International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 1
ISSN: 2088-5334

Using Variation in Weighting Criteria and String Size Matching on Hybrid Model Schema Matching

Edhy Sutanta[a,1], Erna Kumalasari Nurnawati[a,2], Rosalia Arum Kumalasanti[a,3]

[a]*Department of Informatics Engineering, Institut Sains & Teknologi AKPRIND Yogyakarta, Jl. Kalisahak 28, Yogyakarta, 55222, Indonesia*
Corresponding author: [1]edhy_sst@akprind.ac.id, [2]ernakumala@akprind.ac.id, [3]rosaliaarum@akprind.ac.id

Abstract — **Schema matching plays a vital role in the information integration process from heterogeneous databases. Generally, the process of schema matching is to receive input, which are two databases (one as the source and another as a target), to match similarity attributes, and generate output in the form of mapping the similarity of the attribute pairs that are declared suitable. Furthermore, the user will assess these attribute pairs to determine whether the results obtained are correct or still need to be revised. Our previous study developed a model and software prototype of hybrid schema matching using a combination of constraint-based method and instance-based method. In this study, the model improved by adding new features. This paper discusses the increasing effectiveness of adding the features to customize the weight of matching criteria and string sizes matching. The hybrid model's best effectiveness is obtained when the weight of instance is 0.286, the type is 0.238, width is 0.190, nullable is 0.143, unique is 0.095, and the domain is 0.048. The matching process using a bigger string size increases the model effectiveness with the highest precision of 97.66 when the string size interval is between (length-100) and (length+100). The best combination of weight and string size variation obtains 97.66% precision, a 99.90% recall, and an f-measure of 98.74%.**

Keywords — **Criteria; effectiveness; hybrid schema matching; string size variation; weight variation.**

Manuscript received 26 Aug. 2018; revised 12 Jun. 2020; accepted 17 Nov. 2020. Date of publication 28 Feb. 2021.

I. INTRODUCTION

Integrating all information in an organization is an important process to be a more efficient and effective organization [1]. Schema matching is the main problem in the information integration process [2]. Schema matching plays a vital role in applications that require interoperability from heterogeneous database sources [3], i.e., query mediation and data warehouse [4], or data integration, schema integration, mapping order for e-commerce, and semantic query processing [5]. Schema matching is also required by the users to reveal schema evolution and reuse software [6]. Schema matching is a data integration task performed at the back-end level to solve the problems caused by schematic heterogeneity [7]. The schema matching is a matching process of schema elements to find a similarity between the pairs of attributes [8]. Technically, schema matching is an integration process of the heterogeneous database, which generates generalization or specialization [6]. The cardinality in pairs of similar attributes can form a (1:1) relationship between the local schema and (n:1) or (m:n) on global schema [9].

The models of schema matching can be developed using a single or a combination of methods. Concerning model is developed using a combination of methods, it can be either hybrid or composite. The hybrid model performs multiple methods simultaneously to determine match candidates of attribute pairs based on various criteria or information process [10]-[12]. The composite schema matching is performed method independently collaboratively and combined on the result [9]. The Clio [13]-[16], Cupid [17], SYM [18], as well as [19] are examples of hybrid schema matching. Meanwhile, *composite* combination methods are used in SemInt ([20]-[21]), LSD [9], COMA [22], COMA++ [23], COMA 3.0 ([24]-[25]), IMAP [26], Protoplasm ([27]-[30]), Falcon-AO ([31]-[32]), and ASMOV [33]. Some of the schema matching models using constraint-based methods are combined as a composite with an instance-based method, as performed on SemInt ([11], [20]-[21]), LSD [9], and research by [25]. iMAP uses constraints of schema elements on data types, value range, the uniqueness, the possibility of null values, and foreign key on matching task ([5], [17], [26], [34]). Utilization data types or uniqueness constraints on attributes were applied in [35]-[42]. While, TranScm [12], Autoplex [43], Automatch

[44]-[46], GLUE [47]-[48], SCM [49], and DUMAS [50] are using instance-based method.

Unclear naming on a schema, difficulty finding synonyms names, and differences in language are major problems encountered in the schema matching. These elements cause the schema matching to be not possible to generate the output that is exactly 100% [31]. Since there is no fully automatic possible solution, schema matching should effectively help the user interactively and iteratively solve the matching problem [23]. The efforts to develop a new model or prototype of schema matching are still necessary to find better method combinations that have already existed [34]. According to [5], the use of combination methods should provide a better result and better performance (effectiveness) than the separate execution of a multiple or individual process.

Based on another previous study [51], we have proposed a hybrid schema matching [52]. Our model is combining constraint-based and instance-based methods [34], [53]. Or, based on [54] our model combines three Matchers, i.e., DTM (Data Type Matcher), CM (Constraint Matcher), and IDM (Instance of the Data Matcher). The constraint criteria explored refer to [34], includes data type, width, domain, nullable, and unique, while instance matched by its appearance in the pair of attributes.

The hybrid model of schema matching in [52] still has problems with output effectiveness because each matching criterion in constraint and instance is assumed to have equal weight. These criteria can have different weights when determining the value of similarity (SIM) in the attribute pair. One idea of weighting can be done using the case-based reasoning (CBR) method [55].

Another problem [52] is that the string type's attribute pair will be the same if it has the same size. Every database designer does not have uniformity in determining the size of the string. Some designers define the size strictly according to the data's contents, and some others define very loose data sizes using the maximal size within the string data limit. The main contribution of this paper provides an alternative solution to the two problems. First, modified the model by adding weight variation to matching criteria and the second is adding a variety of the string size during the matching process.

II. Materials and Methods

A. The Datasets

In this study, the datasets are 1) simulation database for testing the logical model validity and weighing the matching criteria, and 2) real database as datasets for testing the model effectiveness. The model's effectiveness was analyzed to find the facts of experimental results. The datasets are the relational database as input to the model, one as DBSource (reference database in the matching process) and the other one is DBTarget (a database to match). The simulation database consists of 4 databases arranged in varying on constraint and or instance. Therefore, it will serve as much as possible to show various possible errors in the model. The simulation database contains predefined code and location data in e-government applications within the Ministry of Home Affairs of The Republic of Indonesia. Each database consists of 3 connections, eight attributes, and a 9,953 *instance*. The datasets for testing include 30 databases that are real data from surveys that matched the criteria (schema, *constraints,* and *instances*) and heterogeneous (based on DBMS software, application domains, and capacity). So, that is worth used to test the model. Based on the DBMS application used, the datasets are consisting of 8 databases, which developed by using Ms Access and 22 using MySQL. Based on the application domains, the datasets are composed of 8 academic colleges, 12 academic databases for Senior High School, eight databases of *e-government* applications, and two databases of *e-commerce*. Based on capacity, the largest dataset is 79,769 kiloByte, contains 204 tables, 1,851 attributes, and 232,893 data items. Whereas the smallest dataset is measuring 115 kiloByte, composed of 1 connection, 16 attributes, and lists 480 data items.

