# Business Process Model Similarity Analysis Using Hybrid PLSA and WDAG Methods

Indra Gita Anugrah, Riyanarto Sarno
Department of Informatics, Faculty of Information Technology
Sepuluh Nopember Institute of Technology Surabaya, Indonesia
indra14@mhs.if.its.ac.id, riyanarto@if.its.ac.id

*Abstract* - **Business process modeling means to describe the set of activities either within the companies or organizations. A variety of approaches used and the level of complexity is a problem that often occurs within applicability therefore needed a way to measure the degree of correspondence between the model of Business Process. By measuring the level of compatibility model in Business Process expected company can be easier to analyze. The need in the analysis of business process models are expected of a company or organization can understand the business processes that are running and can be used as a tool to help companies in the face of change and development so as to facilitate in making policies that are needed quickly. In this paper would propose merging structural analysis with semantic analysis where semantic analysis performed using Probabilistic Latent Semantic Analysis (PLSA), and then every method both structural and semantic analysis will be represented into Weighted Directed Acyclic Graph (WDAG) and to calculate, a combined with the aim to generating methods of measuring the degree of correspondence between business process models are better than just using structural analysis.**

***Keywords – similarity process model; topic mining; weighted directed acyclic graph; probabilistic latent semantic analysis; reusable process.***

## I. Introduction

The business process is an instrument to regulate all activities and to improve understanding of linkages between the parts or processes, to run the business processes optimally takes the approach of using Business Process Management (BPM). BPM is a comprehensive approach to improve the effectiveness and efficiency in the business of a company. BPM aims to make the process run optimally, flexible, reusable and fully integrated with information technology, examples of reusable business processes within the company is implementing Enterprise Resource Planning (ERP), one of the challenges of ERP implementation in the enterprise is, the ERP should be able to be reconfigured quickly and flexibly [1]. Application of reusable business processes, can be done by finding the process that has a value of similarity among a collection of other processes, this is done by analyzing every existing processes [2]. Business process overall has a high level of complexity, it makes it difficult to analyze directly, so that the necessary methods to break down the entire business process into several sub process called fragment to make it easier to be analyzed in order to find the sub process that has the same similarities [3].

Various approaches to analyze similarities sub process done, one approach taken by Djikman et. al, to analyze the similarity of the process model can be done by using three approaches: analytical approach to the similarity in syntax, semantics and contextual [4], in addition to approaches syntax, semantics and contextual structural approach can be performed to analyze the similarity of the process. According Djikman et. al. [5] is based on the topology of the model process is represented using a graph, for example, structural analysis conducted by Zhang et. al use Process Structure Tree (PST) [6], some of the above methods sometimes stand alone so that problems arise, for example, when using an analysis of structural in finding the value of similarity model of the process, generating high similarity values are structurally but in reality the process model being compared have different results, of this phenomenon this research approach merger analysis to seek common ground in order to get the value of better accuracy than using a similarity analysis approach alone.

This research proposes the incorporation of similarity between the analysis model business processes by combining structural analysis with textual analysis (string matching), then the combined analysis model represented into Weighted Directed Acyclic Graph (WDAG) and calculate the similarity between the model WDAG. This systematic writing as follows, in Part 1 will discuss the background of the research, part two will discuss the study of literature, section 3 will discuss methods and case studies are used, section 4 will discuss the conclusions of the study, and section 5 is the future work.

## II. Literature Study

### A. Decomposition Process Model

As explained in Part 1, the problems that arise in analyzing the business process model within a company or organization one of them is the level of complexity of the business process itself, in which the business processes of an enterprise or organization composed of many processes that illustrates process linkages existing sections so it will be very difficult if analysis directly to business processes overall, so we need a

978-1-5090-1381-4/16/$31.00 ©2016 IEEE

method to decompose/ break the whole business into several parts called fragments, with the aim of reducing the value of the complexity of a process to be more easily analyzed that are beneficial to the maintenance process and the implementation of reusable business process [7].

