# Development of Distance Measures for Process Mining, Discovery, and Integration

Joonsoo Bae[1], Ling Liu[2], James Caverlee[3], Liang-Jie Zhang[4], Hyerim Bae[5]
[1] Chonbuk National Univ, South Korea
jsbae@chonbuk.ac.kr
[2] Georgia Institute of Technology, USA
lingliu@cc.gatech.edu
[3] Georgia Institute of Technology, USA
caverlee@cc.gatech.edu
[4] IBM T.J. Watson Research Center, USA
zhanglj@us.ibm.com
[5] Pusan National Univ, South Korea
hrbae@pusan.ac.kr

## ABSTRACT:

Business processes continue to play an important role in today's service-oriented enterprise computing systems. Mining, discovering, and integrating process-oriented services has attracted growing attention in the recent years. In this paper we present a quantitative approach to modeling and capturing the similarity and dissimilarity between different process designs. We derive the similarity measures by analyzing the process dependency graphs of the participating workflow processes. We first convert each process dependency graph into a normalized process matrix. Then we calculate the metric space distance between the normalized matrices. This distance measure can be used as a quantitative and qualitative tool in process mining, process merging, and process clustering, and ultimately it can reduce or minimize the costs involved in design, analysis, and evolution of workflow systems.

## KEY WORDS:

*Business Process, Similarity, Process Mining*

# 1. Introduction

With the increasing interest and wide deployment of web services, we see a growing demand for service-oriented architectures and technologies that support enterprise transformation. Effective enterprise transformation refers to strategic business agility in terms of how efficiently an enterprise can respond to its competitors and how timely an enterprise can anticipate new opportunities that may arise in the future. In the increasingly globalized economy, enterprises face complex challenges that can require rapid and possibly continual transformations. As a result, more and more enterprises are focused on the strategic management of fundamental changes with respect to markets, products, and services (Rouse, 2005). Such transformation typically has a direct impact on the business processes of an enterprise. Enterprise transformation may range from traditional business process improvement to wholesale changes to the processes supported by the enterprise – from performing current work in a new fashion to performing different work altogether. Each of these challenges may lead to a different degree of enterprise transformation.

Fundamental to enabling the transformation of an enterprise is the development of novel tools and techniques for transforming the business processes of an enterprise. In this paper, we present a critical component to the problem of process transformation from a web services point-of-view. In particular, we present a novel process difference analysis method using distance measures between process definitions of two transactional web services. The process difference analysis focuses on process activity dependencies and process structure to identify distance measures between processes.

The proposed difference analysis method achieves three distinct goals. First, by analyzing the attributes of process models, we present a quantitative process similarity metric to determine the relative distance between process models. This facilitates not only the comparison of existing process models with each other, but also provides the flexibility to adapt to changes in existing business processes. Second, the proposed method is quick and flexible, which reduces the cost of both the analysis and design phases of web service processes. Third, the proposed method enables the flexible deployment of process mining, discovery, and integration – all key features that are necessary for effective transformation of an enterprise.

## 2. Web Service Process Reference Model

The web service process reference model consists of business process definitions and the specification of workflows among the processes with respect to data flow, control flow, and operational views (Rush, 1997; Schimm, 2004). We define a business process in terms of business activity patterns. An activity pattern consists of objects, messages, message exchange constraints, preconditions and postconditions (WfMC, 2005), and is designed to specify the service actions and execution dependencies of the business process. An activity pattern can be viewed as a web service process when it is executable as a web service. We consider two types of activity patterns – elementary activity patterns and composite activity patterns (Aalst, 2003a; Bae, 2004). An elementary activity pattern is an atomic unit. A composite activity pattern consists of a one or more elementary activity patterns or other composite activity patterns. The dependencies could capture complex interactions between activities.

We define a business process as a collection of business activities connected by data flow and control flow, where each represents a business process. A process definition can be seen as a web service (or a collection of web services). We use data flow among processes to define the data dependencies among processes within a given business process. We use control flow to capture the operational structure of the business process service, including the process execution ordering, the transactional semantics and dependencies of the process. A number of workflow specifications have gathered attention, including BPEL4WS (BEA, IBM, Microsoft), WSFL (IBM), XLANG (Microsoft), and XPDL (WfMC) (WfMC, 2005). In our prototype development, we choose to use a variant of BPEL4WS.

Formally, each workflow service is specified in terms of process definitions. We can model each process definition as a process model using activities, precedence relation between activities, and their properties.

**Definition 1 (Process Model, PM)**
A process model *PM* consists of tasks, links, and attributes. That is, *PM*= <*A*, *L*, *Attr*>.

- A set of activities: $A = \{a_i \mid i= 1, \dots, I\}$, where, $a_i$ represents *i*-th activity and *I* is the total number of tasks in a process.
- A set of links: $L = \{l_k= (a_i, a_j) \mid a_i, a_j$ □$A, i\neq j \}$, where, $l_k$ represents a link between two activities, $a_i$ and $a_j$. A link also represents a precedence relation. The link $(a_i, a_j)$ indicates that $a_i$ immediately precedes $a_j$.
- A set of attributes: *Attr* is a set of attributes ($attr_l$), whose element represents feature of objects such as process, activity and link. An attribute of an object is represented using

common dot(.) notation. For example $a_i.attr_{Name}$ represents *name* attribute of activity $a_i$. ■

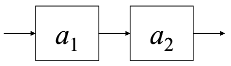
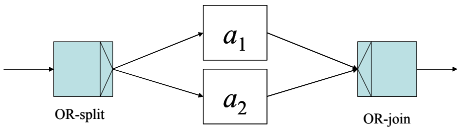
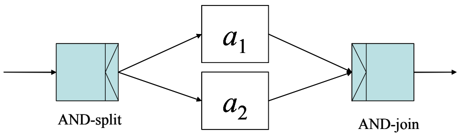
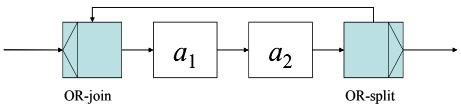
In our process model definition, structural information is specified using activities and links. All the other information related to time properties, business logic, correctness, and split/merge pattern is assumed to be presented with attributes. In order to execute the process model after being designed, it should be in a computer readable format. We store process models in an XML format, and they can be exported into BPEL4WS codes automatically, to be accessible via web services.

