributes s that are valid. Evaluation process is performed to calculate the values of model performance parameters, which are P (Precision), R (Recall), and F (F-measure). The values of P, R, and F are calculated by comparing the preliminary result and the verified result.

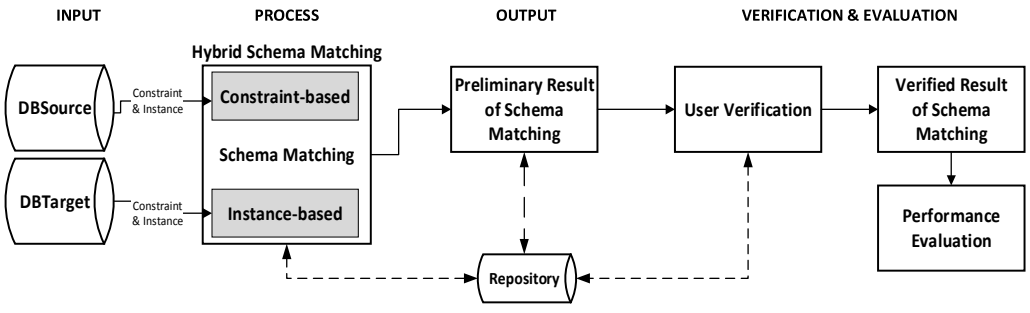


Figure 1. The proposed model of hybrid schema matching

3. MODEL IN DETAIL

3.1. Requirement Analysis

Requirement analysis is conducted on five aspects that are functional requirement, input document, output document, database, and model evaluation. Functional requirements of the proposed model are as follows:

1. Input of The model is `DBSource` and `DBTarget`.

2. The model can extract information schema to find the names of the tables, attribute names, and constraints (type, width, domain, nullable, unique) in DBSource and DBTarget.
3. The model can convert the data types on the attributes used by the DBMS on DBSource and DBTarget into new data type used by the model.
4. The model can extract instances in DBSource and DBTarget.
5. The model can match and compute the value of similarity between each pair of attributes on DBSource and DBTarget.
6. The model is able to determine the pair of attributes, by comparing the value of SIM_{MN} of each pair of attributes and find a partner with the largest similarity value ($SIM_{MN}MAX$) or a pair of attributes with the similarity value is equal to 1 (SIM_{MN}=1).
7. The model can receive the user verification to the *preliminary result* similarity mapping pair of attributes s.
8. The model can calculate and show the value of the parameter that indicates the effectiveness of the model.

Input documents required by the model include;

1. User name, date of analysis, the type of DBMS, domain of application, and size of DBSource and DBTarget.
2. Information schema document which contains the database name, table names, attribute names, and constraints in DBSource and DBTarget, and instances in DBSource and DBTarget.
3. User verification on the *preliminary result.*

Requirements output of documents generated by the model are as follows;

1. Information about the user, the type of DBMS, database name, database size, table names, attribute names, and constraints and instances in DBSource and DBTarget.
2. Results of the data type conversion according used in the model.
3. The $SIM_{MN}Max$ value for each attributes pair, and the preliminary result and verified result similarity mapping.
4. The test results of the model parameters that are P, R, and F.

3.2. Data Type Conversion

Data type conversion is required to change the data type on the DBMS used by DBSource and DBTarget into new data type used by the model. This process is meant to facilitate the matching process. For example, in the MySQL, data type *char(n)* or *varchar(n)* will be converted into *string*, while the data type *int(n)* or *float(n,d)* will be converted into *numeric*.

3.3. Matching Mechanism and Computing the Similarity of Attribute Pair

Matching mechanism and similarity value calculation carried out at every possible pair attribute in DBSource (AS_i) and in DBSource (AT_i), and every pair will provide a SIM_{MN} value. Each type of constraints (type, width, domain, nullable, unique) and the matching instances will be given a value according to weight that predetermined, whereas if it does not match then it will be assigned a null value. Constraints on DBTarget are the same as constraints on DBSource if both have the same constraints definitions. Meanwhile, the same instance will be stated if the instance in DBTarget appears in DBSource. A matching mechanism and computing the SIM_{MN} is shown in Figure 2.

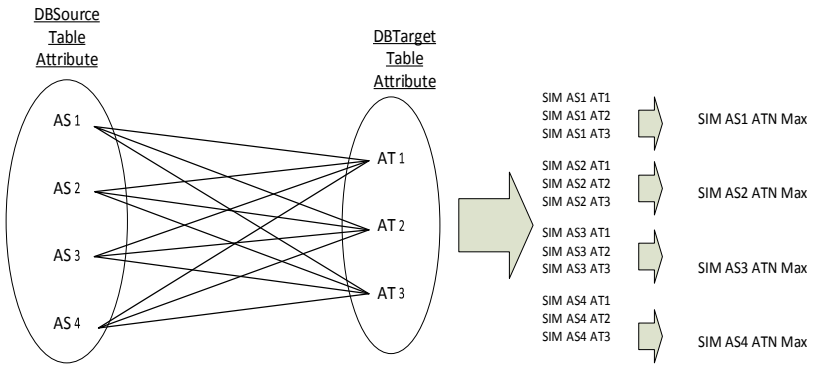


Figure 2. Matching mechanism and computing the similarity value (SIM_{MN})

3.4. Support Database

All of data input, process, and results on the proposed hybrid model will be stored into a relational database model named `dbhybridschematch`. The `dbhybridschematch` consists of 15 tables in the third norm form, which the use of each tables as listed in Table 1. The support database is intended to minimize the computational load, especially during the lasted process of matching.