Several methods can be used to decompose the process model, including the method proposed by Vanhatalo et. al. of The Refined Process Structured Tree (RPST) [8], the decomposition process based on Two Terminal Graph (TTG), whereby the structure graph that has a connectedness that weakness can be decomposed into sub graph TTG called Triconnected Component, then this method was developed by Polyvyanyy [9] in which the graph structure which has a weak linkages can be decomposed into sub graph MTG (Multi Terminal graph). The decomposition process is aimed to get the unique and modular, the decomposition method using RPST produce fragment Triconnected Component, Triconnected Component some kind fragment is Polygon fragments, Bond fragment, Rigid fragment, and Trivial fragment.

### B. *Abstraction From Behavioral Models*

Behavioral Model developed for a variety of purposes, one of which is used as a reference in the design process and as instruction in the proper execution of a process. Behavioral Model describes the working procedures in detail. Polyvyanyy [10] proposed a method of abstraction from triconnected component to describe the behavior model by applying transformation rules that can be adjusted to the desired needs. Abstraction is the approach used to reduce unwanted details and store important information only. Essential information needed by certain parties in performing its duties, the purpose of abstraction is to provide an overview of the process can be simplified, as decomposition produces triconnected component fragment, then this model is called triconnected abstraction, kind of triconnected abstraction is Polygon abstraction, Bond abstraction, Rigid abstraction, and Trivial abstraction. Figure 1 is an example of transformation rule of Polygon abstraction, abstraction model is used as a rule of fragment decomposition transformation to Weighted Directed Acyclic Graph (WDAG).

### C. *Process Vectors*

Approach to structural analysis can be done using several ways, as described in section 1, the others being the analytical approach used to determine the mining process branching, as research conducted Sarno et. al. [11], using the mining process and workflow branching pattern in the decision making. In this study, structural analysis done by using the method proposed by Jung et. al. [12], wherein the process model is built based on the tuple W = (A, T, Split and Join). Process vector consists of a collection activity, transition, branching Split and Join, a set of activity and the transition should have a value of dependency and control flow that describes the relationship between, for example, the pattern sequence and branching pattern (OR, XOR and AND) [13], where the value of execution probability calculated using equation 1 while the value of dependency is calculated by weighting approach using weighted Completed Dependency Graph (WCDG) and is calculated using equation 2.

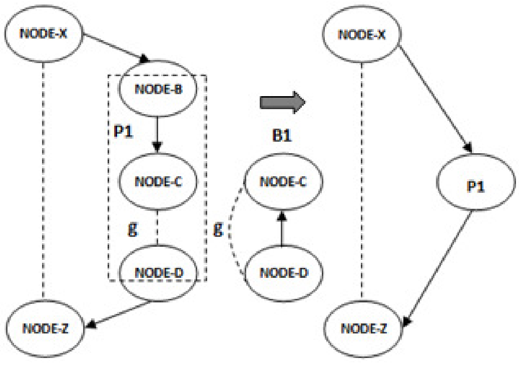


Fig. 1, Transformation Rule From Polygon Abstraction.

$$a_x = (a_{i,x}), a_{ix} = e_{i,x} \tag{1}$$

Where i : 1,..., n, a $_x$ is execution probability activity, e$_{i,\, x}$ is execution probability, *Pr(a)* is Probability activity, s is number of branches.

$$t_x = (t_{ji,x}), t_{ix} = \frac{1}{d_{ijx}} e_{i,x}\, e_{j,x} \tag{2}$$

Where i : 1,..., n, t $_x$ is execution probability transition, t$_{i,\, x}$ is distance weight.

### D. *Probabilistic Latent Semantic Analysis (PLSA)*

PLSA represents a document that involves modeling topics where the original idea of PLSA is the aspect of models built from statistical models, according to Hofmann [14], the aspect of the model is a bridge that connects between the document and the word of a keyword. Aspect model is an invisible variable (latent) of a document. Relationships between documents, topics and words can be represented in Figure 2.

PLSA generate topics and probability value, to obtain the probability value, PLSA algorithm and Expectation Maximization (EM), which consists of two phases:

1. Expectation Step, this phase is used to find the approximate value of the probability of the topics in the document originating from the initial probability value, the stage of expectation can be calculated using equation 3.

$$P(c_i|d_j) = \frac{p(c_i)\prod_{k_i}^{|d_j|} p(c_i)}{\sum_{r=1}^{|c|} \quad p(c_i)\prod_{k_i}^{|d_j|} p(c_i)} \tag{3}$$

Where:

- *P(c$_i$)* is probability of category *c$_i$*,
- $P(c_i|d_j)$ is probability of category $c_i$ in document d$_j$,
- $P(c_i)\prod_{k_i}^{|d_j|} P(c_i)$ is probability of category $c_i$ in term $k_i$ to document $d_j$.