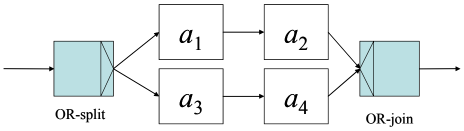
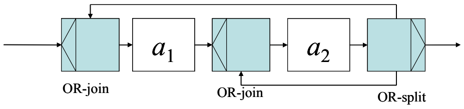
As a real-life example of business process, there are many PIPs (Partner Interface Processes) as defined by RosettaNet (RosettaNet). PIPs define business processes between trading partners. PIPs fit into seven Clusters, or groups of core business processes, that represent the backbone of the trading network. Each Cluster is broken down into Segments and within each Segment are individual PIPs. RosettaNet standards provide the infrastructure for integrating business processes with trading partners across the globe, delivering essential value to industries and proven real-world business results. Fig. 1 shows a standard process of procurement order by buyer, which is in Segment 3A(Quote and Order Entry) of Cluster 3(Order Management). This example process has 13 activities, 22 links, and many attributes, which can be presented with our formal model in the following.

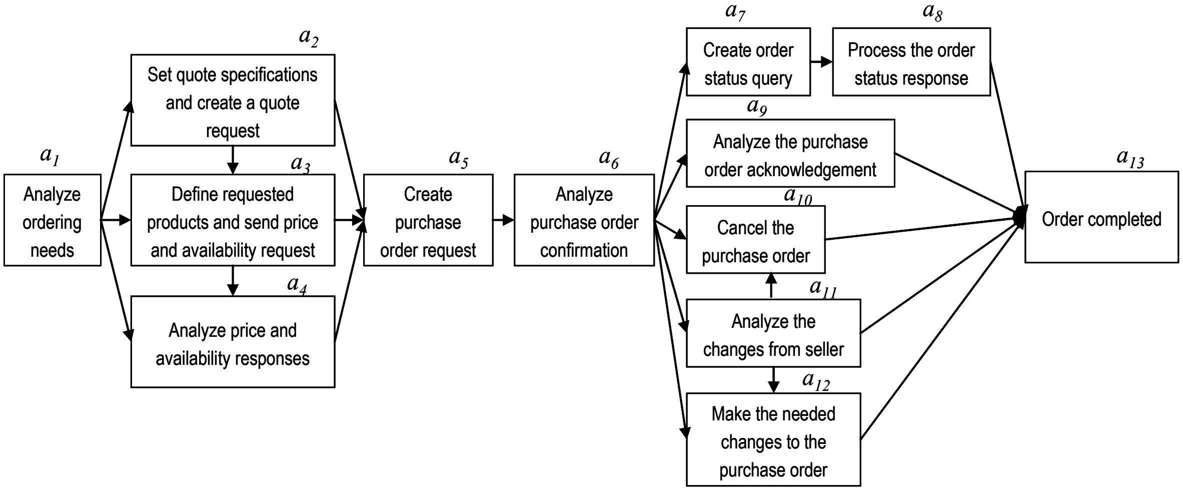
$$
\begin{aligned}
A &= \{a_1, a_2, a_3, \ldots, a_{13}\} \\
L &= \{l_1, l_2, l_3, \ldots, l_{22}\} = \{(a_1, a_2), (a_1, a_3), (a_1, a_4), \ldots, (a_{12}, a_{13})\} \\
Attr &= \{a_1.attr_{\text{TaskName}}(= \text{“Analyze ordering needs”}), a_2.attr_{\text{TaskName}}, \ldots, \\
&\quad a_1.attr_{\text{ExpTime}}, \ldots, l_1.attr_{\text{TransCond}}, \ldots\}
\end{aligned}
$$

Recent business environments impel enterprises to interface with each other, and SOA (Service Oriented Architecture) is considered as a natural tool for B2B (Business to Business) collaboration. For our model to be used in such computing environments, we transform our process model into XML based language, that is, BPEL4WS codes. Rules used for our transformation are summarized in Table 1.

Table 1. Rules for transforming process model into BPEL4WS codes

| Pattern | Graph | (Structured) BPEL4WS |
|---|---|---|
| Sequence |  | <sequence> <$a_1$> <$a_2$> </sequence> |
| Parallel Flow |  | <switch><br><case condition=”condition”><$a_1$></case><br><case condition=”condition”><$a_2$></case><br></switch> |
| |  | <flow> <$a_1$> <$a_2$></flow> |
| Loop |  | <loop condition=”condition”> <$a_1$> <$a_2$></loop> |

| | | |
|---|---|---|
| Switch-sequence |  | <switch><br><case condition=”condition”> <sequence><br><$a_1$><$a_2$></sequence> </case><br><case condition=”condition”> <sequence><br><$a_3$><$a_4$></sequence></case><br></switch> |
| Complex Flow |  | <sequence><$a_1$><$a_2$></sequence><br><loop><br><switch><br><case><$a_1$></case><br></switch><br><$a_2$><br></loop> |



$a_1$
Analyze ordering needs
$a_2$
Set quote specifications and create a quote request
$a_3$
Define requested products and send price and availability request
$a_4$
Analyze price and availability responses
$a_5$
Create purchase order request
$a_6$
Analyze purchase order confirmation
$a_7$
Create order status query
$a_8$
Process the order status response
$a_9$
Analyze the purchase order acknowledgement
$a_{10}$
Cancel the purchase order
$a_{11}$
Analyze the changes from seller
$a_{12}$
Make the needed changes to the purchase order
$a_{13}$
Order completed

Fig. 1 A real-life example of business process

# 3. Process Dependency Graph

From a process model, we can extract a graph which presents dependencies among activities. We call the graph ‘Dependency Graph’ in Definition 2. The process dependency graph captures information about how activities share information and how data flows from one activity to another. Depending on whether the edges indicate execution dependencies or data flow dependencies, we have a process aggregation hierarchy, which captures the hierarchical execution ordering of activities.

**Definition 2 (Dependency Graph, *DG*)**
A dependency graph *DG* is defined by a binary tuple <*DN*, *DE*>, where

- $DN = \{nd_1, nd_2, ..., nd_n\}$ is a finite set of activity nodes where $n \geq 1$.
- $DE = \{e_1, e_2, ..., e_m\}$ is a set of edges, $m \geq 0$. Each edge is of the form $nd_i \rightarrow nd_j$. ■

Note that in the dependency graph formulation, self-edges are disallowed since edges are intended to denote data flow dependencies between different activities (nodes). Additionally, a dependency graph must be a connected graph. Unconnected nodes and isolated groups of nodes are disallowed in the graph, as isolated nodes or groups of nodes are considered a separate service process in our reference model.