B. The Methods

Our hybrid schema matching model is described as follows. If DMATCH is declaring the result of schema matching process for DS and DT pair, x is the number of attributes in DS, and y is the number of attributes in DT, then;

$$\text{DMATCH}= \{(\text{AS}_1,\text{AT}_1),(\text{AS}_1,\text{AT}_2),..\ (\text{AS}_x,\text{AT}_y)\}$$

If T denotes a type, W declares a width, N denotes nullable, U denotes unique, D denotes domain, I denotes instance, and C is the set of matching criteria of constraint and instance, then;

$$\text{C} = \{\text{T,W,N,U,D,I}\}$$

Suppose SIMT states the value of similarity of T. In that case, SIMI states the value of the similarity of I, SIMW states the value of the similarity of W, SIMN states the value of the similarity of N, the SIMU states the value of similarity U, SIMD states the value of similarity D, SIMI states the value of equality I. The value of similarity of any pair of attributes on the sequence a-th on the DS and the b sequence attribute on the DT, then the values of similarity for each criterion are calculated as follows:

$$\mathrm{SIMT}(\mathrm{AS_a},\mathrm{AT_b}) = \begin{cases}1, \mathrm{T}(\mathrm{AS_a}) = \mathrm{T}(\mathrm{AT_b})\\ 0, \mathrm{other}\end{cases} \quad (1)$$

$$\mathrm{SIMW}(\mathrm{AS_a},\mathrm{AT_b}) = \begin{cases}1, \mathrm{W}(\mathrm{AS_a}) = \mathrm{W}(\mathrm{AT_b})\\ 0, \mathrm{other}\end{cases} \quad (2)$$

$$\mathrm{SIMN}(\mathrm{AS_a},\mathrm{AT_b}) = \begin{cases}1, \mathrm{N}(\mathrm{AS_a}) = \mathrm{N}(\mathrm{AT_b})\\ 0, \mathrm{other}\end{cases} \quad (3)$$

$$\mathrm{SIMU}(\mathrm{AS_a},\mathrm{AT_b}) = \begin{cases}1, \mathrm{U}(\mathrm{AS_a}) = \mathrm{U}(\mathrm{AT_b})\\ 0, \mathrm{other}\end{cases} \quad (4)$$

$$\mathrm{SIMD}(\mathrm{AS_a},\mathrm{AT_b}) = \begin{cases}1, \mathrm{D}(\mathrm{AS_a}) = \mathrm{D}(\mathrm{AT_b})\\ 0, other\end{cases} \quad (5)$$

$$\mathrm{SIMI}(\mathrm{AS_a},\mathrm{AT_b}) = \begin{cases}1, \exists\ \mathrm{I}(\mathrm{AS_a}) = \mathrm{I}(\mathrm{AT_b})\\ 0, other\end{cases} \quad (6)$$

In general, the calculation of the attribute pair's similarity value for any matching criterion in C is:

$$SIM_c(AS_a,AT_b) = \begin{cases}1, x(AS_a) = x(AT_b)\\ 0, other\end{cases} \quad (7)$$

where x(A)= criterion x in A.

If WI denotes weight at I, WT is the weight of T, WW is the weight of W, WN is the weight of N, WU is the weight of U, and WD is the weight at D, the AS_a and AT_b pair's similarity values are calculated as follows:

$$SIM(AS_a, AT_b) = \sum_{y \in C}^{\cdot} SIM_y\,(AS_a, AT_b)\, W_y \qquad (8)$$

The pair of attributes to be declared matched by the model, AT_a that matches the AS_b is taken AT_b according to the following conditions:

$$SIM(AS_a, AT_b) = \underset{z=1}{\overset{m}{Max}}\ SIM(AS_a, AT_z) \qquad (9)$$

The model is then modified by adding weight variations to matching criteria (Alt_Weight) and string-matching size variations (Alt_Length). This modification affects the calculation of SIM values. In general, the calculation of the similarity value for any pair of attributes a-in the DS and the b attribute in the DT values of AS_a and AT_b attribute pair attribute (= SIM') is calculated as follows:

$$SIM'(AS_a, AT_b) = \sum_{j \in C-\{cw\}}^{\cdot} SIM_j\,(AS_a, AT_b)\, W_j + SIM_{cw}(AS_a, AT_b, alt\_length) W_{cw} \qquad (10)$$

$$SIM_{cw}(AS_a, AT_b, alt\_length) = \begin{cases} 1, cw(AS_a) \pm alt\_length = cw(AT_f) \\ 0, other \end{cases} \qquad (11)$$

where Alt_Length is either 0, 25, 50, or 100.

The first step in this study is to do tests to ensure our model's procedure is logically valid. Testing is applied 128 times by using a combination of 16 pairs of simulation databases combined by two weighting variations (Alt_Weight1 and Alt_Weight2) and four variations of string size (Alt_Length1, Alt_Length2, Alt_Length3, and Alt_Length4). At Alt_Weight1 it is assumed the weights on the constraint and instance are the same, each 0.50, so the weight of each criterion in the constraint is 0.10, and the instance weight is 0.50. At Alt_Weight2 it is assumed that all criteria' weights are the same, so each has the same weight of 0.166. Furthermore, the result is compared with the result of a manual process to determine whether the procedure is running correctly.