2. Maximization Step, this phase is used to update the probability value so get the maximum probability value, the stage of maximization can be calculated using equation 4.

$$P(w_{kj}|c_i) = \frac{\sum_{r=1}^{|D|} N(w_{kj}, d_j) p(c_i|d_j)}{\sum_{s=1}^{|w|} \sum_{j=1}^{|D|} N(w_s, d_j) p(c_i|d_j)} \quad (4)$$

Where:

- $N(w_{kj}.d_j)$ is number of words $w_k$ in the document $d_j$,
- $|W|$ is total number of words / features used,
- $N(w_{kj}.d_j)$ is total number of training documents.

*E. Similarity of Weighted Directed Acyclic Graph (WDAG)*

Of the graph theory and terminology as we know, it is explained that a set of graphs which have the directions and label in each arc of WDAG. Calculation of similarity of WDAG has similarities with the weighted tree similarity, which use the same calculation of the weighted average of the pair of arc multiplied by the weighted average recursively [15]. In the research we do to represent the model and calculate the value of similarity of our combined analysis using Weighted Directed Acyclic Graph (WDAG). Application of Similarity WDAG one of which is also used to check the similarity ontology as done by Djoko Pramono et. al. [16]. To calculate the similarity of WDAG value is based on research conducted by Sarno et. al. [17], where the similarity algorithm of WDAG influenced by the value of similarity and simplicity, the value of similarity is calculated using equation 5.

$$\sum \begin{cases} \text{if root node label of g and g ' are not identical} \quad (5) \\ \text{if g and g ' is a leaf node} \\ \begin{cases} wDAGsim(g_i, g'_j) \frac{(w_i + w'_j)}{2} \\ wDAGsim(g_i, \varepsilon) \frac{(w_i + 0)}{2} \\ wDAGsim(\varepsilon \ , g'_j) \frac{(0 + w'_j)}{2} \end{cases} \\ \sum_{j-1}^{breadth\ of\ g'} wDAGsim(\varepsilon \ , g'_j) \frac{(0 + w'_j)}{2} \\ \sum^{breadth\ of\ g} wDAGsim(g_i, \varepsilon) \frac{(w_i + 0)}{2} \end{cases}$$

The value of simplicity calculated using equation 6,

$$\begin{cases} D_i \text{ If the root node label of g and g 'are not identical} \quad (6) \\ \frac{D_f}{m} \sum_{j=1}^{|D|} wDAGplicity(g_j) \text{ for g except leaf node} \end{cases}$$

DOKUMEN
(D)
TOPIK
(Z)
TERM
(W)

Fig. 2, The Relationship Between Documents, Topics And Term.

## III. The Proposed Method

In this study, we propose a method which is combining structural and semantic analysis in finding the level of similarity between the business process purpose of increasing the reusability of the process running. The case study we use is the business process of New Student Admission PPDB (Penerimaan Peserta Didik Baru).

*A. New Student Admission : A Case Study*

PPDB business processes that became a case study in this research includes three main processes Pre Registration, Registration and Student Collection, where each main process has wide variety of models and processes. Each model and variation of three main processes will be decomposed into sub-processes and then stored in a repository, where each sub-process will be analyzed to look for similarity values between processes.

*B. Refined Process Structured Tree And SESE*

As explained in section 2A, analysis of the overall business process directly is particularly difficult because it has a high level of complexity, therefore in this case study every model of the process and variation of three main processes will be decomposed using RPST and SESE. For example in Figure 3 is a model of Pre Registration process and in Table 1 are sub processes that results from the decomposition of Figure 3.