Given two processes and their respective dependency graphs, there are numerous ways these two graphs may differ. Typically, it makes more sense to compare only those graphs that have sufficient similarity in terms of their dependency graphs. Consider two extreme cases: one is when there is no common node between two graphs and the other is when the two dependency graphs have the same set of nodes. By assigning 0 for the first case and 1 for the latter case, we define a comparability measure that indicates the ratio of common nodes in two graphs. One way to measure the extent of comparability between two graphs is to use a user-controlled threshold, called δ-Comparability, which is set to be between 0 and 1. Because this value represents the ratio of common nodes over the union of all nodes in two graphs, the larger the value is, the greater degree of comparability between the two graphs. Note that δ value can not be 0 since δ = 0 means that there is no common node between two graphs, i.e., $DN_1 \cap DN_2 \neq \varnothing$.

**Definition 3 (δ-Comparability of *DG*)**

Let $DG_1 = (DN_1, DE_1)$ and $DG_2 = (DN_2, DE_2)$ be two dependency graphs, and δ be a user-defined control threshold. We say that $DG_1$ and $DG_2$ are *δ-comparable* if the condition $\frac{|DN_1 \cap DN_2|}{|DN_1 \cup DN_2|} \geq \delta$ holds, where $0 < \delta \leq 1$ ■

If we apply the δ-Comparability to the example graphs shown in Fig. 2 with δ=0.5, $g^0$ and $f^2$ are not comparable because the number of common nodes is only one but the number of total nodes is 7, that is $\frac{|DN_1 \cap DN_2|}{|DN_1 \cup DN_2|} = \frac{1}{7} < 0.5$. On the other hand, $g^0$ and $g^2$ are δ-comparable because there are 3 common nodes and the total number of nodes is 5, thus the two graphs satisfy the δ-comparability condition $\frac{|DN_1 \cap DN_2|}{|DN_1 \cup DN_2|} = \frac{3}{5} \geq 0.5$ and δ = 0.5.

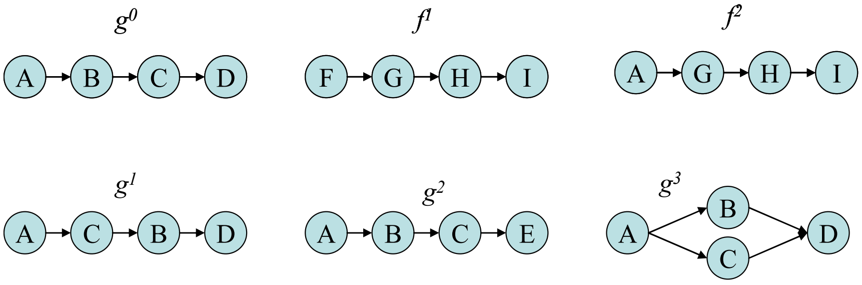


Fig. 2 Examples of δ-Comparability

# 4. Motivating Scenarios

Given the process reference model, we consider two motivating scenarios that benefit from the difference analysis methodology introduced in this paper. Consider a scenario where a company has maintained a warehouse of existing processes used in various business locations. *Process mining* (Aalst, 2003b; Aalst, 2004) of the process warehouse can help the enterprise to discover interesting associations or classifications among business processes running at different locations or branches of the company.

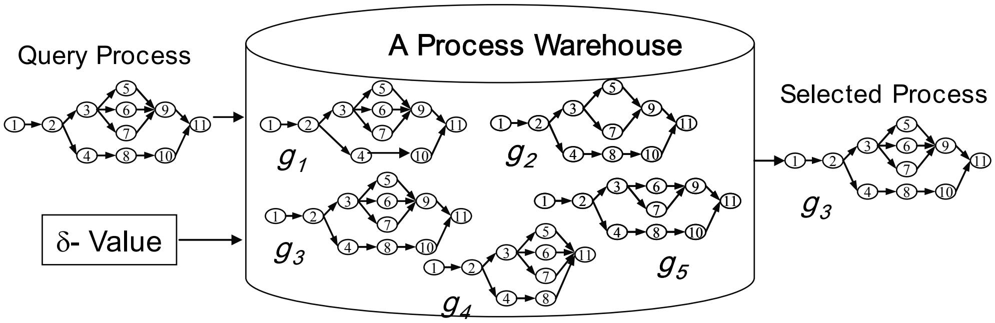


Fig. 3 Process mining example

In Fig. 3, we show a process warehouse that contains many types of processes (for example, $g_1$, $g_2$, $g_3$, $g_4$, $g_5$). A typical process mining scenario is the identification of the processes most similar to a query process template in the process warehouse. Given a query process and a comparability threshold δ-value, the process mining will identify ($g_3$) as the process that is most similar based on the comparability criterion. It is obvious that the concept of process similarity (or distance) is critical to the effectiveness of process mining.

## 5. Process Difference Analysis

In this section, we present the process difference analysis method for evaluating the distance between two processes. We first define the concept of a process matrix and introduce the concept of a normalized matrix. And then, we define the dependency distance measure by measuring the difference between the normalized matrices.

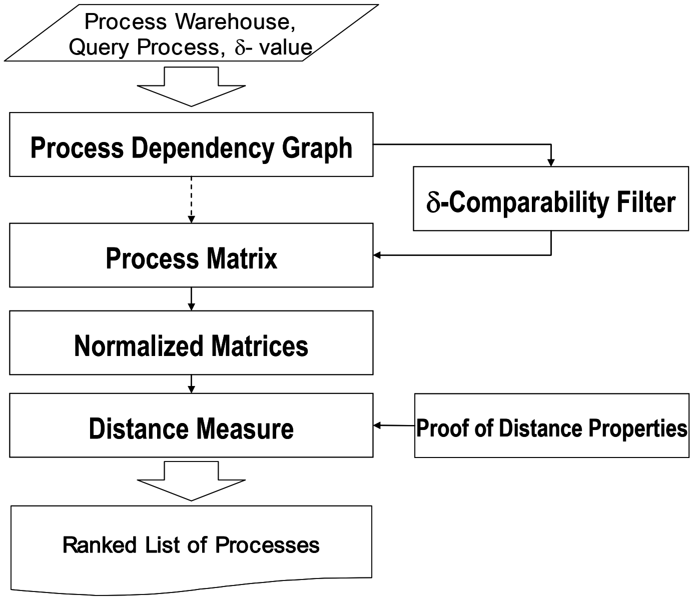
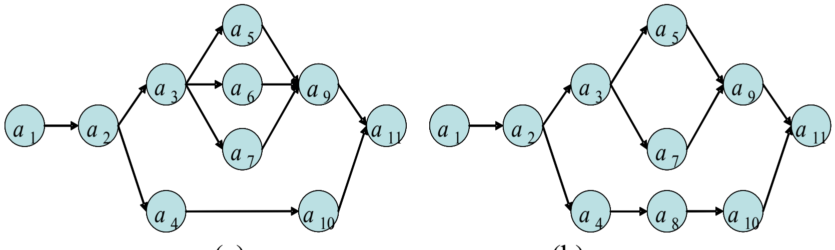