TABLE I
TEST RESULTS FOR DETERMINING THE WEIGHT RATING OF MATCHING CRITERIA

Matching Criteria	Weight Variation	Results (%)		
		P	R	F
instance	Alt_Weight4	98.83	100.00	99.37
type	Alt_Weight5	98.44	100.00	99.17
width	Alt_Weight6	98.44	100.00	99.14
domain	Alt_Weight7	91.10	100.00	94.53
nullable	Alt_Weight8	98.25	100.00	99.23
unique	Alt_Weight9	97.18	100.00	98.41

The next step is to identify the weight of matching criteria. Each criterion tested for 64 times using a combination of weight valued 1.00 and four variations in the string size. Alt_Weight4 encodes the weight value for type, Alt_Weight5 for width, Alt_Weight6 for a domain, Alt_Weight7 for nullable, Alt_Weight8 for unique, and Alt_Weight9 for instance. Table I shows the test results of each matching criterion for determining the weight rating.

Based on the precision (P) value, the test results are used to determine the matching criteria' weight rank. The test results provide the highest P has given the first position, and so on, the smallest P means having the lowest level. The result obtained then encoded by Alt_Weight3. In another variation, encoded by Alt_Weight1, the weights are assigned based on the assumption the instance has the same weight with a constraint. Thus, the weight of instance is obtained 0.5 and 0.10 on each constraint criteria. While the variation Alt_Weight2, the weights are assigned based on the assumption that each criterion has the same weight, it is 0.167. The values of the matching criteria weight used in each variation are shown in Table II.

TABLE II
VARIATION OF WEIGHT VALUE ON MATCHING CRITERIA FOR MATCHING ATTRIBUTE PAIRS

Weight Variation	Value on Criteria					
	instance	type	width	unique	nullable	domain
Alt_Weight1	0.500	0.10	0.10	0.10	0.10	0.10
Alt_Weight2	0.167	0.167	0.167	0.167	0.167	0.167
Alt_Weight3	0.286	0.238	0.190	0.095	0.143	0.048

The last stage is to conduct effectiveness tests for the model that is done for 384 times by using 32 pairs of databases which are combined by three variations in weight on matching criteria (Alt_Weight1, Alt_Weight2, and Alt_Weight3) and four-string size variations (Alt_Length1, Alt_Length2, Alt_Length3, and Alt_Length4). The string size variations that were used are shown in Table III.

TABLE III
VARIATION OF STRING SIZE FOR MATCHING ATTRIBUTE PAIRS

string size variation	String Size Interval
Alt_Length1	(StringLength - 0) to (StringLength + 0)
Alt_Length2	(StringLength - 25) to (StringLength + 25)
Alt_Length3	(StringLength - 50) to (StringLength + 50)
Alt_Length4	(StringLength - 100) to (StringLength + 100)

The SIM_{CW} values of the pair of attributes were calculated based on 3 variations on weights of matching criteria and 4 string sizes. Furthermore, the model effectiveness is measured using precision (P), recall (R), and F-measure (F) parameters [2]-[3], [11], [23], [32], [56]-[62], which is calculated by equation as follows,

$$P = \frac{|TP|}{|TP| + |FP|} \qquad (12)$$

$$R = \frac{|TP|}{|TP| + |FN|} \qquad (13)$$

$$F = \frac{2x(PxR)}{P+R} \qquad (14)$$

In equation (12) and (13), TP is true positive, FP is false positive, and FN is a false negative.

III. RESULTS AND DISCUSSION

First, we tested the logical validity of the hybrid schema matching model. Testing was done 16 times using a combination of 4 database simulation. The results are shown in Table IV.

TABLE IV
PRELIMINARY RESULTS TESTING OF HYBRID SCHEMA MATCHING

DBSource	DBTarget	Results (Average, %)		
		P	R	F
db_location1	db_location1	100.00	100.00	100.00
db_location1	db_location2	100,00	100.00	100.00
db_location1	db_location3	98.44	100.00	99.17
db_location1	db_location4	100.00	100.00	100.00
db_location2	db_location1	100.00	100.00	100.00
db_location2	db_location2	98.44	100.00	99.17
db_location2	db_location3	98.61	100.00	99.27
db_location2	db_location4	98.44	100.00	99.17
db_location3	db_location1	100.00	100.00	100.00
db_location3	db_location2	98.44	100.00	99.17
db_location3	db_location3	98.21	100.00	99.04
db_location3	db_location4	98.44	100.00	99.17
db_location4	db_location1	98.44	100.00	99.17
db_location4	db_location2	98.44	100.00	99.17
db_location4	db_location4	100.00	100.00	100.00
db_location4	db_location4	100.00	100.00	100.00
Average:		99.12	100.00	99.53

The results of this test obtained the average value of effectiveness parameters, namely P = 100%, R = 99.12%, and F = 99.53%. This result is the same as the values done manually, so it is concluded that the model is logically valid.

The next section highlights the increasing effectiveness of adding the features to customize matching criteria weight and string sizes matching. The brief results of the model tested using 32 pairs of the real database is shown in Table V.

TABLE V
RESULTS TESTING OF HYBRID SCHEMA MATCHING

DBSource	DBTarget	Results (Average, %)		
		P	R	F
db01_sipt_admision	db01_sipt_admision	90.22	100.00	94.80
db01_sipt_admision	db02_sipt_academic	89.63	100.00	94.51
db02_sipt_academic	db03_sipt_payroll	92.36	98.88	95.42
db02_sipt_academic	db04_sipt_employ	92.11	98.25	94.97
db02_sipt_academic	db05_sipt_tax_pph	95.64	99.47	97.48
db02_sipt_academic	db07_sipt_workshop	92.99	100.00	96.23
db02_sipt_academic	db09_sipt_library	89.74	100.00	94.59
db02_sipt_academic	db11_sipt_user	94.73	100.00	97.28
db22_egov_dptkp	db16_lisence	96.73	100.00	98.31
db22_egov_dptkp	db17_lisence_ol	93.31	100.00	96.52
db22_egov_dptkp	db19_egov_dptbgcp	100.00	100.00	100.00
db22_egov_dptkp	db20_quickcount_bgcp	95.96	100.00	97.92
db22_egov_dptkp	db21_egov_dptbtl	100.00	100.00	100.00
db22_egov_dptkp	db22_egov_dptkp	100.00	100.00	100.00
db25_egov_dptkdy	db61_ecomm_rsmitra	96.42	100.00	98.17
db25_egov_dptkdy	db64_ecomm_motorcredit	95.34	100.00	97.59
db30_nuptk	db30_nuptk	99.56	100.00	99.78
db30_nuptk	db32_hs_sinisa	90.37	100.00	94.93
db30_nuptk	db33_hs_sipp	95.29	100.00	97.58
db30_nuptk	db34_hs_psb	98.38	100.00	99.18
db33_hs_sipp	db32_hs_sinisa	99.79	99.97	99.88
db33_hs_sipp	db33_hs_sipp	99.51	100.00	99.75
db33_hs_sipp	db34_hs_psb	92.80	100.00	95.83
db33_hs_sipp	db35_hs_grade	96.26	100.00	98.08
db33_hs_sipp	db36_hsgrade_ol	93.74	100.00	96.74
db33_hs_sipp	db37_hs_report	99.41	100.00	99.70
db33_hs_sipp	db39_hs_sma2pwt	95.50	99.98	97.67
db33_hs_sipp	db41_hs_forum	90.90	100.00	95.23
db33_hs_sipp	db42_hs_announcement	94.23	100.00	96.97
db33_hs_sipp	db43_hs_webinfo	98.32	100.00	99.15
db33_hs_sipp	db44_hs_osis	93.91	100.00	96.84
db33_hs_sipp	db45_hs_elearning	99.13	100.00	99.56
	Average:	95.38	99.89	97.52