3.5. Information Schema

Information schema in a database contain all the metadata information of all database objects stored, example for the proposed model has 28 tables in the information schema. Some of the information that can be explored from the information schema and useful in the process of matching schema, such as table, table_constraints, referential_constraints, and statistics. Thus, the proposed model does not use XML as an intermediary language ever developed by [43],[44].

3.6. Constraint and Instance Extraction

A constraint extraction is a process to obtain the data type, width, domain value, nullable, as well as on the unique nature of each attribute in `DBSource` and `DBTarget`. Constraints can be explored from table_constraints in the information schema or directly from each table in the database. In many cases, database designers often are not explicitly defined the constraints, so it will not be found in the information schema and it will be ignored in the matching process. An instance extraction is a process to obtain instances on each attribute in `DBSource` and `DBTarget`. The instance can be explored from each table that is in `DBSource` and `DBTarget`. Normally, the number of instances in each table is equal to the multiplication of the number of records with the number of attributes. However, no guarantee that the value is correct, so it is necessary to find the correct number of instances.

Table 1. Database support for the proposed model

Table Name	Usage
mst_user	Store the user application data
mst_dbms_type	Store the data types of DBMS
mst_data _type_conversion	Store the conversion of origin data types to the data types used by the DBMS in the model
mst_application_domain	Store the types of applications field on DBSource and DBTarget
mst_alt_weight_match	Store the alternative matching criteria weights (type, width, domain, nullable, unique, & instance) which specified by the user
mst_alt_string_size_match	Store the alternative string size difference matching which specified by the user
source_database	Store data about a database on DBSource
source_table	Store data about the tables in DBSource
source_attribute	Store data constraints (type, width, domain, nullable, unique) and instance for each attribute in DBSource
target_database	Store data about a database on DBTarget
target_table	Store data about the tables in DBTarget
target_ attribute	Store data constraints (type, width, domain, nullable, unique) & instance for each attribute in DBTarget
matching_preliminary	Store the preliminary result for all alternative weighting criteria & differences size of string that has not been verified by the user
matching_final	Store the verified result for all alternative weighting criteria & differences size of string that has been verified by the user
matching_report	Store the summary data of preliminary and verified result, & the evaluation of schema matching model

3.7. Computing the Value of Similarity Pair of Attribute (SIM_{MN})

The value of SIM_{MN} for each pair of attributes on `DBSource` and `DBTarget` is determined based on the similarity of the constraints (data type, width, domain value, nullable, unique) and instances. Problems that happen in the process of matching are no limited and very open database designers to specify and define the size of the data in string data type. To overcome it, the proposed model provides features that allow the user to choose an alternative difference data size (width) of the string data types before the matching process is done. Options provided include, *`ALT_1`* (default) the string size of attribute in `DBSource` and `DBTarget` must be exactly the same; *`ALT_2`* the string size of attribute in `DBSource` and `DBTarget` has the difference width is 5; *`ALT_3`* the string size of attribute in `DBSource` and `DBTarget` has the difference width is 15; and *`ALT_4`* the string size of attribute in `DBSource` and `DBTarget` has the difference width is 25. SIM_{MN} value calculation process also faces problems related to the administration of the weight value to each matching criteria. Assuming that the similarity pair of attributes can be specified by constraint or instance only, or both simultaneously, then the proposed model provides features that allow users to select alternative values on the weight of the matching criteria before the calculation is done. By default

(INDEX_1), the weights used in each matching criteria is 0.1 on the constraints (type, width, domain, nullable, unique) and 0.5 on the instance. The values are given with the assumption that the matching process will be done only based on the similarity constraints or instances only. The second alternative (INDEX_2), the weights used in each matching criteria is 0.17. This value is given on the assumption that each criterion has the same role in determining the similarity of attributes.

Different combinations on the choice of string size and weight to the matching criteria will give 8 different results on SIM_{MN} and $SIM_{MN}MAX$ as shown in Table 2. These results will be useful as a material for evaluating the performance of the model and determining the best alternative combinations. SIM_{MN} value is in the range between 0 and 1, where for SIM_{MN}=1 means that the value of an attribute on DBSource match with the attributes on DBTarget, for SIM_{MN} value=0 means the attribute on DBSource not match with the attributes on DBTarget, and to value $0<SIM_{MN}<1$ means that the attributes on DBSource matches with the attributes on DBTarget with similarity level is SIM_{MN}.