Fig. 4 Flow chart of Difference Analysis

In order to show the proposed procedure, we use two derived processes that are variations of procurement order process in Fig. 1. These two processes have 10 activities respectively but have different activities with each other. The first process ($g_{11}$) has $A_6$ but does not have $A_8$, and the second process ($g_{22}$) has $A_8$ but does not have $A_6$. These two graphs satisfy δ-Comparability as $\frac{|DN_1 \cap DN_2|}{|DN_1 \cup DN_2|} = \frac{9}{11} \geq 0.5$ and $\delta = 0.5$.



(a) $g_{11}$ (b) $g_{22}$

Fig. 5 Two extended examples of Fig. 1

## 5.1. Comparison Matrices

Two dependency graphs are said identical if the two graphs have the same set of nodes and the same set of edges. Formally we define identical dependency graphs as follows:

**Definition 4 (Identical dependency graphs)**
Let $DG_1 = (DN_1, DE_1)$ and $DG_2 = (DN_2, DE_2)$ be two dependency graphs. We say that $DG_1$ and DG$_2$ are identical if the two graphs have the same set of nodes and the same set of edges.

i) $DN_1 =_{Set} DN_2$ ii) $DE_1 =_{Set} DE_2$ ■

One way to compare and rank a set of similar process definitions is to transform each dependency graph into a numerical representation. This allows us to compare the dependency graphs using similarity distance in Euclidian distance metric space. This leads us to introduce the concept of a process matrix. A process matrix *M* is established in order to describe the precedence dependencies between two activities (tasks). The size of *M* is determined by the number of nodes in the dependency graph and each cell in the matrix denotes an element of *M.* The value of cell *M*(*i,j*) is set either to 1 or 0 depending on whether or not there is a precedence dependency between the two nodes *i* and *j*.

## Definition 5 (Process matrix, *M*)

Let $g = (DN, DE)$ be a dependency graph with $|DN| = n$ nodes. A process matrix *M* of *g* is *n*-by-*n* matrix with *n* rows and *n* columns, and each row is named after the node name. Let $M_g(i,j)$ denote the value of the $i^{\text{th}}$ row and the $j^{\text{th}}$ column in *M,* $1 \leq i, j \leq n$. We define $M_g(i,j)$ as follows:

$$M_g(i,j) = \begin{cases} 1 & if\ \exists\ nd_i, nd_j \in DN\ such\ that\ (nd_i, nd_j) \in DE \\ 0 & else \end{cases} \quad ■$$

Fig. 6 depicts the transformation of a process dependency graph $g_{11}$ shown in Fig. 5 (a) into its process matrix *M*, a 10×10 matrix. Each element of *M* is determined according to whether or not the corresponding two activities have precedence dependency. An edge between nodes $a_1$ and $a_2$ shows that activity $a_1$ precedes activity $a_2$. Thus, $M_g(a_1, a_2)$ is set to a value of 1. There is no direct edge between nodes $a_1$ and $a_3$. Thus $M_g(a_1, a_3)$ is set to a value of 0.

| $M_{11}$ | TO | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FROM | | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_9$ | $a_{10}$ | $a_{11}$ |
| | $a_1$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $a_2$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $a_3$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| | $a_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | $a_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | $a_6$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | $a_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | $a_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | $a_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | $a_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 6 Process matrix of $g_{11}$

In order to compare the two process dependency graphs $g_{11}$ and $g_{22}$, we need to further normalize each process matrix that participates in the similarity computation. Each normalized process matrix includes the union of all sets of nodes, each from one participating process dependency graph. We formally introduce the concept of normalized process matrix in Definition 6 by extending the definition of a process matrix to include the entire union of nodes in the two graphs. The size of the normalized matrix is increased to the size of the union of the sets of nodes in both graphs. For those nodes that exist in a process matrix before normalization, the corresponding elements in the normalized matrix are the same as those in the process matrix. For those nodes added through the normalization, the corresponding elements in the normalized matrix are set to a value of 0. After normalization, both matrices have the same number of rows and columns, and share the same row and column names and sequences. The normalized matrices can then be used as an input to calculate distance.

**Definition 6 (Normalized Matrix, *NM*)**

Let $DG_1 = (DN_1, DE_1)$ and $DG_2 = (DN_2, DE_2)$ be two dependency graphs. Let $NM_1$ and $NM_2$ denote the normalized matrices for $DG_1$ and $DG_2$ respectively. We generate $NM_1$ and $NM_2$ from $DG_1$ and $DG_2$ as follows.

i) The number of rows and columns are computed by $m = |DN_1 \cup DN_2|$

ii) Let $DN_1 \bigcup DN_2 = \{a_1, a_2, ..., a_m\}$. Note that the row and column names of $NM_1$ and $NM_2$ are now normalized into the same node names $a_1, a_2, ..., a_m$ in the union of $DN_1$ and $DN_2$.

iii) Let $NM_1(i, j)$ denote the value of the $i^{th}$ row and the $j^{th}$ column in $NM_1$, and $NM_2(i, j)$ denote the value of the $i^{th}$ row and the $j^{th}$ column in $NM_2$

$$NM_1(i, j) = \begin{cases} 1 & \text{if } (a_i, a_j) \in DE_1 \\ 0 & \text{otherwise} \end{cases},$$

$$NM_2(i, j) = \begin{cases} 1 & \text{if } (a_i, a_j) \in DE_2 \\ 0 & \text{otherwise} \end{cases} \quad \blacksquare$$