Each pair of the database tested for 12 times uses 3 weights of matching criteria and 4 string size variations. Summary of the test results show the effectiveness comparisons based on variety in weight matching criteria presented in Table VI. While Table VII shows the comparison of model effectiveness based on a difference in the string size.

TABLE VI
HYBRID MODEL SCHEMA MATCHING EFFECTIVENESS BASED ON WEIGHT MATCHING CRITERIA VARIATION

Alt_Weight	Results (Average, %)		
	P	R	F
Alt_Weight1	95.03	99.88	97.33
Alt_Weight2	94.97	99.90	97.28
Alt_Weight3	96.16	99.90	97.95

TABLE VII
HYBRID MODEL SCHEMA MATCHING EFFECTIVENESS BASED ON STRING SIZE VARIATION

Alt_Length	Results (Average, %)		
	P	R	F
Alt_Length1	93.55	99.87	96.55
Alt_Length2	94.28	99.89	96.96
Alt_Length3	96.04	99.90	97.84
Alt_Length4	97.66	99.90	98.74

A. Model Effectivenesss Based on Weight Variation on Matching Criteria

Referring to Table VI, the highest P achieved by Alt_Weight3, that is 95.38%, followed by Alt_Weight1 that is 95.03%, and the lowest by Alt_Weight2 that is 94.97%. Comparing the P value on Alt_Weight3 to Alt_Weight1 increase to 1.13%, while comparing to Alt_Weight2 increase by 1.19%.

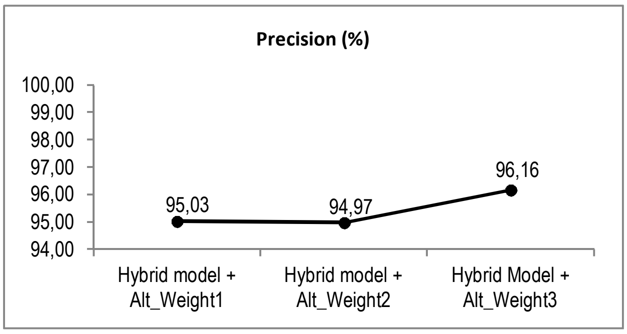

Fig. 1 Hybrid model effectiveness (P) based on the weight variation matching criteria

These indicate the use of an appropriate weight variation on matching criteria can increase the P value. The best variation gain the highest P values was obtained in Alt_Weight3 as shown in Fig. 1. Considers Table VI, the highest R is obtained in Alt_Weight2 and Alt_Weight3. Both of them are in the same value of 99.90%, followed by

Alt_Weight1 that is 99.88%. Rated R on Alt_Weight2 compared with Alt_Weight3 there is no any increase, when compared with Alt_Weight1 there is an increase of 0.02%. These results indicate a variation in weights of matching criteria affects the R-value. The highest R is reached in Alt_Weight2 and Alt_Weight3 as shown in Fig. 2.

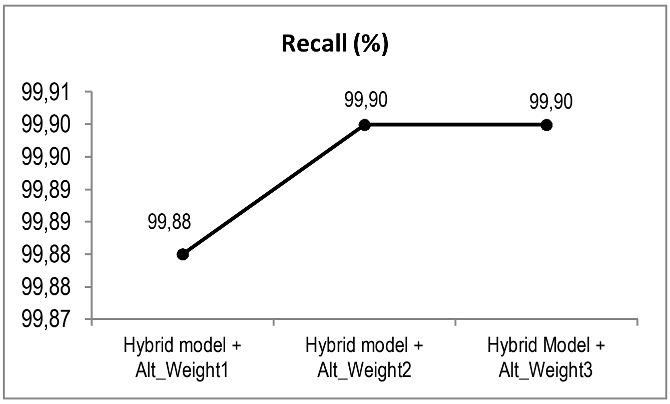

Fig. 2 Hybrid model effectiveness (R) based on the weight variation matching criteria

Based on Table VI, the highest F value obtained in testing by Alt_Weight3 that is 97.95%, followed by Alt_Weight1 that is 97.33%, and the lowest occurred in Alt_Weight2 that is 97.28%. F values at Alt_Weight3 when compared with Alt_Weight1 increased 0.62%, while compared with Alt_Weight2 increased 0.67%. These things indicate a variety of weights on criteria matching effects on the F value. The highest achieved by Alt_Weight3, as shown in Fig. 3.

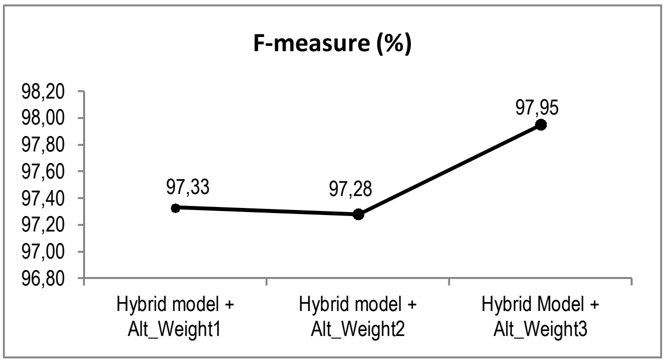

Fig. 3 Hybrid model effectiveness (F) based on the weight variation matching criteria

Based on these results, the average values of the highest P, R, and F are obtained at the Alt_Weight3, such as I = 0.286, T = 0.238, W = 0.190, U = 0.143, N = 0.095, and D = 0.048. The increase was due at Alt_Weight3 and was determined according to rank obtained based on the results of the previous testing and not merely considered as the Alt_Weight1 and Alt_Weight2.

