워크플로우 마이닝 : 휴리스틱 접근

이명희[*] ,유철중[**], 장옥배[***]

Workflow Mining based on Heuristic Approach using Log data

Myoung-Hee Lee[*] , Cheol-jung Yoo[**], Ok-bae Jang[**]

Abstract

As the workflow systems are becoming complex and obscure, there are discrepancies between actual workflow process and designed process. Therefore, we have developed techniques for discovering workflow models. The starting point for such techniques is a so-called "workflow log" containing information about the workflow process as it is actually being executed. This paper presents an algorithm of workflow process mining based on heuristic approach from the workflow log, which can be happen to business process system.

Key Word : Workflow Mining, Heuristic Approach

[*] 전북기능대학 멀티미디어과 조교수

[**] 전북대학교 컴퓨터과학과 조교수

[***] 전북대학교 컴퓨터과학과 교수

1. Introduction

1.1 Background and Necessity

Generally, business activity has developed as process in the form of collective form in the order of tasks that are most simple, efficient, easy to understand, concerning stages it must take to produce, or make business with others. Also, such process has developed as business process to achieved organization or corporation goal along with development of humanity, and into a workflow in a larger aspect.

Recently, transaction in many forms such as government to business (G2B), business to business (B2B), business to consumer (B2C), and more based on government, corporation, consumer is especially attempting modification into workflow by being activated into electronic commerce transaction, which is a new concept in commerce, and being connected to expanded infrastructure of information network.

On the other hand, such meaning of this process in the aspect of software engineering, must go through a series of process to achieve a task such as providing service, preparing report, or developing a software while producing. This task is executed in the same order every time; for example, building fences after completing house construction, or mixing all ingredients before baking a cake. In other words, it is a set of ordered tasks, and a series of process including activities, constraints, and resources to achieve results in such form [1]. These processes are important since they can increase efficiency by emphasizing on the structure and consistency of a series of activities.

Moreover, defining process in the aspect of business means to define how the organization can achieve its goal [2]. As can be found in this definition, business process has become an inevitable core idea to electronic commerce between corporations. Also, interest for workflow is increasing recently along with business process, as service between e-market places is becoming serious in all businesses.

This paper analyzes workflow mining to deduct and support more efficient process using log file of workflow, and suggests rules to manage workflow more efficiently by applying mining algorithm based on this analysis. After mining according to the rules of mining, achievable visibility and efficiency among tasks are searched, and then workflow mining of a corporation is explained and analyzed as an example.

2. Related Study

2.1 Workflow

Workflow chosen by WMC is one that benefits or automated whole or part of business process. In other words, workflow is an information technology supporting accurate and quick task management through automation of business process.

2.2 Workflow Mining

Workflow mining is a technique analyzing trend of process of workflow system, or providing other data using workflow monitoring. In other words, workflow mining must be able to provide other data to corporations or customers through analysis of task process [4].

2.3 Data Mining Algorithm

Data mining is a series of process finding significant correlation, pattern, trend, and more included in a massive material, and uses diverse techniques such as statistic and patter recognition, neural network, and more. This is the representative algorithm;

CART – making classification and regression analysis through decision making tree.

KMEANS – divides community according to similarity as much as input number of community.

PCA – modified into principal component according to the degree of contribution influencing independent variable on dependent variable.

2.3 Analysis of Correlation

Shows correlations between each other concerning all fields in data range (excluding data field in category) using correlation function.

2.4 PCA (Principal Component Analysis) Method

Principal component analysis is a multivariate data analysis technique clarifying relation of variables in low dimension through dimensional decreasing.

$$X = t_1p_1^T + t_2p_2^T + \cdots\cdots + t_tp_t^T$$

t_h is score, $p_h{}'$ is weight

3. Purpose and Method

This study focuses on deducting strongly influential process for efficient optimization using statistical analysis and Heuristic approach among diverse mining approaches.

These are the 4 stages for the whole order;

1st stage : data filtering and pretreatment process

2nd stage : rule definition

3rd stage : case study of applying PCA Heuristic

4th stage : capacity evaluation

4. Heuristic Approach of Workflow Mining

Workflow mining method is progressed in the order of data filtering and disposition, which is a stage to make use of basic data, rule definition of data analysis using principal component and correlation analysis, case study of applying PCA Heuristic, and function evaluation in the end.

4.1 Data Filtering and Pretreatment Process

Remove node unnecessary to PCA.

Designate all variables as independent and continuous variable, since PCA only analyzed with independent variables (X) in ‘type node’.

4.2 Rule Definition

These are the rules;

(1) Statistical analysis of continuous data

(2) Analysis of principal component

(3) Analysis according to groups by designating group data (set groups)
(4) Analysis of correlation
(5) Set coefficient of correlation (0.0 <= p <= 1.0)
(6) Analysis of data using analysis of correlation and principal component

The rule is progressed in 6 stages;

(1) Statistically analyze continuous data

Statistical analysis of continuous data defines variable names, forms of variables, and the input and output forms of variables. Since PCA analyzes only with independent variables (x), designate variable names accordingly from A1 to A54, set the form of variables and continuous, and set the form of input and output accordingly as independent variables.