Consider processes in Fig. 5 as an example. By constructing normalized matrices for $g_{11}$ and $g_{22}$, denoted by $NM_{11}$ and $NM_{22}$ respectively, the size of $NM_{11}$ of $g_{11}$ is increased to 11 because $NM_{11}$ should include node $a_8$, which was not originally included in $g_{11}$. All the elements of the newly added column for node $a_8$ are set to a value of 0 because there is no dependency between any node of $g_{11}$ and node $a_8$. Similarly, node $a_6$ is added in $NM_{22}$. Now $NM_{11}$ and $NM_{22}$ have the same row names and column names: $a_1$ through $a_{11}$. We can use $NM_{11}$ and $NM_{22}$ to compare $g_{11}$ and $g_{22}$.

| $NM_{11}$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $A_1$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $A_2$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $A_3$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| $A_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $A_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $A_6$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $A_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $A_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $A_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $A_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $A_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

(a) $NM_{11}$

| $NM_{22}$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_2$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_3$ | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| $a_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| $a_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $a_6$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $a_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $a_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $a_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $a_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

(b) $NM_{22}$

Fig. 7 An example of comparison matrices

The algorithm for construction of normalized process matrices consists of three steps. First, we must determine whether or not $DG_1$ and $DG_2$ are δ-comparable for the given δ value. Second, we compute the size of the normalized $NM$ by $m = |DN_1 \cup DN_2|$ and label nodes in $\{DN_1 \cup DN_2\}$ as $\{a_1, a_2, ..., a_m\}$ using a uniform naming scheme. Third, we create the matrix data structures for $DG_1$ and $DG_2$: $NM_1(i, j)$ and $NM_2(i, j)$, where $i, j$ = 1, 2, ..., m, and assign a value of 1 or 0 to each element in the two normalized matrices.

## 5.2 Distance-based Process Similarity Measures

With the concept of a normalized matrix, we now transform the problem of comparing two processes into the problem of computing the distance-based similarity of the two normalized process matrices. One obvious idea is to compute the distance of two normalized matrices using matrix subtraction.

Consider the example processes $g_{11}$ and $g_{22}$ in Fig. 5. One way of computing the distance between $g_{11}$ and $g_{22}$ by matrix subtraction is to simply perform subtraction element by element. By subtracting $NM_{22}$ from $NM_{11}$, we can see only five elements have values 1 and -1 respectively and the rest of the elements are 0. This means that five elements are unmatched between the two dependency graphs $g_{11}$ and $g_{22}$.

$NM_1 – NM_2 =$

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_2$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_3$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| $a_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 1 | 0 |
| $a_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_6$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $a_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| $a_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

A drawback of this approach is that both 1 and -1 values in the resulting matrix represent the fact that there are some discrepancies between two graphs $g_{11}$ and $g_{22}$ in five elements. But it does not tell the degree of such discrepancies in terms of concrete distance measure. Thus we need an efficient way to represent the total number of non-zero values in the resulting matrix.

One obvious way to capture the degree of the difference between $NM_{11}$ and $NM_{22}$ is to use the sum of the squares of elements in $NM_1 – NM_2$ as shown below, which is $(1)^2 + (-1)^2 + (1)^2 + (1)^2 + (-1)^2 = 5$ because only five elements have non-zero values 1 and -1.

$(NM_{11} - NM_{22})(NM_{11} - NM_{22})^T =$

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_2$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_3$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_4$ | 0 | 0 | 0 | 2 | 0 | 0 | 0 | -1 | 0 | 0 | 0 |
| $a_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_6$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| $a_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_8$ | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Interestingly, we can calculate the sum of the squares of elements in a matrix by the notion of trace in linear algebra. According to (Anton, 1994), the sum of diagonal elements in a matrix is defined as the trace of the matrix. The best way to calculate the sum of the squares of elements in a matrix is using the concept of inner products, which is defined by the trace concept.

### Definition 7 (Dependency Difference Metric, *d*)

Let $DG_1 = (DN_1, DE_1)$ and $DG_2 = (DN_2, DE_2)$ be two dependency graphs. Let $NM_1$ and $NM_2$ be the normalized matrix of $DG_1$ and $DG_2$ respectively. We define the symmetric difference metric on graphs $DG_1$ and $DG_2$ by the trace of the difference matrix of $NM_1$ and $NM_2$ as follows:

$$d(DG_1, DG_2) = tr[(NM_1 - NM_2) \times (NM_1 - NM_2)^T]$$

where $tr[\cdot]$ denotes the trace of a matrix, i.e., the sum of the diagonal elements. ■

This distance function counts the number of edge discrepancies between $DG_1$ and $DG_2$. Now, we want to show that the dependency difference metric *d* satisfies the distance measure properties. The function *d* is called a metric if and only if for all graphs $g_1$, $g_2$, $g_3$, the following conditions hold (Banks, 1994):

i) $d(g_1, g_2) = 0$ iff $g_1$ and $g_2$ are identical
ii) $d(g_1, g_2) = d(g_2, g_1)$
iii) $d(g_1, g_2) \leq d(g_1, g_3) + d(g_3, g_2)$.

**Theorem 1. *d*(*DG*$_1$,*DG*$_2$) satisfies Distance Measure Properties.**
Proof:
Concretely, we want to prove that if $A = NM_1 - NM_2$ and

$d(DG_1, DG_2) = < A, A^T > = tr(A \times A^T) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^{\ 2}$ , then this distance $d(DG_1, DG_2)$ satisfies the three distance measure properties:

i) $d(DG_1, DG_2) = 0$ iff $DG_1$ and $DG_2$ are identical, because the matrix A becomes 0.
ii) $d(DG_1, DG_2) = d(DG_2, DG_1)$ by the *d* definition.
iii) $d(DG_1, DG_2) \leq d(DG_1, DG_3) + d(DG_3, DG_2)$

For any two nodes *i*, *j*, let

$$NM_k(i, j) = \begin{cases} 1 & \text{if } (a_i, a_j) \in DE_1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k\text{=1, 2, 3}$$

Then we can show the property iii) holds.

$$\begin{aligned} d(DG_1, DG_2) &= tr[(NM_1 - NM_2) \times (NM_1 - NM_2)^T] \\ &= \sum\nolimits_{i,j} \{NM_1(i, j) - NM_2(i, j)\}^2 \quad . \\ &= d(DG_2, DG_1) \end{aligned}$$

Now we show that the property iii) holds as well, because $NM_1(i,j) - NM_2(i,j)$ is either 0 or ±1, thus we have $d(DG_1, DG_2) = \sum_{i,j} \left| NM_1(i,j) - NM_2(i,j) \right|$.