*C. WDAG Transformation From Triconnected Abstraction*

In section 2B has explained method of triconnected component abstraction is used to describe behavioral models using transformation rule, as a result of decomposition that produces triconnected fragment RPST then triconnected abstraction component can be used to represent behavioral models of a fragment of decomposition For example in Figure 4 is result of the decomposition of which consists of 5 fragment is P0, P1, P2, B0 and T0, where P is a polygon fragment, B is a B fragment and fragment T is nontrivial. The fragment of triconnected transformed into abstraction, suppose that will be transformed fragment of type polygon will rule polygon transformation abstraction, then be transformed into WDAG form where the number of leaf on WDAG in accordance with the number of sub-processes of each fragment for example in Figure 5. To fragment P0 has three sub-nodes that process, node A, node B and node GA.

*D. Weighting Using Process Vector*

In section 2E., has described the notion WDAG, where on each arc has a label and the weight, so the weight of the WDAG may represent a relationship with a specific amount. In this study, the weight of WDAG used to represent relationships between nodes in a business process, where a node in the business process form of activity or tansition where between activity and transition into business processes

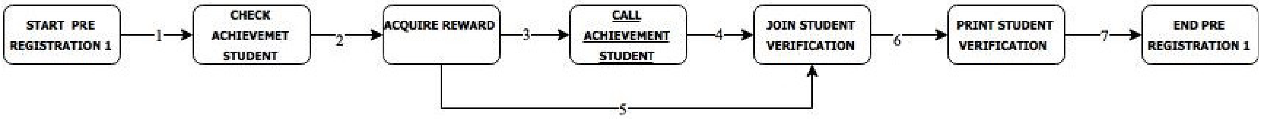


Fig. 3, One Model of The Process From Pre Registration.

have dependency and describes the relationship control flow such that the method proposed by Jung et. al [12] can be used. For example in Table 1 is a table of weighting sub processes on Pre Registration fragment in which the value of the node value obtained by execute the probability and the value of dependency acquired by dependency of the process vector. To get the weight of WDAG, where the value of a branch is 1, then we have normalized so that the weight of WDAG between to 0.

### E. *Textual Analysis Using PLSA*

One approach that can be done to analyze business processes, one of them by performing matching between the label text (string matching) of the node or the edge of the business process being compared. Many ways can be done to make the process of string matching one of them is by using the cosine similarity based on term frequency, but the cosine similarity has a weakness in checking different words but the meaning is the same, for example string: "check student achievement" with "verify student accession" , using the cosine similarity of the string similarity value is 0, in reality meaning the two strings are the same, so that its similarity value close to 1.

From this background, it takes an approach aspect that is latent variable models and serves to connect between documents and words in a keyword. PLSA is a representation of the document by applying aspect models PLSA using algorithms Expectation and Maximization (EM), as can be seen from Table 2 is the value of initial PLSA using algorithms Expectation and Maximization (EM), as can be seen from Table 2 is the value of initial probability of a term derived from equations 3 and 4, so we can calculate the value of similarity string with cosine tfidf 0.167 while using the cosine PLSA 0.783.

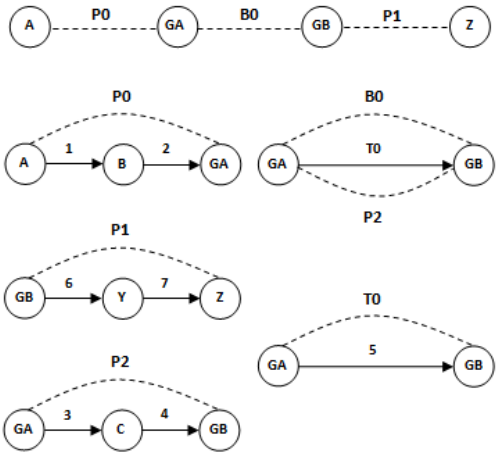


Fig. 4, Result Decomposition Using RPST And SESE From Fig. 3.