(2) Analysis of principal component

Decide how many principal components should be divided for analysis of principal component. Generally, the numbers of principal components are decided by selecting a factor number corresponding to the eigen value after showing a sudden declination after deciding according to eigen value (essential price).

(3) Analysis according to groups by designating group data (set groups)

Designate group data of principal component set above, and analyze correlation.

(4) Analysis of correlation

Correlation between each other concerning all fields inside data range (excluding data field in category) can be seen.

(5) Set coefficient of correlation (0.0 <= p <= 1.0)

Setting coefficient of correlation is made appropriately between 0 to 1.0, and then comprehends correlation.

(6) Analysis of data using analysis of correlation and principal component

Mine the best process using correlation table, using contribute table and correlation through analysis between groups of principal component analysis.

5. Case Study applying PCA Heuristic

This is an example of mining shown in production process. It is trying to maintain a certain quality by finding out the possibility of capacity deviation between installations in production process, and its factors. Also, compare and analyze capacity deviation using PCA after understanding data characteristics through the pretreatment procedure.

5.1 Select Domain

(1) The experiment data is a process data of 54 capacity processes in a certain production factory.
(2) Analyze deviation between business processes.
(3) Deduct more influential process.

5.2 Assumption and Constraint

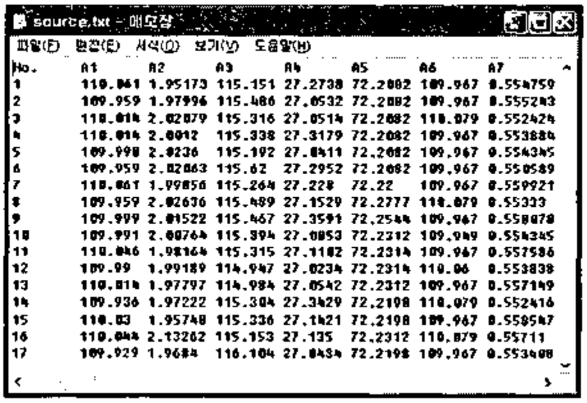
These are the assumption and constraints;

(1) The data must be continuous ones.
(2) Each process is independent.
(3) The size of the process is 54, and the number

is 7596.

(4) Log data does not have noise.

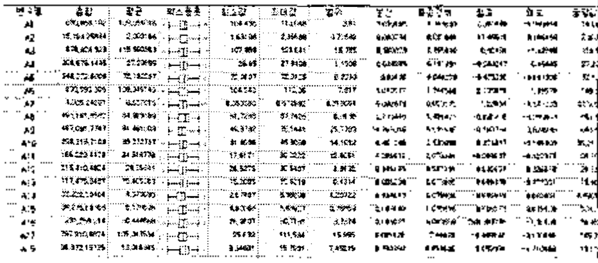
5.3 Original Data



```
No.     A1      A2      A3      A4      A5      A6      A7
1       110.041 1.95173 115.151 27.2738 72.2082 109.967 0.554759
2       109.959 1.97996 115.486 27.0532 72.2082 109.967 0.555243
3       110.014 2.02079 115.316 27.0514 72.2082 110.079 0.552424
4       110.014 2.0012  115.338 27.3179 72.2082 109.967 0.553884
5       109.998 2.0236  115.192 27.0411 72.2082 109.967 0.554345
6       109.959 2.02063 115.62  27.2952 72.2082 109.967 0.550589
7       110.061 1.99856 115.264 27.228  72.22   109.967 0.559921
8       109.959 2.02636 115.489 27.1529 72.2777 110.079 0.55333
9       109.999 2.01522 115.467 27.3591 72.2544 109.967 0.558078
10      109.991 2.00764 115.394 27.0053 72.2312 109.949 0.554345
11      110.046 1.98164 115.315 27.1102 72.2314 109.967 0.557586
12      109.99  1.99189 114.947 27.0234 72.2314 110.06  0.553838
13      110.014 1.97797 114.984 27.0542 72.2312 109.967 0.557149
14      109.936 1.97222 115.304 27.3429 72.2198 110.079 0.552416
15      110.03  1.95748 115.336 27.1421 72.2198 109.967 0.558547
16      110.044 2.13262 115.153 27.135  72.2312 110.079 0.55711
17      109.929 1.9684  116.104 27.0434 72.2198 109.967 0.553408
```

(1) Statistical analysis of continuous data

Analyze distribution, average deviation, kurtosis, skewness through statistical analysis of continuous data.