Table 2. Combination of string size, index of matching criteria, and similarity value

Alternative of the string size (width)	INDEX_1 (Default)	INDEX_2
ALT_1 **(**Default)	$SIM_{MN}MAX_{11}$	$SIM_{MN}MAX_{12}$
ALT_2	$SIM_{MN}MAX_{21}$	$SIM_{MN}MAX_{22}$
ALT_3	$SIM_{MN}MAX_{31}$	$SIM_{MN}MAX_{32}$
ALT_4	$SIM_{MN}MAX_{41}$	$SIM_{MN}MAX_{42}$

3.8. Preliminary Result, User Verification, and Verified Result

The model developed provides a list of pair of attribute and similarity value generated by the model namely *preliminary result*. Pair of attributes is declared match if it has value SIM_{MN}=1 or $SIM_{MN}MAX$ between each pair of attributes . User verification is done by providing an assessment and then determines whether the results of mapping similarity of each pair of attributes have been as expected. The results of the assessment will give users 4 types of possible values, namely TP (true positive), FP (false positive), FN (false negative), or TN (true negative) as shown in Table 3 [45]-[46]. Verified result of the model is mapping of schema matching results that have been verified by the user, and the values of the parameters P, R, and F which showed the model's performance.

Table 3. The contingency table for examining result of hybrid model schema matching

	Relevant	Non Relevant
Retrieved	True Positive	False Positive
Not Retrieved	False Negative	True Negative

3.9. The Dataset

Hybrid model schema matching will be tested using the test data in the form of a relational database models that meets the heterogeneous nature, form it is has differences in terms of application domains, as well as different DBMS being used. The proposed model is tested on 30 database in relational models that are fulfilled the criteria of heterogeneous, that is, different DBMS platforms (MS Access and MySQL) and different application domains (academic application in higher education and high school, egoverment, and commerce). The largest data capacity is 172,441.6 KB while the smallest is 12.2 KB; the largest table number is 163 while the smallest is 2 tables; the largest number of attributes is 1,642, while fewest is 16; as well as the the largest number of instances is 3,596,857 while fewest is 231, as shown in Table 4. The entire database for testing models derived from survey at 11 institutions, including the universities, government institutions, senior high schools, software developers company, and commercial enterprises.

Table 4. The datasets for testing of proposed hybrid model schema matching

No	Database Name	DBMS Name	Application Domain	Capacity (KB)	Σ Table	Σ Atrribute	Σ Instance
1	db01_sipt_admision	MS Access	HE Academic	75.0	25	193	199,064
2	db02_sipt_academic	MS Access	HE Academic	42.6	69	451	135,319
3	db03_sipt_payroll	MS Access	HE Academic	12.2	16	97	8,827
4	db04_sipt_employ	MS Access	HE Academic	17.8	16	97	6,607
5	db05_sipt_tax_pph	MS Access	HE Academic	1,331.2	10	57	627
6	db06_sipt_research	MS Access	HE Academic	326.6	9	63	3,150
7	db07_sipt_labwork_registration	MS Access	HE Academic	171,056.7	26	162	443,448
8	db08_sipt_library	MS Access	HE Academic	9,932.8	53	435	188,415
9	db09_sipt_menwa_registration	MS Access	HE Academic	144,00	8	42	231
10	db10_nuptk	MySQL	Egoverment	240.0	53	607	1,700,195
11	db11_poor_dss	MySQL	Egoverment	214.0	14	64	429,602
12	db12_office_letter	MySQL	Egoverment	224.1	8	71	710
13	db13_lisence	MySQL	Egoverment	578.5	2	31	6,200
14	db14_lisence_sms	MySQL	Egoverment	172,441.6	140	687	3,596,857
15	db15_dpt_bgcipto	MySQL	Egoverment	79,769.6	4	19	2,721
16	db16_quickcount_bgcipto	MySQL	Egoverment	138,854.4	15	88	7,313
17	db17_dpt_kp	MS Access	Egoverment	76,697.6	7	46	334,270
18	db18_hs_sinisa	MySQL	HS Academic	77,246.0	6	71	2,010
19	db19_hs_sipp	MySQL	HS Academic	656.4	18	151	737,909
20	db20_hs_psb	MySQL	HS Academic	540.6	10	63	564
21	db21_hs_schoolgrade	MySQL	HS Academic	49,049.6	22	190	12,391
22	db22_hs_schoolgrade_online	MySQL	HS Academic	256.2	4	27	567
23	db23_hs_raport	MySQL	HS Academic	1,024.0	44	311	745,655
24	db24_hs_eraport	MySQL	HS Academic	4,558.1	32	233	381,900
25	db25_hs_websma2pwt	MySQL	HS Academic	2,047.5	100	1,642	980,475
26	db26_elearning	MySQL	HS Academic	78.8	163	1,423	163,645
27	db27_elearning_homeschooling	MySQL	HS Academic	1,433.6	105	748	20,205
28	db28_motorcycle_loan	MySQL	Commerce	432.0	10	57	3,879
29	db29_cust_telkomvision	MySQL	Commerce	75.0	5	31	2,916
30	db30_rsmitra_pharmacy	MySQL	Commerce	42.6	14	66	7,453