TABLE 1

| PROCESS1_PR | | | | |
|---|---|---|---|---|
| **Name of Sub Process** | **Node Value** | **Distance** | **Dependency** | **W WDAG** |
| START PRE REGISTRATION | 1 | | 1 | 0,333 |
| CHECK ACHIEVEMENT STUDENT | 1 | 1 | 1 | 0,333 |
| ACQUIRE REWARD | 1 | | 1 | 0,333 |
| | | | **3** | **1** |
| **PROCESS2_PR** | | | | |
| **Name of Sub Process** | **Node Value** | **Distance** | **Dependency** | **W WDAG** |
| ACQUIRE REWARD | 1 | | 1 | 0,4 |
| CALL ACHIEVEMENT STUDENT | 0,5 | 1 | 0,5 | 0,2 |
| JOIN STUDENT VERIFICATION | 1 | | 1 | 0,4 |
| | | | **3** | **1** |
| **PROCESS3_PR** | | | | |
| **Name of Sub Process** | **Node Value** | **Distance** | **Dependency** | **W WDAG** |
| JOIN STUDENT VERIFICATION | 1 | | 1 | 0,333 |
| PRINT STUDENT VERIFICATION | 1 | 1 | 1 | 0,333 |
| END PRE REGISTRATION | 1 | | 1 | 0,333 |
| | | | **3** | **1** |

TABLE 2

| **NO** | **TERM** | **TOPIC PROBABILITY** | | |
|---|---|---|---|---|
| | | **PR** | **R** | **SC** |
| 4 | ACCESSION | 0,037 | 0 | 0 |
| 73 | CHECK | 0,020 | 0,020 | 0,047 |
| 85 | STUDENT | 0,082 | 0 | 0,136 |
| 87 | ACHIEVEMENT | 0,044 | 0 | 0 |
| 93 | VERIFY | 0 | 0,010 | 0,004 |

### F. *WDAG Similarity*

WDAG similarity calculate the value of similarity of process models, to calculate the value of similarity of WDAG can be done using equation 5, and to calculate the value of simplicity of WDAG can be done using equation 6, In Figure 6 is a representation of a fragment WDAG Pre Registration 1 and Pre Registration 2, so the value of similarity can be calculated as follows:

$$sim(DAG5A,\ DAG5B):\ \frac{(0\times 1)+(0\times 0.667)}{2}\times 0 = 0$$

$$plicity(DAG5A,\ DAG5B):0.5\times 0.3333 = 0.1665$$

$$sim(DAG5A,\ DAG5B):\ \ 0+(0.5\times 0.1665) = 0.08325$$

$$sim(DAG3A,\ DAG3B):$$

$$\frac{0.4+0.4}{2}\times 1+\frac{0.0165+0.0165}{2}\times 1+\frac{0.4+0.4}{2}\times 1$$
$$= 0.817$$

$$sim(DAG2A,\ DAG2B):$$

$$\frac{0.333+0.333}{2}\times 1+\frac{0.333+0.333}{2}\times 0+\frac{0.333+0.333}{2}\times 0$$
$$= 0.333$$

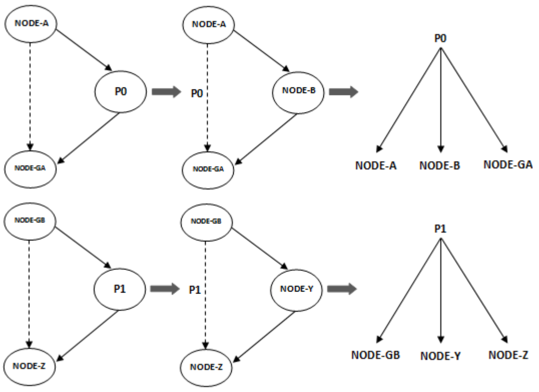
$$sim(DAG4A,\ DAG4B):$$

$$\frac{0.333+0.333}{2}\times 0+\frac{0.333+0.333}{2}\times 1+\frac{0.333+0.333}{2}\times 0$$
$$= 0.333$$

$$sim(DAG1A,\ DAG1B):$$

$$\frac{0.111+0.111}{2}\times 1+\frac{0.271+0.271}{2}\times 1+\frac{0.111+0.111}{2}\times 0$$
$$= 0.382$$

From the calculation using above pure WDAG similarity values obtained for 0.382, while using WDAG PLSA obtained similarity value of 0.78. Calculation WDAG similarity above shows that the similarity of the fragment Pre Registration 1 with fragment Pre Registration 2 amounted to 0.78, where the closer value 1, then similarity of the model has evaluated have many suitability. In the application process reusability, shows that more same similarities between the process then it will be easier to do reusability process.