$$\begin{aligned}
& d(DG_1, DG_3) + d(DG_3, DG_2) \\
& = \sum_{i,j} \left| NM_1(i,j) - NM_3(i,j) \right| + \sum_{i,j} \left| NM_3(i,j) - NM_2(i,j) \right| \\
& = \sum_{i,j} \left\{ \left| NM_1(i,j) - NM_3(i,j) \right| + \left| NM_3(i,j) - NM_2(i,j) \right| \right\} \\
& \geq \sum_{i,j} \left| NM_1(i,j) - NM_3(i,j) + NM_3(i,j) - NM_2(i,j) \right| \\
& = \sum_{i,j} \left| NM_1(i,j) - NM_2(i,j) \right| \\
& = d(DG_1, DG_2)
\end{aligned}$$

So the new process distance measure is, in fact, a distance metric. ■

Since the dependency distance metric *d(*$g^1$*,* $g^2$*)* counts the number of asymmetric arcs, it can reflect the difference of some characteristics between two processes, such as activity precedence, activity commonality, flow structure, etc. Activity precedence describes how the activities are linked and sequenced in terms of execution ordering. The dependency distance metric denotes the disparity of sequence between two activities and can be extended to represent the sequence disparities between all activities. In Fig. 8, the distance of two processes $g^0$ and $g^1$, denoted by *d(*$g^0$*,* $g^1$*)*, illustrates the difference of activity precedence. Activity commonality means how many activities are shared between two process models. This counts the different activities or new activities of two processes, as illustrated by processes $g^0$ and $g^2$ in Fig. 8. In addition, flow structure denotes the difference between serial and parallel flows. Two processes $g^0$ and $g^3$ show the difference measurement of flow structures, serial and parallel flows.

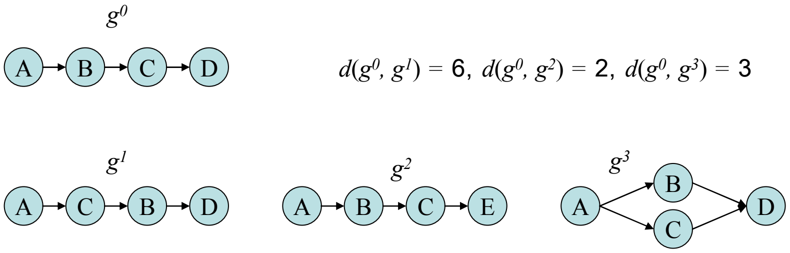


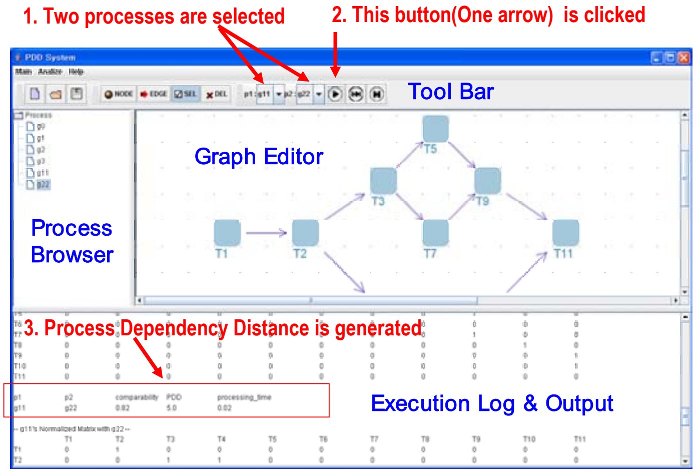
Fig. 8 Examples of dependency distance

In Fig. 8, if we follow the previous procedure to calculate the dependency distance, all of the graphs are transformed to process network matrices and normalized process matrices. Then the distance of dependency between $g^0$ and $g^1$ is 6, the distance of $g^0$ and $g^2$ is 2, and the distance of $g^0$ and $g^3$ is 3. This means that $g^0$ and $g^2$ are the most similar, which is intuitively correct because the first three activities are in the same sequence but only the last activity is different. $g^0$ and $g^1$ are mostly different because the sequence of the activities in $g^1$ is quite different from $g^0$. In this dependency distance measure, the parallel execution in $g^3$ is not considered important and only the precedence relationships and common activities are considered important.

If we look into more extended examples in Fig. 5 again, each graph is transformed into process matrix, and then normalized matrix. These two normalized matrices are subtracted and squared. Finally we can get the proposed dependency distance 5 by obtaining the trace of it.

## 6. Prototype Implementation and Experiments

The presented concepts of this paper were implemented to analyze the similarity of processes in process warehouse. This system, called “BPSAT(Business Process Similarity Analysis Tool)”, is developed by using Java language. This prototype system has three windows: process browser, graph editor, and execution log output window. We can select some processes in the left process browser, and the selected process is shown and modified in the right graph editor. All the execution log and analysis outputs are displayed in the bottom window. There are also necessary buttons in tool bar. The basic manipulation such as creating and editing of process graph can be done in this prototype system, and the functionality of similarity analysis methods proposed in this paper can be done in this system. Also other new similarity criteria can be added in this system. The current version of this system can be downloaded at http://it.chonbuk.ac.kr/~jsbae/BPMstuff/BPMstuff.html.



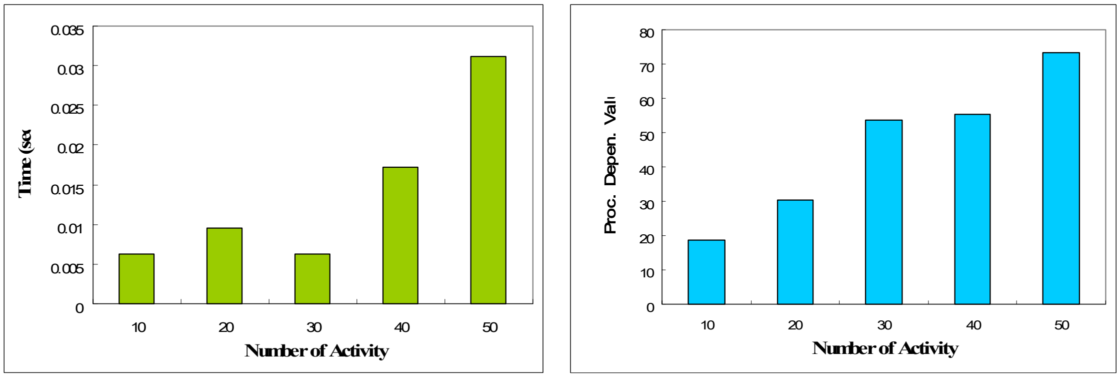
```
T8     0    0    0    0    0    0    0    0    0    1    0
T9     0    0    0    0    0    0    0    0    0    0    1
T10    0    0    0    0    0    0    0    0    0    0    1
T11    0    0    0    0    0    0    0    0    0    0    0

p1     p2     comparability  PDD   processing_time
g11    g22    0.82           5.0   0.02

-- g11's Normalized Matrix with g22 --
       T1   T2   T3   T4   T5   T6   T7   T8   T9   T10  T11
T1     0    1    0    0    0    0    0    0    0    0    0
T2     0    0    1    1    0    0    0    0    0    0    0
```