3.10. Performance Measurement

Evaluation of the model is run to measure the model performance. The evaluation will be run using the parameters *P*, *R*, and *F* obtained from the simulation of prototypes on test data. The values of these parameters are calculated based on the value of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) as the evaluation of performance used in the information retrieval (IR) field research [45]-[46], and then calculated the value of precision (P), recall (R), and f-measure (F) using equation (1) for P, (2) for the R, and (3) to F [7],[15],[21],[39],[42],[47]-[53], that is:

$$Precision = \frac{|TP|}{|TP+FP|} \quad (1)$$

$$Recall = \frac{|TP|}{|TP+FN|} \quad (2)$$

$$F - Measure = \frac{2*Recall*Precision}{Recall+Precision} \quad (3)$$

4. RESULT AND DISCUSSION

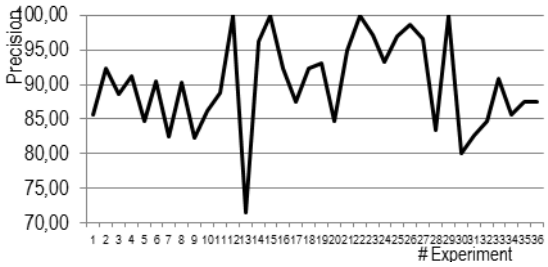
Hybrid model schema matching has been tested for 36 times in pair of `DBSource` and `DBTarget`. Test was performed by using the default matching mechanism that is a combination of `ALT_1` and `INDEX_1`, in three variations of pair of `DBSource` and `DBTarget`. The first test was conducted 30 times in pair of similar `DBSource` and `DBTarget`, the second test was conducted 3 times in pair of `DBSource` and `DBTarget` in the same application domain, and the last test was performed three times in pair of `DBSource` and `DBTarget` in different application domains.

The first experimental step is to read two databases through the import process, it acts as `DBSource` and another as `DBTarget`. If the type of DBMS on `DBSource` and `DBTarget` different from the DBMS used in the model, it is necessary do the data type conversion as described in Section 3.2. The next step, the data constraints are obtained based on the extraction of the information schema in `DBSource` and `DBTarget`, while the database instance is extracted from each of these databases. The process is

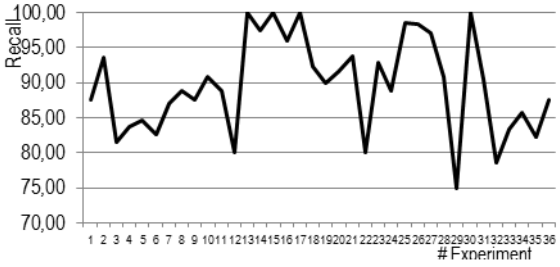
continuing to do matching and calculation SIM_{MN} across all possible pairs of attributes in DBSource and DBTarget, then determining the pair proved a match that which has maximum SIM_{MN} value. The end of this step will generate an output called preliminary result containing pairs of attributes declared matched and the SIM_{MN} value. A preliminary result is verified by a user thereby providing the verified result. The verification process performed on each pair of attributes in the preliminary result. Each verification process generates values of TP, FP, FN, or TN as stipulated in Table 3. Based on such values, it was then computed across values of P, R, and F which indicates the performance of the proposed model.

By using equation (1), (2), and (3), it has been obtained the experimental result values the highest P value was 100.00% while the lowest was 71.43% (Figure 3 (a)); the highest R value was 100.00% while the lowest was 75.00% (Figure 3 (b)); and the highest F-Measure value is 100.00% while the lowest was 81.48% (Figure 3 (c)).