B. Model Effectiveness Based on String Size Variation

Based on Table VII, the highest P obtained at Alt_Length4 is 97.66%, followed by Alt_Length3 that is 96.04%, followed by Alt_Length2, 94.28 and the lowest is in Alt_Length1 of 93.55%. The P value of Alt_Length4 when compared by Alt_Length3 is increasing at 1.62%, in relation to Alt_Length2 there is increasing by 3.38%, and compared to Alt_Length1 is increasing by 4.11%. These indicate the use of a longer string size will increase the P value. The best variation of string size, which the highest P, is obtained on the Alt_Length4, as shown in Fig. 4.

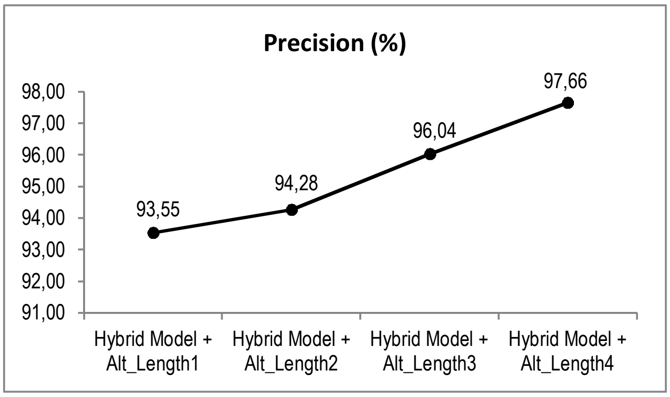

Fig. 4 Hybrid model effectiveness (P) based on the string size variation

Based on Table VII, the highest F obtained in Alt_Length4 is 98.74%, followed by Alt_Length3 that is 97.84%, followed by Alt_Length2, 96.96 the lowest is in Alt_Length1 that is 93.55%. Rated R on Alt_Length4, when compared with Alt_Length3 is not an increase, when compared with Alt_Length2 there is an increase of 0.01%, and Alt_Length1 also occurs an increase of 0.03%. These results indicate a string size variation in a matching process affects the R-value. The highest value is reached in Alt_Length4 and Alt_Length3. The results are shown in Fig. 5.

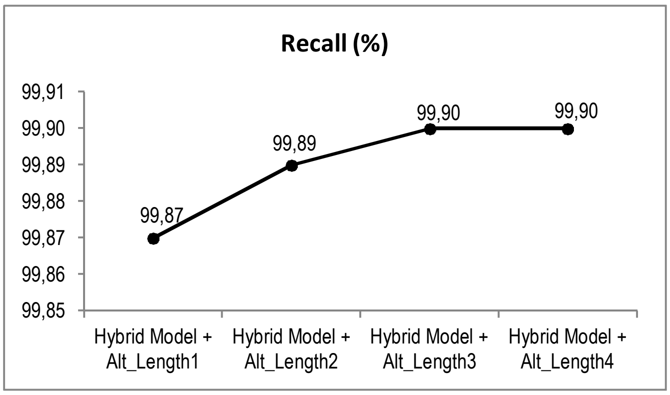

Fig. 5 Hybrid model effectiveness (R) based on the string size variation

Table VII shows that the highest F value is obtained in Alt_Length4 that is 98.74%, followed by testing by using Alt_Length3 that is 97.84%, followed by Alt_Length2 96.96, and the lowest is occurring in Alt_Length1 that is 99.87%. The F value on Alt_Length4, when compared by the same method as Alt_Length3, there is an increase of 0.90%, compared with Alt_Length2 an increase in 1.78%, while comparing with Alt_Length1 there is an increase 2:19%. These results indicate a string size variation affects the F value. The highest F value achieved in Alt_Length4 as shown in Fig. 6.

Based on these results, the highest average values of P, R, and F reach on the weight variation Alt_Length4, and the matching was done by varying the string size (length-100) to (length+100). The effectiveness model increased causes of database designers who they may define size freely so they can describe different size. For example, attributes for the person's name on multiple datasets are defined in various ways, as follows:

- In db01_sipt_admission, admission_name attribute is defined as varchar(50)
- In db02_sipt_academic, the name is defined as varchar(37), while student_name is defined as varchar(100)
- In db03_sipt_payroll, attribute of c_name is defined as varchar(36)
- In db04_sipt_employ, attribute such as employee_name is defined as varchar(150)
- In db05_sipt_taxe_pph, attribute as like emp_name is defined as varchar(55)

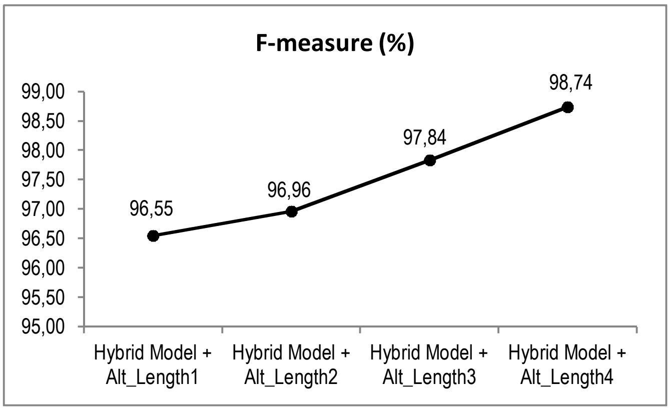

Fig. 6 Hybrid model effectiveness (F) based on the string size variation

The examples above show a name of a person's description in varying sizes. The shortest name defined as 36 characters, the other ways the longest name defined as 150 characters. Suppose that matching on W criteria must use the same size then the value SIMW of the entire matching process will be worth 0. By adding features of string size variation in instance matching, pairing these attributes will likely be worth > 0. It means that there is a possibility to be considered a pair of attributes that matched. This case requires the flexibility of matching criteria on the string size. The use of variation in string size will obtain different effectiveness. The matching by using a bigger of string size will increase the F value. Based on our experiment, the use Alt_Length4 obtain the best results on the precision (P) and the estimated value of the level of effort of adding FN and removes FP (F). In general, the use of a longer string size will provide The better effectiveness of the model. However, variations in the string size still need to be restricted, otherwise as ignoring the `width` criteria. And, it is contrary to the concept of constraint-based method.

