The Research of Process Mining Assessment used in Business Intelligence

Zhenyu Wang, Qing Yao, Yuqing Sun
Department of Computer Science and Technology
Shandong University
Jinan, China
dayu.office@gmail.com, yaoqing, sun_yuqing@sdu.edu.cn

***Abstract*—In order to do process mining better in the Business Intelligence environment, the proper process mining model and the right evaluation method to assess the mining results become the key issues. This paper is based on both theoretical and technological achievements in process mining area, putting forward a process mining model under event log mode, providing the analysis process of selecting evaluation indicators by AHP, and gives out a set of formulas for assessing process instance combined with the vector model. For illustration, a case recorded by an insurance company handling claims lodged over the phone is utilized to show the feasibility of the given model and the formulas in solving assessment problems. Empirical results is presented as numerical data by formula assessing effectively to the new process records under specific target, it can be displayed directly for its consequence; process screening and classification, including excavation work for the follow-up by the threshold value setting can evaluate the process improvement and provide basis for decision making in further process mining work.**

Keywords-process mining; assessment; AHP; vector model; WPPDD

I. INTRODUCTION

The optimization and reengineering of the business processes is one of the key problems which promotes enterprise's competitiveness. Business Intelligence (BI) that deals with all kinds of data in enterprise's operation course and excavates the knowledge pattern on the basis of here, and then obtained intellectual technology of commercial affair of decision support has risen and become the important sign of enterprise's modernized information management styles progressively. The Business Process Management (BPM) which is managed and run through in commercial links such as materials procurement, producing, providing and delivering, selling and enterprise's office automation,etc., customized goal and cooperative partner to produce the requirement which hit the agility, pressure of the variable cost to cooperation in the face of the individualized, pluralistic customer, and meet the enterprise process change under the particular environment, demand recombinated, optimized, the BPM as a variable and flexible deployment of enterprise BI strategy plays the most important role.

The existing BPM that is on the basis of Work Flow (WF), overweight performance optimizing, connecting logic and safeguarding the work of the function integrality of the whole business process in the certain procedure, there is not enough description and excavate thoroughly to manage the work of the BPM rules itself, performance optimize and assessment of the process. To make BPM in BI field work well, one set of business process assessment standards should be provided. In order to analyse each assessment index better, one evaluation model used in Process Mining (PM) seems particularly necessary and important. In this paper, Section II presents related work on PM and evaluation. Section III puts forward the process mining model under event log mode based on BI used for PM with thought and method of Data Mining (DM), and Section IV provides the analysis process of selecting evaluation indicators which is the complement to the model before. Combined this model with further study and assessing the evaluation system that comes with several formulas is arranged in section V with a case recorded by an insurance company handling claims lodged over the phone while presents and discusses the result. Finally, the last section highlights the main points of this work and presents its conclusions.

II. RELATED WORKS

Process mining, or workflow mining, risen at the end of last century, according to the description of the literature [1][2][11], can summarize for mode on the basis of WF pattern rediscovering course which integrated the method of automatic treatment for transaction and the standardized data storage solutions that the computer deal with, where noise data such as the unusual, special case,etc. is extracted, transformed and reloaded before analyzing a large number of process data, and thus find effective output about business process reengineering and processes optimization. The assessment that excavate to the process of the literature [11] involves the process model, process modeling rules, including the one that adopted excavated problems such as algorithms,etc. carried on research. It is obvious, under the BI environment, that realizing the effective process evaluation is the key to proving the process implemented integrality and validity.

Through the research of WF models, groups represented by W.M.P. van der Aalst have proposed some application schemes regarding modeling method of the process mining, data storage form, algorithm and relevant on the basis of Petri network in the literature [1][2][3], which provide meta data storage model based on daily event logs, and then a realization scheme based on MXML as XES (eXtensible Event Stream) that is accepted by IEEE as standard format through limiting the essential condition of the data property; In addition, group of this research also provides the instrument to dig up of a process and collects called ProM, in order to carry on the implementation

978-0-7695-4694-0/12 $26.00 © 2012 IEEE
DOI 10.1109/ICIS.2012.109

with a large number of instance, algorithms and scenes how the process mining exactly working in the literature [11]. Literature [4] has made detailed analysis and introduced the main task of process mining in frame. The literature [5] has proposed a decision making model under the BI environment, use AHP (Analytic Hierarchy Process) method to chase layer analysing to the goal and subdivide the complicated goal. The work of Lucene and analysis of search engine in the literature [7] use vectorial model to get feedback result sorted and provide a assessment scheme with stronger compatibility.