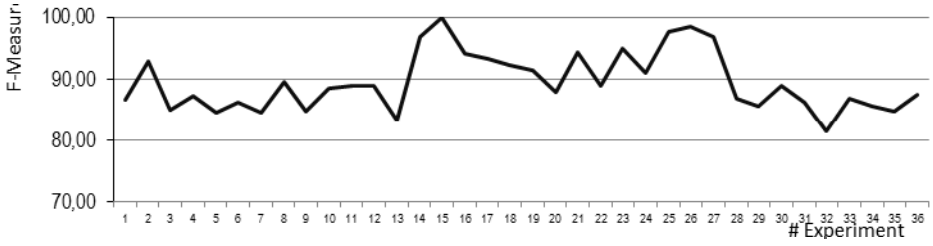
The highest P value was 100% obtained on the four matching experiments on the similar DBSource and DBTarget, namely db12_office_letter, db15_dpt_bgcipto, db22_hs_schoolgrade_online, and db29_cust_telkomvision, while the lowest P value was 71.43% obtained in experiments on the same DBSource and DBTarget at db13_lisence. The highest R value was 100% obtained on the four matching experiments on the similar DBSource and DBTarget, namely db13_lisence, db15_dpt_bgcipto, db17_dpt_kp, and db30_rsmitra_pharmacy, while the lowest R value was 75.00% obtained in matching experiment on the similar DBSource and DBTarget at db29_telkomvision. The highest F value was 100% obtained on the matching pairs on the similar DBSource and DBTarget, namely db15_dpt_bgcipto, while the lowest F value was 81.48% obtained in experiment on the different application domain for DBSource and DBTarget, i.e. on the matching between db02_sipt_academic as DBSource and db08_sipt_library as DBTarget. Based on the experimental results known that errors in the results of hybrid models schema matching occurs in three cases, i.e. the use of an id attribute with data type auto increment; the use of code on data that is defined in the same way (type, width, domain, nullable, unique) but has a different meanings; and if encountered the same instances and the data defined in the same way but actually have different meanings.



(a) Precision



(b) Recall



(c) F-Measure

Figure 3. The experimental results of hybrid models schema matching

Compared with the results of the hybrid schema matching models having been developed previously by [26] which obtained a value of P = 70.00%, R = 75.00%, and F = NA and [27] which obtained a value of P = 90.00%, R = 80.00, and F = 84.00%; it means that the proposed model has fairly good result. To increase the value F-Measure, the proposed model would still be enhanced by providing variation of weighting on constraints, where in general constraint of data type is more dominant as a determinant in common pair of

attributes than constraint of width; and constraint of domain value is more decisive than the constraint of width; whereas constraint of nullable and unique have a similar roles. This research also will be further developed to implement the model into a software prototype by applying all size variations at length of character (`ALT_1`, `ALT_2`, `ALT_3`, and `ALT_4`) and the variety weights used on each matching criteria (`INDEX_2` and `INDEX_2`). The model will be re-evaluated to determine whether there is influence of variety length of data and variety weights used on each matching criteria, in order to know the best variation to obtain the most precise results on the model of schema matching. Improvement efforts are expected to be able to increase the value by F, so that the proposed model can generate similarity on mapping pair of attributes better.

5. CONCLUSION

A hybrid model schema matching by combining the two methods of constraint-based and inctance-based simultaneously has been developed. The model has four main parts, namely *input*, *process*, *output*, and *verification and evaluation*. Based on experiment known that the proposed model has fairly good result, compared with the results of the hybrid schema matching models the have been developed by previously researcher. Errors results on the proposed model occurs in three cases, including use of an id attribute with data type as auto increment; using codes that are defined in the same way (type, width, domain, nullable, unique) but differences meanings; and if encountered in common instances with the same definitions on the attributes but different meaning. Our future work are to providing variation of weighting on constraints and instances, so that the model can generate similarity on mapping attribute pair better.

REFERENCES

[1] He B. and Chang K. C. C., "Statistical schema matching across web query interfaces," *The ACM SIGMOD Int'l Conf. Management of Data. San Diego, CA, USA*, pp. 217-228, 2003. DOI: 10.1145/872757.872784.

[2] Engmann D. and Massmann S., "Instance matching with COMA++. Datenbank Systeme in Business, Technologie und Web," *Procediings of the Model Management & Metadata. Aache, Germany*, pp. 28-37, 2007. URL: http://ceur-ws.org/Vol-814/ om2011 _Tpaper5.pdf.

[3] Villanyi B., *et al.*, "A novel framework for the composition of schema matchers," *The 14th WSEAS Int'l Conf. on Computers, Latest Trends on Computers. Corfu Island, Greece*, pp. 379-384, 2010. URL: http://dl.acm.org/citation .cfm ?id=1981573.1981641.

[4] Velicanu M., *et al.*, "Ways to increase the efficiency of information systems," *The 10th WSEAS Int'l Conf. on Artificial Intelligence, Knowledge Engineering and Databases. Cambridge, UK*, pp. 211-216, 2011. URL: http://www.wseas.us/e-library/conferences/2011/Cambridge/AIKED/AIKED-36.pdf.

[5] Kim W. and Seo J., "Classifying schematic and data heterogeneity in multidatabase systems," *IEEE*, vol/issue: 24(12), pp. 12-18, 1991. DOI: 10.1109/2.116884.

[6] Kavitha C., *et al.*, "Ontology based semantic integration of heterogeneous databases," *European Journal of Scientific Research*, vol/issue: 64(1), pp. 115-122, 2011.

[7] Algergawy A., *et al.*, "Combining effectiveness and efficiency for schema matching evaluation," *The 1st Int'l Workshop on Model-Based Software and Data Integration (MBSDI 2008). Communications In Computer and Information Science (CCIS). Berlin, Germany*, pp. 19-30, 2008. DOI: 10.1007/978-3-540-78999-4_4.