IV. Conclusions

Our study shows that using proper weight for the pair attribute matching criteria has increased the effectiveness of the model. The best weighting in hybrid schema matching model is Alt_Weight3, i.e., instance = 0.286, type = 0.238, width = 0.190, unique = 0.143, nullable = 0.095, and domain = 0.048. Additional features of string size variations in certain limits also improve the model of effectiveness. The string size matching of the attribute pairs yields the best effectiveness in Alt_Length4, that is, using a matching string size (length-100) to (length+100). Combining the best weighting and the string size matching obtained the average P value is 97.66%, the R-value is 99.90%, and the F value is 98.74%.

Furthermore, our study will focus on analyzing the effect of adding features the usage a threshold value of SIM associated with the verification process by a user, the similarity checking inter attributes in the database are matched, and the selection of appropriate databases is placed as DBSource or DBSource.

Acknowledgment

The authors would like to thanks the Directorate of Research and Community Service, Directorate General of Research and Development Strengthening at the Ministry of Research, Technology and Higher Education of the Republic of Indonesia under Fundamental Research Grant Contract No.: 109/SP2H/LT/DRPM/2018.

References

[1] D. Sulisworo, Tawar, and U. Ahdiani, "ICT Based Information Flows and Supply Chain in Integrating Academic Business Process," *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, vol. 2, no. 6, p. 454-458, 2012, DOI:10.18517/ijaseit.2.6.243.

[2] H. H. Do, S. Melnik, and E. Rahm, "Comparison of Schema Matching Evaluations," in *The 2nd International Workshop Web and Databases, In Lecture Notes In Computer Science (LNCS) 2593*, Springer-Verlag, Germany, 2002, p. 221-237, DOI: 10.1007/3-540-36560-5_17.

[3] A. Algergawy, E. Schallehn, and G. Saake, "Combining Effectiveness and Efficiency for Schema Matching Evaluation," in *Proceedings of The 1st International Workshop on Model-Based Software and Data Integration (MBSDI 2008)*, vol. 8 (Communications In Computer And Information Science (CCIS), Berlin, Germany, 2008, p. 19-30, DOI: 10.1007/978-3-540-78999-4_4.

[4] L. A. P. P. Leme, M. A. Casanova, K. K. Breitman, and A. L Furtado, "OWL Schema Matching," *Journal of the Brazilian Computer Society*, vol. 16, no. 5, p.21-34, 2010, DOI: 10.1007/s13173-010-0005-3.

[5] E. Rahm and P. A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," *Very Large Databases (VLDB) Journal*, vol. 10, no. 4, p. 334-350, 2001, DOI: 10.1007/s007780100057.

[6] C. Kavitha, G. S. Sadasivam, and S. N. Shenoy, "Ontology-Based Semantic Integration of Heterogeneous Databases," *European Journal of Scientific Research*, vol. 64, no. 1, p. 115-122, 2011.

[7] B. Villanyi, P. Martinek, and B. Szikora, "A Framework for Schema Matchers Composition," *WSEAS Transactions on Computers Journal*, vol. 9, no. 10, p. 1235-1244, 2001, URL: http://dl.acm.org/citation.cfm?id=1865307.1865322.

[8] B. He and K.C.C. Chang, "Statistical Schema Matching Across Web Query Interfaces," in *Proceedings of The ACM SIGMOD International Conference on Management of Data*, San Diego, California, USA, 2003, p. 217-228, DOI: 10.1145/872757.872784.

[9] A. H. Doan, P. Domingos, and A. Y. Halevy, "Reconciling Schemas of Disparate Data Sources-A Machine-Learning Approach," in *Proceedings of The ACM SIGMOD International Conference on Management of Data*, Santa Barbara, California, USA, 2001, p. 509-520, DOI: 10.1145/376284.375731.

[10] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, "Semantic Integration of Heterogeneous Information Sources," *Data and Knowledge Engineering*, vol. 36, no. 3, p. 215-249, 2001, DOI: 10.1016/S0169-023X(00)00047-1.

[11] W. S. Li and C. Clifton, "SEMINT-A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Network," *Data and Knowledge Engineering Journal*, vol. 33, no. 1, p. 49-84, 2000, DOI: 10.1016/S0169-023X(99)00044-0.

[12] T. Milo and S. Zohar, "Using Schema Matching to Simplify Heterogeneous Data Translation," in *Proceedings of The 24th International Conference on Very Large Data Bases (VLDB)*, New York, USA, 1998, p. 122-133, http://dl.acm.org/citation.cfm?id=645924.671326.

[13] M. A. Hernández, R. J. Miller, and L. M. Haas, "CLIO-A Semi-Automatic Tool for Schema Mapping, Software Demonstration," in

Proceedings of The ACM SIGMOD International Conference on Management of Data, Santa Barbara, California, USA, 2001, p. 607, DOI: 10.1145/376284.375767.

[14] F. Naumann, C. T. Ho, X. Tian, L. Haas, and N. Megiddo, "Attribute Classification Using Feature Analysis (Poster)," in *Proceedings of The 18th International Conference on Data Engineering (ICDE)*, San Jose, California, USA, 2002, p. 271, DOI: 10.1109/ICDE.2002.994725.

[15] L. Popa, M. A. Hernández, F. Naumann, Y. Velegrakis, H. Ho, R. J. Miller, "Mapping XML and Relational Schemas with CLIO (Software Demonstration)," in *Proceedings of The International Conference on Data Engineering (ICDE)*, San Jose, California, USA, 2002, p. 498-499, DOI: 10.1109/ICDE.2002.994768.

[16] L. M. Haas, M. A. Hernández, H. Ho, L. Popa, and M. Roth, "CLIO Grows Up: From Research Prototype to Industrial Tool," in *Proceedings of The ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, USA, 2005, p. 805-810, DOI: 10.1145/1066157.1066252.

[17] J. Madhavan, P. A. Bernstein, and E. Rahm, "Generic Schema Matching with CUPID," in *Proceedings of The 27th International Conference on Very Large Data Bases (VLDB)*, Roma, Italy, 2001, p. 49-58, http://dl.acm.org/citation.cfm?id=645927.67219.