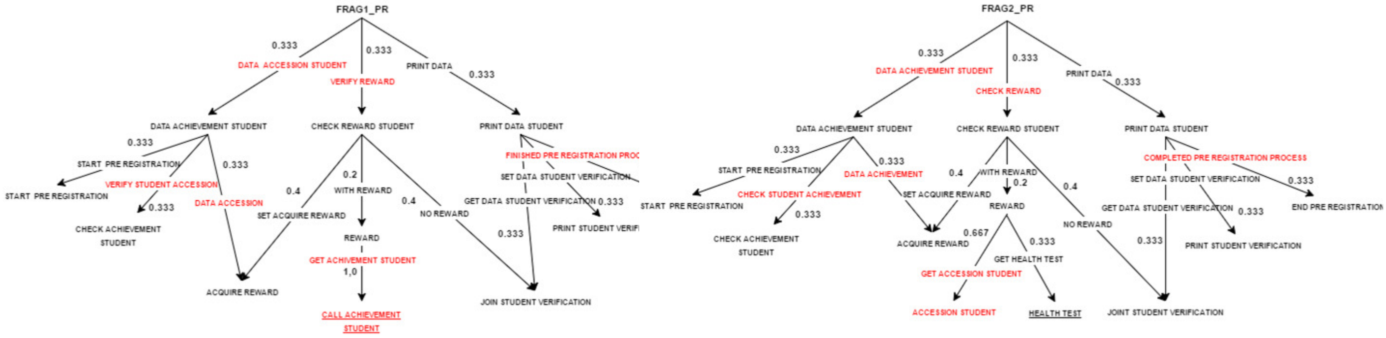
a. Triconnected Abstraction b. WDAG Transformation

Fig. 5, Representation WDAG Using Triconnected Abstraction.

For example in Table 3 is the result of the calculation similarity of some models Pre Registration, from Table 3 rank obtained similarity value of the model Pre Registration 2 (PR2) is PR3, PR4, PR5 and PR1, from these results, ranked by similarity values can be seen the largest to the smallest variant suitable for models Registration Pre 2, as research by Setiawan et. al [18], in the selection of variant web service is done by selecting the greatest ranked based on criteria that fit.

TABLE 3

| WDAG | | SIMILARITY | |
|---|---|---|---|
| MODEL 1 | MODEL 2 | WDAG PURE | WDAG PLSA |
| PR2 | PR1 | 0,3828 | 0,7747 |
| PR2 | PR3 | 0,3793 | 0,7928 |
| PR2 | PR4 | 0,3804 | 0,7899 |
| PR2 | PR5 | 0,3805 | 0,7881 |



(a) Pre Registration 1

(b) Pre Registration 2

Fig. 6, Representation WDAG From Fragment Pre Registration.

# IV. Evaluation

Evaluation used in the system that we have built in measuring the value similiaritas of the business process is to perform a test to measure the value of similarity between the fragment in the same model, for example fragment models Pre Registratin compared with fragment variations, and the value of similarity between the fragment in a different class, suppose the fragment is in a class Pre Registration compared to fragments that are at or Student Registration fragment Collection then we evaluate the test results using the confusion matrix to measure the value of his accuracy.

From our evaluation, we get value of accuracy of the method of analysis that we submitted using WDAG PLSA and WDAG Pure, and evaluation of results we get that value accuracy WDAG PLSA (0.45) while the value WDAG Pure accuracy (0.05), indicating that WDAG PLSA has better accuracy than WDAG Pure in getting value similarity on case study data for new students. whereas for textual similarity evaluation using a dataset of label string PPDB business process model, we get the results of the methods Cosine Vector Space Model (0.548), Cosine Latent Semantic Analysis (0.5) and Cosine Probabilistic Latent Semantic Analysis (0.595).

# V. Conclusion

From the research that we do can be concluded, by using the combined analysis of structural and textual could improve accuracy values than simply using structural analysis alone, the accuracy of WDAG PLSA (0.45) is better than the WDAG Pure (0.05). WDAG we use to analyze the structure of business processes as well as the representation for the combined analysis method we proposed, where as we use to check the PLSA Label String node and edge of the structure of business processes. PLSA method can check that the word has the same meaning (synonim) during the process of String Matching. PLSA will show a better value compared to methods VSM and LSA in the words that are in the same topic. Some things that affect the value of probabilistic PLSA which set of words (parts of words) that are used in the training phase to get the value of probabilistic words topic.