[8] Bernstein P., *et al.*, "The Microsoft repository," *The 23rd Int'l Conf. Very Large Databases (VLDB). Athens, Greece*, pp. 3-12, 1997. DOI: 10.1.1.50.8527.

[9] Bernstein P. A., "Applying model management to classical meta data problems," *The 1st Int'l Conf. Innovative Data Systems Research (CIDR). Asilomar, CA, USA*, pp. 209-220, 2003. DOI: 10.1.1.12.2729.

[10] Stabenau A., *et al.*, "An overview of ensemble," *Genome Research Journal*, vol/issue: 14(5), pp. 929-933, 2004. DOI: 10.1101/gr.1857204.

[11] Bernstein P. A., *et al.*, "Generic schema matching, ten years later," *The VLDB Endowment. Seattle, Washington, USA*, vol/issue: 4(11), pp. 695-701, 2011. URL: http://www.vldb.org/pvldb/vol4/p695-bernstein_madhavan_rahm.pdf.

[12] Do H. H. and Rahm E., "COMA: A system for flexible combination of schema matching approach," *The 28th Conf. on Very Large Data Bases (VLDB). Hong Kong, China*, pp. 610-621, 2002. URL: http://dbs.uni-leipzig.de/file/COMA.pdf.

[13] Do H. H., "Schema matching and mapping-based data integration," *Ph.D. Thesis*. Interdisciplinary Center for Bioinformatics & Dept. of Computer Science, Univ. of Leipzig. Leipzig, Germany, 2005. URL: lips.informatik.uni-leipzig.de/files/2006-4.pdf.

[14] Massmann S., *et al.*, "Evolution of the COMA match system," *The 6th Int'l Workshop on Ontology Matching (OM). Bonn, Germany*, pp. 49-60, 2011. URL: http://ceur-ws.org/Vol-814/om2011_Tpaper5.pdf.

[15] Milo T. and Zohar S., "Using schema matching to simplify heterogeneous data translation," *The 24th Int'l Conf. on Very Large Data Bases (VLDB). NY, USA*, pp. 122-133, 1998. URL: http://www.vldb.org/conf/1998/p122.pdf.

[16] Özsu M. T. and Valduriez P. P., "Principles of distributed database systems," *3rd edition. Pearson Education Inc. NY, USA*, 2011. DOI: 10.1007/978-1-4419-8834-8.