[18] B. C. Chien and S. Y. He, "A Hybrid Approach for Automatic Schema Matching," in *Proceedings of The 9th International Conference on Machine Learning and Cybernetics*, Qingdao, China, 2010, p. 2881-2886, DOI: 10.1109/ICMLC.2010.5580776.

[19] J. Kang and J. F. Naughton, "On Schema Matching with Opaque Column Names and Data Values," in *Proceedings of The ACM SIGMOD International Conference on Management of Data*, San Diego, California, USA, 2003, p. 205-216, DOI: 10.1145/872757.872783.

[20] W. S. Li and C. Clifton, "Semantic Integration in Heterogeneous Databases Using Neural Networks," in *Proceedings of The 20th International Conference on Very Large Data Bases (VLDB)*, Santiago de Chile, Chile, 1994, p. 1-12.

[21] W. S. Li, C. Clifton, and S. Y. Liu, "Database Integration Using Neural Networks: Implementation and Experiences," *Knowledge and Information Systems Journal*, vol. 2, no. 1, p. 73-96, 2000, DOI: 10.1007/s101150050004.

[22] H. H. Do and E. Rahm, "COMA-A System for Flexible Combination of Schema Matching Approach," in *Proceedings of The 28th Conference on Very Large Data Bases (VLDB)*, Hong Kong, China, 2002, p. 610-621, DOI: 10.1016/B978-155860869-6/50060-3.

[23] H. H. Do, "Schema Matching and Mapping-Based Data Integration," Interdisciplinary Center for Bioinformatics and Department of Computer Science, University of Leipzig, Leipzig, Germany, *Ph.D. Thesis,* 2005.

[24] E. Rahm, "Schema Matching and Mapping: Towards Large-Scale Schema and Ontology Matching," in *Data-Centric Systems and Applications*, Z. Bellahsene, A. Bonifati, and E. Rahm, New York: Springer, 2011, p.3-27, DOI: 10.1007/978-3-642-16518-4_1.

[25] J. Madhavan, P. A. Bernstein, K. Chen, A. Halevy, and P. Shenoy, "Corpus-Based Schema Matching," in *Proceedings of The IJCAI-03 Workshop on Information Integration on the Web (IIWeb)*, Acapulco, Mexico, 2003, p. 59-63, DOI: 10.1109/ICDE.2005.39.

[26] R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos, "IMAP-Discovering Complex Semantic Matches between Database Schemas," in *Proceedings of The ACM SIGMOD International Conference on Management of Data*, Paris, Franc, 2004, p. 383-394, DOI: 10.1145/1007568.1007612.

[27] P. A. Bernstein, S. Melnik, M Petropoulos, and C Quix, "Industrial-Strength Schema Matching," *ACM SIGMOD Record*, vol. 33, no. 4, p. 38-43, 2004, DOI: 10.1145/1041410.1041417.

[28] E. Dragut and R. Lawrence, "Composing Mappings Between Schemas Using a Reference Ontology," in *Proceedings of The International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE)*, Larnaca, Cyprus, 2004, p. 783-800, DOI: 10.1007/978-3-540-30468-5_50.

[29] P. Mork and P. A. Bernstein, "Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy," in *Proceedings of The 20th International Conference on Data Engineering (ICDE)*, Boston, Massachusetts, USA, 2004, p. 787-790, DOI: 10.1109/ICDE.2004.1320047.

[30] K. W. Tu and Y. Yu, "CMC: Combining Multiple Schema-Matching Strategies Based on Credibility Prediction," in *Proceedings of The 10th International Conference on Database Systems for Advanced Applications (DASFAA)*, Beijing, China, 2005, p. 888-893, DOI: 10.1007/11408079_80.

[31] D. Engmann and S. Massmann, "Instance Matching with COMA++," in *Datenbank Systeme in Business, Technologie und Web (BTW Workshop): Model Management and Metadata*, Aachen, Germany, 2007, p.28-37, https://dbs.uni-leipzig.de/file/BTW-Workshop_2007_EngmannMassmann.pdf.

[32] N. Jian, W. Hu, G. Cheng, and Y. Qu, "Falcon-AO: Aligning Ontologies with Falcon," in *Proceedings of The K-CAP Workshop on Integrating Ontologies (K-CAP'05)*, Banff, Canada, 2005, p. 85-91, http://ceur-ws.org/Vol-156/paper13.pdf.

[33] Y. R. Jean-Mary, E. P. Shironoshita, and M. R. Kabuka, "Ontology Matching with Semantic Verification," *Web Semantics Journal*, vol. 7, no. 3, p. 235-251, 2009, DOI: 10.1016/j.websem.2009.04.001.

[34] P. A. Bernstein, M. Jayant, and E. Rahm, "Generic Schema Matching, Ten Years Later," in *The 33th International Conference on VLDB Endowment*, vol. 4, Seattle, Washington, 2011, p.695-701, http://www.vldb.org/pvldb/vol4/p695-bernstein_madhavan_rahm.pdf.

[35] J. A. Larson, S. B. Navathe, and R. Elmasri, "A Theory of Attribute Equivalence in Databases with Application to Schema Integration," *IEEETrans Software Engineering Journal*, vol. 16, no. 4, p. 449-463, 1989, DOI: 10.1109/32.16605.

[36] S. Hayne and S. Ram, "Multi User View Integration System (MUVIS): An Expert System for View Integration," in *Proceedings of The 6th International Conference Data Engineering (ICDE)*, Los Angeles, California, 1990, p. 402-409, DOI: 10.1109/ICDE.1990.113493.

[37] W. Gotthard, P. C. Lockemann, and A. Neufeld, "System Guided View Integration for Object Oriented Databases," *Journal of IEEE Transaction Knowledge and Data Engineering*, vol. 4, no. 1, p. 1-22, 1992, DOI: 10.1109/69.124894.

[38] S. Spaccapietra and C. Parent, "View Integration: A Step Forward in Solving Structural Conflicts," *IEEE Transaction Knowledge and Data Engineering*, vol. 6, no. 2, p. 258-274, 1992, DOI: 10.1109/69.277770.

[39] B. S. Lerner, "A Model for Compound Type Changes Encountered in Schema Evolution," *ACM Transaction Database Systems*, vol. 25, no. 1, p. 83-127, 2000, DOI: 10.1145/352958.352983.