(2) Analysis of principal component

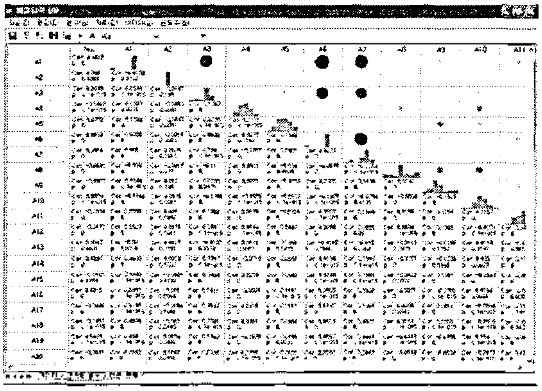
Select 3 factors corresponding to the eigen value after showing sudden declination through bar graph, which shows eigen values in the order of large value when deciding number of principal components.

(3) Analysis according to groups by designating group data (set groups)

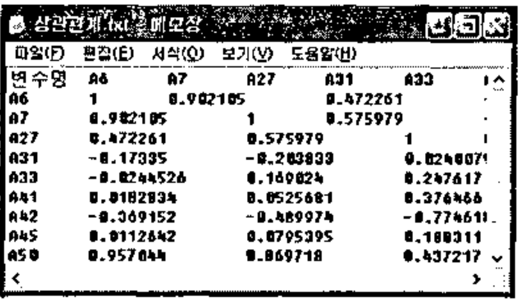
5.4 Data Filtering and Pretreatment Process

Remove 'filter node' since 'No.' is unnecessary in PCA. Also, designate all variables as independent and continuous variables since it only analyzes independent variables (X) in 'type node'.

(4) Analysis of correlation (p = 0)



(5) Set coefficient of correlation (p = 0.9)



```
변수명  A6          A7          A27         A31         A33
A6      1           0.902105                0.472261
A7      0.902105                1           0.575979
A27     0.472261                0.575979                1
A31     -0.17335                -0.203833               0.0240071
A33     -0.0244526              0.169024                0.247617
A41     0.0182834               0.0525681               0.376466
A42     -0.369152               -0.489974               -0.774611
A45     0.0112642               0.0795395               0.180311
A50     0.957044                0.869718                0.437217
```

Correlation table (p = 0.9)

(6) Analysis of data using analysis of correlation and principal component

These are the rules for data analysis;

```
input : workflow event log
output : principal Component process
Rule 1. Definition Correlation table
Given Process A
IF (coefficient of correlation > 0.9) THEN
process candidate
```

This is the process using rule no.1 above;

A6.A7.A27.A31.A33.A41.A45.A50.A53

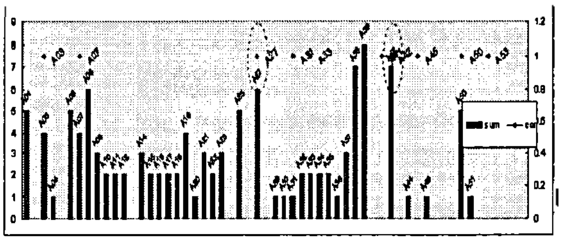
```
Rule 2. Definition Contribution table
Given Process A
For each Principal Component and Group
 (p > 0.8) THEN Process CANDIDATE
p is contribution value
```

```
Rule 3. Given ProcessA
For I = 0 To PrincipalComponentNumber{
        For I = 0 To GroupNumber{
                IF (p > 0.9) && ( count > 5){
                SELECT PROCESS
                }
}}
```

This is the prepared PCA table using rule no.3 above;

Relation	1--2			1--3			2--3			sum	Correlation
Group	g1	g2	g3	g1	g2	g3	g1	g2	g3	n > 0.8	p>0.9
Name	1	2	3	4	5	6	7	8	9		
A01	1	1	1			1			1	5	
A02											
A03	1		1			1			1	4	1
A04				1						1	
A05											
A06	1	1	1			1			1	5	
A07	1		1				1		1	4	1
A08	1	1		1	1		1	1		6	
A09						1		1	1	3	
A10				1	1					2	

This is the graph drawn using a table using rule no.1 and no.2 above;



6. Conclusion and Future Study

In this paper, workflow mining is a method for mining that uses Heurastic approach that can apply according to the rules. This paper mines process more influentially and efficiently through correlation analysis and PCA.

It is thought that mining using more detailed business data applied in actual field is recommendable in the future, and further studies are necessary for more typical rule and algorithm for Heuristic application.

Reference

[1] F. Casati, “Workflow Evolution. Data and Knowledge Engineering,” 24(3):211-238, 1998.

[2] Bussler, C., "B2B Protocol Standards and their Role in Sementic B2B Integration Engines," March 2002, Vol.24, No.1. IEEE Computer Society.

[3] W.M.P van der Aalst, “Process Mining : Discovering Workflow Models from Event-Based Data,” BNAIC 2001, pp.283-290, 2001.

[4] Workflow Management Coalition Specification Document, "The Workflow Reference Model," Document Number TC00-1003 Version1.1, Jan 1995.