[17] Rahm E. and Bernstein P. A., "A survey of approaches to automatic schema matching," *Very Large Databases (VLDB) Journal*, vol/issue: 10(4), pp. 334-350, 2001. DOI: 10.1007/s007780100057.
[18] Madhavan J., *et al.*, "Generic schema matching with Cupid," *The 27th Int'l Conf. on Very Large Data Bases (VLDB). Roma, Italy*, pp. 49-58, 2001. URL: http://dl.acm.org/citation.cfm?id=645927.672191.
[19] Gross A., *et al.*, "On matching large life science ontologies in parallel," *The 7th Int'l Conf. Data Integration in the Life Sciences (DILS). Gothenburg, Sweden*, pp. 35-49, 2010. DOI: 10.1007/978-3-642-15120-0_4.
[20] Bergamaschi S., *et al.*, "Semantic integration of semistructured and structured data sources," *ACM SIGMOD Record*, vol/issue: 28(1), pp. 54-59, 1999. DOI: 10.1145/309844.309897.
[21] Li W. S. and Clifton C., "Semint: A tool for identifying attribute correspondences in heterogeneous databases using neural network," *Data and Knowledge Engineering Journal*, vol/issue: 33(1), pp. 49-84, 2000. DOI: 10.1016/S0169-023X(99)00044-0.
[22] Hernández M. A., *et al.*, "CLIO: A semi-automatic tool for schema mapping (software demonstration)," *The ACM SIGMOD Int'l Conf. Management of Data. Santa Barbara, CA, USA*, pp. 607, 2001. DOI: 10.1145/376284.375767.
[23] Naumann F., *et al.*, "Attribute classification using feature analysis," *IBM research report. IBM Research Division*. San Jose, CA, USA, 2002. URL: www.hpi.uni-potsdam.de/fileadmin/hpi /FG_Naumann/publications/ ICDE02 Poster.pdf.
[24] Popa L., *et al.*, "Mapping XML & relational schemas with CLIO (software demonstration)," *The Int'l Conf. on Data Enginnering (ICDE). San Jose, CA, USA*, pp. 498-499, 2002. URL: http://disi.unitn.it/~velgias/docs/ PopaHVMNH02.pdf.
[25] Haas L. M., *et al.*, "CLIO grows up: from research prototype to industrial tool," *The ACM SIGMOD Int'l Conf. Management of Data. Baltimore, Maryland, USA*, pp. 805-810, 2005. DOI: 10.1145/1066157.1066252.
[26] Kang J. and Naughton J., "On schema matching with opaque column names & data values," *The ACM SIGMOD Int'l Conf. Management of Data. San Diego, CA, USA*, pp. 205-216, 2003. DOI: 10.1145/872757.872783.
[27] Chien B. C. and He S. Y., "A hybrid approach for automatic schema matching," *The 9th Int'l Conf. on Machine Learning and Cybernetics. Qingdao, China*, pp. 2881-2886, 2010. DOI: 10.1109/ICMLC.2010.5580776.
[28] Li W. S. and Clifton C., "Semantic integration in heterogeneous databases using neural networks," *The 20th Int'l Conf. on Very Large Data Bases (VLDB). Santiago de Chile, Chile*, pp. 1-12, 1994. URL: https://www.cerias.purdue.edu/assets/pdf/bibtex _archive/2001-86-report.pdf.
[29] Li W. S., *et al.*, "Database integration using neural networks: implementation and experiences," *Knowledge and Information Systems Journal*, vol/issue: 2(1), pp. 73-96, 2000. DOI: 10.1007/s101150050004.
[30] Doan A. H., *et al.*, "Reconciling schemas of disparate data sources-a machine-learning approach," *The ACM SIGMOD Int'l Conf. Management of Data. Santa Barbara, CA, USA*, pp. 509-520, 2001. DOI: 10.1145/376284.375731.
[31] Madhavan J., *et al.*, "Corpus-based schema matching," *The IJCAI-03 Workshop on Information Integration on the Web (IIWeb). Acapulco, Mexico*, pp. 59-63, 2003. DOI: 10.1109/ICDE.2005.39.
[32] Rahm E., "Towards large-scale schema and ontology matching," in Bellahsene Z, Bonifati A, Rahm E. *Schema matching and mapping, data-centric systems & applications*. Springer. NY, USA, pp. 3-28, 2011. DOI: 10.1007/978-3-642-16518-4_1.
[33] Dhamankar R., *et al.*, "IMAP: discovering complex semantic matches between database schemas," *The ACM SIGMOD Int'l Conf. Management of Data. Paris, France*, pp. 383-394, 2004. DOI: 10.1145/1007568.1007612.
[34] Bernstein P. A., *et al.*, "Industrial-strength schema matching," *ACM SIGMOD Record*, vol/issue: 33(4), pp. 38-53, 2004. DOI: 10.1145/1041410.1041417.
[35] Dragut E. and Lawrence R., "Composing mappings between schemas using a reference ontology," *The Int'l Conf. on Ontologies, Databases, & Applications of Semantics (ODBASE). Larnaca, Cyprus*, pp. 783-800, 2004. DOI: 10.1007/978-3-540-30468-5_50.
[36] Mork P. and Bernstein P. A., "Adapting a generic match algorithm to align ontologies of human anatomy," *The 20th Int'l Conf. on Data Engineering (ICDE). Boston, Massachusetts, USA*, pp. 787-790, 2004. DOI: 10.1109/ICDE.2004.1320047.
[37] Tu K. W. and Yu Y., "CMC: combining multiple schema-matching strategies based on credibility prediction," *The 10th Int'l Conf. on Database Systems for Advanced Applications (DASFAA). Beijing, China*, pp. 888-893, 2005. DOI: 10.1007/11408079_80.
[38] Jian N., *et al.*, "Falcon-AO: Aligning ontologies with Falcon," *The K-CAP Workshop on Integrating Ontologies (K-CAP'05). Banff, Canada, USA*, pp. 85-91, 2005. DOI: 10.1016/j.websem.2008.02.006.
[39] Jean-Mary Y. R., *et al.*, "Ontology matching with semantic verification," *Web Semantics Journal*, vol/issue: 7(3), pp. 235-251, 2009. DOI: 10.1016/j.websem.2009.04.001.
[40] Sutanta E., *et al.*, "Kajian model dan prototipe schema matching (Studi untuk menemukan peluang pengembangan model dan prototipe baru)," *Prosiding Seminar Nasional Aplikasi Teknologi Informasi (SNATI 2015). Yogyakarta, Indonesia*, pp. J-9-15, 2015. URL: http://journal.uii.ac.id/index.php/Snati/article/viewFile/3556/3147.
[41] Sutanta E., *et al.*, "Survey: Models and prototypes of Schema Matching," *Int'l Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 6(3), 2016. URL: http://www.iaesjournal.com/online/index.php/IJECE/article/view/9789.
[42] Karasneh Y., *et al.*, "Integrating schemas of heterogeneous relational databases through schema matching," *The 11th Int'l Conf. on Information Integration and Web-based Applications and Services (iiWAS). Kuala Lumpur, Malaysia*, pp. 209-216, 2009. DOI: 10.1145/1806338.1806380.