This paper combined with the thinking of DM, tries to provide a process mining assessment method based on layer goals, through excavating the classical sample process sets to get Panorama Process Directed Digraph (PPDD), with which to evaluate single procedure again, offer the results in order to evaluate the process improvement and provide basis for decision making in further process mining work.

III. ASSESSMENT MODEL

A practical process instance waiting to assess is reflecting the gathering of a whole set of nodes and flow direction control points of the effective goal. This section provides an intact assessment model, all business process instances to be assessed can be mapped upon one to one, realize the assessment method that provided in the subsequent content. The model can regard as the intersection of business process's closure space.

A. Element and Definition in PAM

Process Assessment Model (PAM) puts the emphasis on relationship between reality and the model whether effectively mapped and carried out, avoiding ambiguity, and support to assess the maximization of the result under the optimum situation of the theory, that is to say to optimum situation in reality extremely well to approach, or reflect directly, and also reduce the jitter rate of the result to get rid of the interference. Considering that the single process is unable to represent the whole characteristic of the process set, fuzzy processing to carry on the process detail is given in aspects of model definition, no longer considering the integrality of the process, thus the best optimum process mode in the process instance set can be checked out easily. The following are relevant element and definition of the PAM.

Define 1: Action =< ActionID, Name, RequireMsgList, SendMsgList, Status >

Action is the basic unit of business treatment in the business process, which accomplish a concrete activity in a certain business logic, fulfill for qualification: atomicity and consistency. Action does not appear in the business process directly, and there is no limit of the using scene; the same action is equivalent and having no difference in all areas of BPMM. Action can't be inherited.

Define 2: Node =< NodeID, Name, Status, Async, TransitionLists, ActionLists, Timer, WeightLists, DependNodeIDs, ExceptionTransition, Owner, Goal >

Node is the business unit in the business process, and it is the carriers of relevant actions with business unitarity and some particular attributes itself; more complicated business course is realized through the node nestification. The node is one of the composition structure of the business process, the same node is not equivalent and having discrepancy in its function of different areas of BPMM, which depends on some restrictions of own attribute of the node, namely when changing in node input, its output is reloaded. The node uses the hollow circle to annotate.

- Public attribute: name, asynchronous, state, transition, action lists, timer, role, exception, weight lists
- Private attribute: depends

Define 3: Information =< InformationID, Name, Type, Contents, Operations >

Information relies on transition realizing input and output mainly message and event two kinds. Among them, message is mainly multimedia forms such as some briefing on texts, language, picture, audio and video,etc.; event is mainly some operation, function activity and corresponding movements sets. Information does not possess the initiative, so the function activity in the Event means to the demand of the activity lists under the particular incident or exports.

Define 4: Transition =< TransitionID, Name, PreviousNodeList, NextNodeList, WeightLists, InformationList >

Transition is the node connector, each of which is uniqueness annotated, and it is the carriers of relevant Informations. It is unidirectional, annotated by arrowhead.

Define 5: Process =< ProcessID, Name, NodeList, TransitionList, Description, Status >

Process is the sets of nodes and transitions, in view of the abnormal condition, state attribute (Status) set up for the process. Only complete process is considered while doing PM.

Define 6: Panorama Process Directed Digraph (PPDD) is the maximum directed digraph that contains all the process instances under the particular goal. PPDG is the handling tool of the model, which is the biggest collection of the process set. With the help from it, the certain process instance characteristic can be reflected from the training set effectively, and leave out the unnecessary process detail while offer the most direct, succinct support for better process assessment. When needing to carry on the process assessment, PPDD need to be produced at first, and when each process changed, PPDD should be reproduced before upgrading.

B. Elements Rules

The above is about definition and explanation of basic elements in the PAM. Considering the basic element mappings call for rules when carrying on the process mining, further rules among the elements about the logic connection is given below.