[40] P. Mitra, G. Wiederhold, and M. Kersten, "Graph-Oriented Model for Articulation of Ontology Interdependencies," in *Proceedings of The 7th International Conference Extending Database Technology (EDBT)*, Konstanz, Germany, 2000, p. 86-100, http://dl.acm.org/citation.cfm?id=645339.650198.

[41] S. Castano, V. D. Antonellis, and S. D. C. di Vimercati, "Global Viewing of Heterogeneous Data Sources," *International Journal of IEEE Transaction Knowledge and Data Engineering*, vol. 13, no. 2, p. 277-297, 2001, DOI: 10.1109/69.917566.

[42] E. Bertino, G. Guerrini, and M. Mesiti, "A Matching Algorithm for Measuring the Structural Similarity between an XML Document and a DTD and Its Applications, Information Systems," vol. 29, no. 1, p. 23-46, 2004, DOI: 10.1016/S0306-4379(03)00031-0.

[43] J. Berlin and A. Motro, "Automatch: Database Schema Matching Using Machine Learning with Feature Selection," in *Proceedings of The 14th International Conference on Advanced Information Systems Engineering (CAiSE '02)*, 2002, Toronto, Ontario, Canada, 2002, p. 452-466, http://dl.acm.org/citation.cfm?id=646090.680403.

[44] J. Berlin and A. Motro, "Autoplex: Automated Discovery of Content for Virtual Databases," in *Proceedings of The 9th International Conference Cooperative Information Systems (CoopIS), In Cooperation with VLDB 2001*, Trento, Italy, 2001, p. 108-122, DOI: 10.1007/3-540-44751-2_10.

[45] D. W. Embley, D. Jackmann, and L. Xu, "Multifaceted Exploitation of Metadata for Attribute Match Discovery in Information Integration," in *Proceedings of the International Workshop on Information Integration on the Web (WIIW'01)*, Rio de Janeiro, Brazil, 2001, p. 110-117.

[46] L. Xu and D. Embley, "Discovering Direct and Indirect Matches for Schema Elements," in *Proceedings of The 8th International Conference on Database Systems for Advanced Applications (DASFAA)*, Kyoto, Japan, 2003, p. 39-46, DOI: 10.1109/DASFAA.2003.1192366.

[47] A. H. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to Map between Ontologies on the Semantic Web," in *Proceedings of The 11th International Conference on World Wide Web (WWW)*, Honolulu, Hawaii, 2002, p. 662-673, DOI: 10.1145/511446.511532.

[48] J. Wang, J. Wen, F. Lochovsky, and W. Ma, "Instance-Based Schema Matching for Web Databases by Domain-Specific Query Probing," in *Proceedings of The 13th International Conference on Very Large Databases (VLDB)*, Toronto, Canada, 2004, p. 408-419, http://dl.acm.org/citation.cfm?id=1316689.1316726.

[49] T. Hoshiai, Y. Yamane, D. Nakamura, and H. Tsuda, "A Semantic Category Matching Approach to Ontology Alignment," in *Proceedings of The 3rd International Workshop Evaluation of Ontology Based Tools (EON)*, Hiroshima, Japan, 2004, p. 67-78, http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-128/EON2004_EXP_Hoshiai.pdf.

[50] A. Bilke and F. Naumann, "Schema Matching Using Duplicates," in *Proceedings of The 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, 2005, p. 69-80, DOI: 10.1109/ICDE.2005.126.

[51] E. Sutanta, R. Wardoyo, K. Mustofa, and E. Winarko, "Survey: Models and Prototypes of Schema Matching," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 3, p.1011-1022, 2016, DOI: 10.11591/ijece.v6i3.pp1011-1022.

[52] E. Sutanta, R. Wardoyo, K. Mustofa, and E. Winarko, "A Hybrid Model Schema Matching Using Constraint-Based and Instance-Based," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 3, p. 1048-1058, 2016, DOI: 10.11591/ijece.v6i3.pp 1048-1058.

[53] M. T. Özsu and P. P. Valduriez, *Principles of Distributed Database Systems, 3rd Edition*, New York: Pearson Education, Inc., Springer, 2011.

[54] Y. Karasneh, H. Ibrahim, M. Othman, and R. Yaakob, "Integrating Schemas of Heterogeneous Relational Databases Through Schema Matching," in *Proceedings of The 11th International Conference on Information Integration and Web-based Applications and Service*, 2009, DOI: 10.1145/1806338.1806380.

[55] M. B. Shuaibu, "Determining an Appropriate Weight Attribute in Fraud Call Rate Data Using Case-Based Reasoning," *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, vol. 4, no. 1, p.34-36, 2014, DOI: 10.18517/ijaseit.4.1.357.

[56] C. J. V. Rijsbergen, *Information Retrieval, 2nd Edition*, London: Butterworths, 1979.

[57] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, London, England: The Massachusetts Institute of Technology Press, 1999.

[58] M. Ehrig and S. Staab, "QOM-Quick Ontology Mapping," in *Proceedings of The 3rd International Semantic Web Conference (ISWC)*, Hiroshima, 2004, p. 683-697, DOI: 10.1007/978-3-540-30475-3_47.

[59] P. Avesani, F. Giunchiglia, and M. Yatskevich, "A Large Scale Taxonomy Mapping Evaluation," in *Proceedings of The 4th International Conference on The Semantic Web Conference (ISWC)*, Galway, Ireland, 2005, p. 67-81, DOI: 10.1007/11574620_8.

[60] J. Li, J. Tang, Y. Li, and Q. Luo, "RiMOM: A Dynamic Multistrategy Ontology Alignment Framework," *IEEE Transaction Knowledge Data Engineering*, vol. 21, no. 8, p.1218-1232, 2009, DOI: 1109/TKDE.2008.202.

[61] P. Martinek, "Schema Matching Methodologies and Runtime Solutions in SOA Based Enterprise Application Integration," Department of Electronics Technology, Faculty of Electrical Engineering & Informatics, Budapest University of Technology and Economics, Hungary, *Ph.D. Thesis*, 2009.

[62] Y. Karasneh, H. Ibrahim, M. Othman, and R. Yaakob, "An Approach for Matching Relational Database Schemas," *Journal of Digital Information Management*, vol. 8, no. 4, p.260-269, 2010, https://dblp.org/rec/bib/ journals/jdim/KarasnehIOY10.