Fig. 9 Prototype system of BPSAT

After we check the candidate processes to be compared, we select two processes to be compared, $g_{11}$ and $g_{22}$. Then we can get the proposed process dependency distance is generated and shown in the output window.

Using the prototype system, we conducted experiments to analyze effectiveness of our method with variation of activity number. We did our experiments for processes including a number of activities. All the processes are generated using random process generator developed in (Ha, 2006). Ten pairs of different distances were calculated for processes with the same number of activities, and an average value was obtained for each number of activities.

First, we observed time required for calculating process dependency distance with increase of activity number. As we expected, more time is required as the number of activities increases, but the increase rate is not so high. The experimental result is presented in Fig. 10 (a).



(a) Calculation time (b) Process dependency distance

Fig. 10 Calculation time and process dependency distance value according to process structure

Next, we examined how values of process dependency distance change as the number of activities increases, which is illustrated in Fig. 10 (b). Our result shows that absolute values of the distance in case of more activities are higher than those of fewer activities. This means that distance values among processes cannot be compared for an arbitrary number of activities, which we reserve as our future research work.

## 7. Related Work

Although business process management systems have been deployed in many industrial engineering fields, research on analysis, mining and integration of business processes are still in its infancy. One of the representative existing studies on process improvement is workflow mining, which investigates the traces and results of workflow execution, and determines significant information in order to improve the existing workflow processes (Aalst, 2003b; Aalst, 2004; Agrawal, 1994; Cook, 1999; Schimm, 2004). However, most of the existing workflow mining research does not provide a quantitative measure to compare and capture the similarity of different workflow designs.

The graph theory in a traditional algorithm textbook is a useful means to analyze the process definitions. Graphs, or representative data structures, are used as an accepted effective tool to represent the problem in various fields, which include pattern matching and machine recognition, such as pattern recognition, web and XML document analysis, and schema integration (Bunke, 1998; Hammouda, 2004; Wombacher, 2004; Zhang, 1989). For example, research on similarities in graph structures can be divided into three categories. The first category of traditional similarity is based on graph and sub-graph isomorphism, which has several weaknesses and distortions in the input data and the models. In order to overcome these weaknesses, other graph similarity analysis techniques, such as the graph edit distance (GED) metric and maximal common sub-graph (MCS) have been introduced (Bunke, 1998; Zhang, 1989). It is also worth mentioning that Bunke (Bunke, 1998) has shown that with generic graphs, under certain assumptions concerning the edit-costs, determining the maximum common sub-graph is equivalent to computing the graph edit-distance. This MCS is a basic concept of workflow similarity that measures the common activities and transitions of workflow processes. In this paper we utilize the graph theory results to derive the metric space distance metric for measuring process similarity and difference.

Our research on workflow similarity measure is mainly inspired by the research results on document similarity analysis and graph similarity measures. A large number of document similarity measures are presented in existing literature for building document management systems, knowledge management systems, as well as search engines (Bunke, 1998; Hammouda, 2004; Lian, 2004).

Finally, in order to support web service composition, an infrastructure for searching and matchmaking of business processes is needed. One example is using annotated deterministic finite state automata (aDFA) to model the business processes (Wombacher, 2004). If a business process is specified as aDFA, the match between two aDFAs is determined by the intersection of their languages. When there is non-empty intersection, the two business processes are matched.

## 8. Conclusion and Future work

We have presented a difference analysis methodology using distance measures between process definitions of web services. The proposed difference analysis method achieves three distinct goals. First, by analyzing the attributes of process models, we can present a quantitative process similarity metric to determine the relative distance between process models. This facilitates not only the comparison of existing process models with each other, but also provides the flexibility to adapt to changes in processes. Second, the proposed method is fast and flexible, which reduces the cost of both the analysis and design phases of complex web service processes. Third, the proposed method enables the flexible deployment of process mining, discovery, and integration – all desirable functionality that are critical for fully supporting the effective transformation of an enterprise.

Our research on process mining, discovering and integration through similarity analysis continues along several directions. First, we are interested in distance measures that can compare workflow designs with complex block structure and various execution constraints. Second, we are interested in developing a prototype system that provides efficient implementation of various similarity analysis methods, including the dependency distance metric presented in this paper. Furthermore we are interested in applying the method developed to concrete case studies of existing enterprise transformations and to evaluate and improve the similarity measures proposed in this paper.

## ACKNOWLEGMENT

The first author was supported by the Korea Research Foundation Grant (KRF-2005-214-D00192). The second author is partially supported by NSF and AFOSR.

### References

van der Aalst, W. M. P., Hofstede, A.H.M. ter, Kiepuszewski, B. Barros, A.P., (2003), Workflow Patterns, *Distributed and Parallel Databases*, 14(3), 5-51.

van der Aalst, W.M.P., van Dongen, B.F., Herbst,,J., Maruster,,L., Schimm, G., Weijters, A.J.M.M., (2003), Workflow Mining: A Survey of Issues and Approaches, *Data and Knowledge Engineering*, 47(2), 237-267.

van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L., (2004), Workflow Mining: Discovering Process Models from Event Logs, *IEEE Transactions on Knowledge and Data Engineering*, 16(9), 1128-1142.

Agrawal,, R. Gunopulos, D., Leymann, F., (1998), Mining Process Models from Work-flow Logs, *6th International Conference on Extending Database Technology*, 469-483.

Anton, H., Rorres, C., (1994), Elementary Linear Algebra: Applications, *John Wiley&Sons*.