- Rule 1: Start Node and End Node do not include any of the activity. Start Node that owns no exception must go with some certain node. End Node has no exception and other node coming after it.
- Rule 2: The number of actions per node depends on the granularity chosen, usually, one node owns only one

action; Actions should be put into the same node when the procedure between some function activities and them can't be replaced.

C. Process Mapping in PAM

The model adopts XML language carried on the event logs based on daily record as its data source, which consults XES standard. The relationship between elements as shown below:

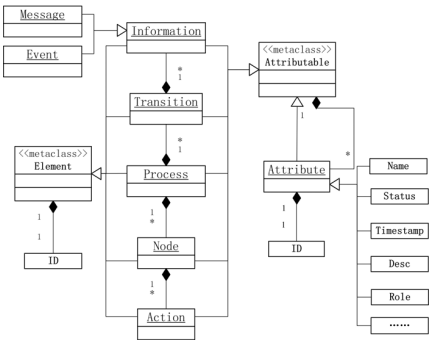

Figure 1. PAM's UML diagram

As to the affairs recorded in the event log, every event can successively be arranged in an order where the start and complete event tracked, merged into one Action carried in a Node.The migration information between different Node is recorded in one Transition, when there is no obvious information, such as restrain time being out, or only the order of activity between the two Nodes, empty Transition is provided. Through mapping, all information guaranteeing to be recorded in the event log can reflect in the model that is established on the basis of this text. The following picture is shown a simple example:

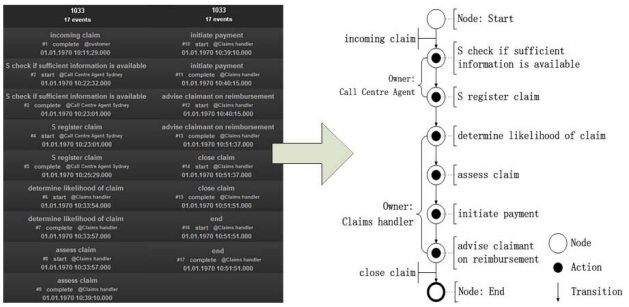

Figure 2. Mapping between Event log and PAM

D. PAM Initialization

Putting up the PPDD with decision tree algorithm, event log is mapped at first, while left out the incomplete process and the process waiting to be handled, and unify data of different form and style in the process.

Algorithm: PPDD Initialization

Input: Process Set P={P_1, P_2, P_3, P_4, …}, N'=Total number of Process, as to a certain process P_i, the method that is described by 2.1 part of this text, P_i={ N_1, N_2, N_3, N_4, …, T_1, T_2, T_3, T_4, …},Among them N is a node, T is transition.

Export: PPDD =< N, T >

```
Object<ProcessSet> P_done = null;
Object<Process> P_target = null;
ArrayList Ce_i = null;
for ( int i; i<N; i++ ){
    Object<ProcessSet> Temp = P - P_done;
    Random Choose Porcess Pi ∈Temp;
    // N_Temp is the amount of Element
    for ( int j; j<N_Temp; j++ ){
        Iterator it = Pi.iterator();
        while(it.hasNext()){
            If E_j ∈ P_target , Then skip;
            If E_j   P_target , Then add to
 P_target ;
            C_Ej++;
        }
    }
}
```

Process Ptarget is the initialized PPDG, strictly speaking, Ptarget is no longer a certain single process, among them variety situations will produce a plurality of branches to some treatment nodes in the process, but unified to the End Node finally. For better description of process set, avoiding violate the definition of the process, PPDD is provided. In theory, PPDD has the nodes and transitions of all process instances, but the inevitable occurring closed loop in it. Because the node is the basic function of handling process instance, the information among the nodes is transmitted by transitions, so when the closed loop appears, must caused by new process instance, but not new node, so that if PPDD will dispel the closed loop, all nodes and most transitions is in PPDD. The existence of the error of this place will be influenced and judged the result, so the best solution is to begin from the longest process instance, which usually includes more classical process elements.