[43] Samini S., *et al.*, "Bridging XML and relational databases: An effective mapping scheme based on persistent," *Int'l Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 2(2), pp. 239-246, 2012. DOI: 10.11591/ijece.v2i2.215.
[44] Win L. H., "XML-based RDF data management for XPath query language," *Int'l Journal of Informatics and Communication Technology (IJ-ICT)*, vol/issue: 2(1), pp. 1-8, 2013. DOI: 0.11591/ij-ict.v2i1.1503.
[45] Manning C. D. and Schutze H., "Foundations of statistical natural language processing," *London The Massachusetts Institute of Technology Press. London, England*, 1999. DOI: 10.1145/601858.601867.
[46] Bellahsene Z., *et al.*, "Schema matching and mapping, data-centric systems and applications," *Springer. New Yor, USA*, 2011. DOI: 10.1007/978-3-642-16518-4.
[47] Rijsbergen C. J. V., "Information Retrieval," *2nd edition. Butterworths, London*, 1979. DOI: 10.1002/asi.4630300621.
[48] Do H. H., *et al.*, "Comparison of schema matching evaluations," *Proceedings of 2nd Int'l Workshop Web & Databases. In: Lecture Notes in Computer Science (LNCS) 2593*. Springer-Verlag, Germany, pp. 221-237, 2003. URL: http://lips.informatik.uni-leipzig.de/files/2002-28.pdf.
[49] Ehrig M. and Staab S., "QOM-quick ontology mapping," *The 3rd Int'l Semantic Web Conf. (ISWC). Hiroshima, Japan*, pp. 683-697, 2004. DOI: 10.1007/978-3-540-30475-3_47.
[50] Giunchiglia F., *et al.*, "A large scale taxonomy mapping evaluation," *The 4th Int'l Conf. Semantic Web Conf. (ISWC). Galway, Ireland*, pp. 67-81, 2005. DOI: 10.1007/11574620_8.
[51] Li J., *et al.*, "RiMOM: A dynamic multistrategy ontology alignment framework," *Journal of IEEE Transaction Knowledge Data Engineering*, vol/issue: 21(8), pp. 1218-1232, 2009. DOI: 0.1109/TKDE.2008.202.
[52] Martinek P., "Schema matching methodologies and runtime solutions in SOA based enterprise application integration," *Ph.D Thesis*. Dept. of Electronics Technology, Budapest University of Technology & Economics. Hungary, 2009. URL: https://repozitorium.omikk .bme.hu/ bitstream/handle/10890/869/ tezis_eng.pdf?sequence=3&isAllowed=y.
[53] Karasneh Y., *et al.*, "An approach for matching relational database schemas," *Journal of Digital Information Management*, vol/issue: 8(4), pp. 260-269, 2010. URL: http://www.dirf.org/jdim/v8i4.asp.

BIOGRAPHIES OF AUTHORS

Edhy Sutanta received Bachelor of Informatics Management & Computer Engineering form IST AKPRIND Yogyakarta, Indonesia in 1996, received Master of Computer Science from Universitas Gadjah Mada, Yogyakarta, Indonesia in 2006. Currently he is a lecturer at Department of Informatics Engineering in IST AKPRIND Yogyakarta Indonesia and pursuing his doctoral program in Computer Science at Department of Computer Sciences & Electronics in Universitas Gadjah Mada, Yogyakarta, Indonesia. His research areas of interest are database systems, database analysis & design, and information systems.
Email : edhy_sst@akprind.ac.id, edhy_sst@yahoo.com

Drs. Retantyo Wardoyo, M.Sc., Ph.D. received Bachelor of Mathematics from Universitas Gadjah Mada, Yogyakarta, Indonesia in 1982, received Master of Computer Science from the University of Manchester, UK in 1990, and received Ph.D. of Computation from University of Manchester Institute of Science and Technology, UK in 1996. Currently he is a lecturer at Department of Computer Science & Electronics in Universitas Gadjah Mada, Yogyakarta, Indonesia. His research area of interest are database systems, operating systems, management information systems, fuzzy logics, and software engineering.

Email : rw@ugm.ac.id

Dr.Techn. Khabib Mustofa, S.Si., M.Kom. received Bachelor of Computer Science from Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia in 1997, received Master of Computer Science from Universitas Gadjah Mada, Yogyakarta, Indonesia in 2001, and receieved Ph.D. from Vienna University of Technology, Austria in 2007. Currently he is a lecturer at Department of Computer Science & Electronics in Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia. His research area of interest are database system**,** semantic web, web engineering, and information management.

Email: khabib@ugm.ac.id

Drs. Edi Winarko, M.Sc., Ph.D. received Bachelor of Statistics from Universitas Gadjah Mada, Yogyakarta, Indonesia in 1996, received Master of Computer Science from Queen's University, Canada, in 2002, and receieved Ph.D. of Computer Sciences from Computer Science School of Informatics and Engineering Flinders University, Australia in 2007. Currently he is a lecturer at Department of Computer Science & Electronics in Universitas Gadjah Mada, Yogyakarta, Indonesia. His research area of interest are data warehousing and data mining, and information retrieval.

Email: edwin@ugm.ac.id