Bae, J., Bae, H., Kang, S., Kim, Y., (2004), Automatic control of workflow process using ECA rules, *IEEE Trans. on Knowledge and Data Engineering*, 16(8), 1010-1023.

Banks, D., Carley, K., (1994), Metric inference for social networks, *Journal of classification*, 11(1), 121-149.

Bunke, H,. Shearer, K., (1998),A Graph Distance Metric based on the Maximal Common Subgraph, *Pattern Recognition Letters*, 19(3-4), 255-259.

Cook, J.E., Wolf, A.L., (1999), Software Process Validation: Quantitatively Measuring the Correspondence of a Process to a Model, *ACM Transactions on Software Engineering and Methodology*, 8(2), 147-176.

Ha, B.-H., Reijers, H. A., Bae, J., Bae, H., (2006), An Approximate Analysis of Expected Cycle Time in Business Process Execution, *Lecture Notes in Computer Science*, 4103, 65-74.

Hammouda, K.M., Kamel, M.S., (2004), Efficient Phrase-Based Document Indexing for Web Document Clustering, *IEEE Transactions on Knowledge and Data Engineering*, 16(10), 1279-1296.

Leymann, F., Roller, D., (2000), Production workflow: concepts and techniques, *Prentice Hall PRT*, New Jersey.

Lian, W., Cheung, W.W., Mamoulis, N., Yiu, S., (2004), An Efficient and Scalable Algorithm for Clustering XML Documents by Structure, *IEEE Transactions on Knowledge and Data Engineering*, 16(1), 82-96.

RosettaNet, RosettaNet Standard (RosettaNet Partner Interface Processes), http://www.rosettanet.org.

Rouse, W. B., (2005), A Theory of Enterprise Transformation, *Systems Engineering*, 8(4), 279-295.

Rush, R., Wallace, W.A., (1997), Elicitation of knowledge from multiple experts using network inference, *IEEE Transactions on Knowledge and Data Engineering*, 9(5), 688-698.

Schimm, G., (2004), Mining exact models of concurrent workflows, *Computers in Industry*, 53(3), 265-281.

WfMC, (2005), Workflow Management Coalition Workflow Standard Process Definition Interface -- XML Process Definition Language, Document Number WFMC-TC-1025 Version 1.13.

Wombacher, A.,Fankhauser, P., Mahleko, B., Neuhold, E., (2004), Matchmaking for Business Processes Based on Choreographies, *International Journal of Web Services*, 1(4), 14-32.

Zhang, K., Shasha, D., (1989), Simple Fast Algorithms for the Editing Distance between Trees and Related Problems, *SIAM Journal of Computing*, 18(6), 1245-1262.

## ABOUT THE AUTHORS

**Joonsoo Bae** is an Assistant Professor of Department of Industrial and Information Systems Engineering in Chonbuk National University. He received PhD, MS, and BS degrees in Industrial Engineering from Seoul National University, South Korea in 2000, 1995, and 1993, respectively. He also completed one year postdoctoral course in College of Computing of Georgia Institute of Technology at 2006. He had industry experience in LG-EDS as a technical consultant of SCM & CRM team from 2000 to 2002. He is interested in system design and integration of management information system and e-Business technology. His research topics include business processes management using workflow systems and advanced internet application.

**Ling Liu** is an associate professor at the College of Computing at Georgia Tech. There, she directs the research programs in Distributed Data Intensive Systems Lab (DiSL), examining research issues and technical challenges in building large scale distributed computing systems that can grow without limits. Dr. Liu and the DiSL research group have been working on various aspects of distributed data intensive systems, ranging from distributed computing systems, enterprise systems to business workflow management systems. Prof. Liu has published more than 160 technical papers in the areas of Internet Computing systems, Internet data management, distributed systems, and information security. She is the recipient of best paper award of WWW 2004 and best paper award of IEEE ICDCS 2003, and a recipient of 2005 Pat Goldberg Memorial Best Paper Award. Her research group has produced a number of software systems that are either open sources or directly accessible online, among which the most popular ones are WebCQ and XWRAPElite. Dr. Liu is currently on the editorial board of several international journals, including IEEE Transactions on Knowledge and Data Engineering, International Journal of Very large Database systems (VLDBJ), International Journal of Web Services Research, and has chaired a number of conferences as a PC chair, a vice PC chair, or a general chair, including IEEE International Conference on Data Engineering (ICDE 2004, ICDE 2006, ICDE 2007), IEEE International Conference on Distributed Computing (ICDCS 2006), IEEE International Conference on Collaborative Computing (CollaborateCom 2005, 2006), IEEE International Conference on Web Services (ICWS 2004). She is a recipient of IBM Faculty Award (2003, 2006). Prof. Liu's current research is partly sponsored by grants from NSF CISE CSR, ITR, CyberTrust, AFOSR, and IBM.

**James Caverlee** is a Ph.D. candidate in the College of Computing at Georgia Tech and a member of the multidisciplinary Tennenbaum Institute for enterprise transformation. His research interests are generally in the areas of Web and Distributed Information Management, with an emphasis on: (1) Enterprise Computing and Workflow Management; (2) Spam-Resilient Web-Scale Computing; and (3) Web Information Retrieval and Management. James graduated magna cum laude from Duke University in 1996 with a B.A. in Economics. He received the M.S. degree in Engineering-Economic Systems & Operations Research in 2000, and the M.S. degree in Computer Science in 2001, both from Stanford University.

**Liang-Jie(LJ) Zhang** is a Research Staff Member (RSM) in Services Technologies Department at IBM T.J. Watson Research Center. He has been leading SOA and Web Services for Business Consulting Services and Industry Solutions research since 2001. He is the founding chair of the Services Computing PIC (Professional Interest Community) at IBM Research and lead professional activities for IBM's Services Computing discipline. In 2004 and 2005, Dr. Zhang was appointed as the Chief Architect of Industrial Standards at IBM Software Group, where he played leadership role in helping define IBM's strategy for industrial standards and open architecture for service-oriented business solutions.

**Hyerim Bae** is an assistant professor in the Industrial Engineering Department at Pusan National University (PNU), Korea. He received PhD, MS, and BS degrees from the Industrial Engineering Department at Seoul National University, Korea. He had been a manager for information strategic planning at Samsung Card Corporation before he joined PNU. He is interested in the areas of Business Process Management (BPM), process-based B2B integration, and ubiquitous business computing. His current research activities include analysis of business process efficiency, controlling of logistics processes with context awareness, and convenient modeling of business processes.