## Future Work

For future work, the author would like to improve the combination analysis method proposed and applied this method to measure the similarity of web service used for selecting the right service.

## References

[1] Sarno, R., Djeni, C. A., Nukhlas, I., & Sunaryono, D. (2015). Developing A Workflow Management System For Enterprise Resource Planning. Journal of Theoretical & Applied Information Technology, 72(3).

[2] Sarno, R., Ginardi, H., Pamungkas, E. W., & Sunaryono, D. (2013, November). Clustering of ERP business process fragments. In Computer, Control, Informatics and Its Applications (IC3INA), 2013 International Conference on (pp. 319-324). IEEE.

[3] Sarno, R., Wibowo, W. A., Sunaryono, D., & Munif, A. (2015, October). Developing workflow patterns based on functional subnets and control-flow patterns. In 2015 International Conference on Science in Information Technology (ICSITech) (pp. 24-29). IEEE.

[4] Dijkman, R., Dumas, M., Van Dongen, B., Käärik, R., & Mendling, J. (2011). Similarity of business process models: Metrics and evaluation. Information Systems, 36(2), 498-516.

[5] Dijkman, R., Dumas, M., & García-Bañuelos, L. (2009, September). Graph matching algorithms for business process model similarity search. In International Conference on Business Process Management (pp. 48-63). Springer Berlin Heidelberg.

[6] Ling, J., Zhang, L., & Feng, Q. (2014). An Improved Structure-based Approach to Measure Similarity of Business Process Models. In SEKE (pp. 377-380).

[7] Anugrah, I. G., Sarno, R., & Anggraini, R. N. E. (2015, September). Decomposition using Refined Process Structure Tree (RPST) and control flow complexity metrics. In Information & Communication Technology and Systems (ICTS), 2015 International Conference on (pp. 203-208). IEEE.

[8] Vanhatalo, J., Völzer, H., & Koehler, J. (2009). The refined process structure tree. Data & Knowledge Engineering, 68(9), 793-818.

[9] Polyvyanyy, Artem, Structuring Process Models, 2012, University of Potsdam

[10] Polyvyanyy, A., Smirnov, S., & Weske, M. (2015). Business process model abstraction. In Handbook on Business Process Management 1 (pp. 147-165). Springer Berlin Heidelberg.

[11] Sarno, R., Sari, P. L. I., Ginardi, H., Sunaryono, D., & Mukhlash, I. (2013, November). Decision mining for multi choice workflow patterns. In Computer, Control, Informatics and Its Applications (IC3INA), 2013 International Conference on (pp. 337-342). IEEE.

[12] Jung, J. Y., Bae, J., & Liu, L. (2009). Hierarchical clustering of business process models. International Journal of Innovative Computing, Information and Control, 5(12), 1349-4198.

[13] Cardoso, J. (2008). Business process control-flow complexity: Metric, evaluation, and validation. International Journal of Web Services Research, 5(2), 49.

[14] Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc.

[15] Sarno, R., Yang, L., Bhavsar, V. C., & Boley, H. (2003). The AgentMatcher architecture applied to power grid transactions. In Proceedings of the First International Workshop on Knowledge Grid and Grid Intelligence, Halifax (pp. 92-99).

[16] Pramono, D., Setiawan, N. Y., Sarno, R., & Sidiq, M. (2013). Physical activity recommendation for diabetic patients based on ontology. In 7th International Conference on Information & Communication Technology and Systems (pp. 27-32).

[17] Sarno, R., Ghozali, K., Nugroho, B. A., & Hijriani, A. (2011). Semantic Matchmaking using Weighted Directed Acyclic Graph. In Proceedings of the third International Seminar on Applied Technology, Science and Arts (APTECS) (pp. 329-334).

[18] Setiawan, N. Y., Sarno, R. (2016). Multi-Criteria Decision Making For Selecting Semantic Web Service Considering Variability And Complexity Trade-Off. Journal of Theoretical & Applied Information Technology 86(2) (pp. 316-326).