E. Revise the PPDD

Adopt the classical ID3 algorithm to delete those nodes and transitions having less information gain by setting the support and confidence. What deserves to be mentioned is that it needs to be verified its related nodes and transitions while deleting a certain one, which guarantee the deletion of this element will not cause the arbitrary element besides this element to be unable to reach the End Node. Thus a modified PPDD is as follows:

PPDD = < $N_{modified}$, $T_{modified}$ >

$N_{modified}$ stores the node revised, $N_{modified}$={ N_1, N_2, N_3, ... N_i, C_{n1}, C_{n2}, C_{n3}, ... , C_{ni} }. As to a certain N_i, the corresponding C_i is its counting in the process instance set. Specially set N_1 as the Start Node, N_i as the End Node, and then $C_1 \geqslant C_i$, and max{C_1,C_2,C_3,...,C_i}= C_1 = The total number of process instances, the size of the train set. Consider the node may include the sub nodes, but for particular analysis, sub node has

no effect on its father node, so further analysis is no longer made from it.

$T_{modified}$ stores the revised transitions, $T_{modified}$ ={ T_1, T_2, T_3, ... T_j, C_{t1}, C_{t2}, C_{t3}, ... , C_{tj} }, As to T_j, its corresponding C_{tj} is its counting in the process instance set. When there is a closed loop, the transition having the lowest C_{tj} will be deleted.

IV. SELECTING ASSESSMENT INDICATORS

Assessment system is constructed by PAM, aim at using element and internal stipulations defined in model, realizing the assessment function of the model. This section provides the way choosing assessment indexes at first, and gives the weight vector depending on it, which improve the PPDG before that turned into Weighted Panorama Process Directed Digraph (WPPDD) given finally to realize the assessment in BI environment.

The choice of assessing index is arbitrary, which needs testers to filter, alter by hand according to the personal experience. Here describes the steps about selecting assessment indicators, making it satisfied and fixed and guaranteeing them to be chosen in minimum relationship, avoid overlapping.

a) List out the assessment indexes as e1, e2, e3 To get accurate effects about assessment, put the indexes to different divide levels order by attributes from top to bottom. Besides the root level, the higher level node includes the set of all the child node's indicators.

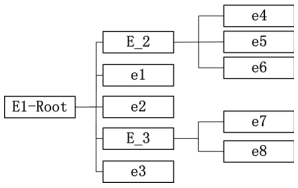

Figure 3. Goals set by levels

E_1 is a root goal, includes E_2, e1, e2, E_3, e3 five assessment indexes, and index E_2, E_3 is the index on the senior level, include the sub goal separately.

b) Construct the pair comparative matrix which list out the indicators on the same level and compare in pairs to get the digit, namely starting from the second level (the one after root node). The value of a_{ij} in the matrix is the weight ratio of the i-th and j-th element whose range is signless integral and reciprocal. When it a_{ij} =1means element i and j has the same importance. The pair comparative matrix is shown as follows:

$$\begin{pmatrix} 1 & 4 & 5 & 8 & 2 \\ 1/4 & 1 & 5/4 & 2 & 1/2 \\ 1/5 & 4/5 & 1 & 8/5 & 2/5 \\ 1/8 & 1/2 & 5/8 & 1 & 1/4 \\ 1/2 & 2 & 5/2 & 4 & 1 \end{pmatrix}$$

c) Calculate the Weight vector and Consistency test.

Calculate the inconsistency index (CI) of the pair comparative matrix M (n>1):

$$CI = \frac{\lambda_{max}(A) - n}{n - 1} \tag{1}$$

λ_{max} is the max eigenvalue of matrix M.The mean random consistency index (RI) only related to matrix order used for consistency test can be checked from material directly.The formula of random coincidence coefficient (CR) for the pair comparative matrix is shown as below:

$$CR = \frac{CI}{RI} \tag{2}$$

As CR<0.1, the pairs comparative matrix M has satisfactory consistency, or its inconsistent intensity can be accepted, that is to say the related intensity between the indexes chosen is minimum; Otherwise adjust should be done until reach satisfactory consistency.

d) Calculate the weight vector: According to the method mentioned above, the eigenvector W=(-0.8549, -0.2137, -0.1710, -0.1069, -0.4275)^T. Standardize this vector: Make their every weight greater than zero, the sum of every weight equals 1. The changed vector W'= (0.4819, 0.1205, 0.0964, 0.0602, 0.2410)^T is called weight vector which reflects the indicators importance order.

V. CASE ANALYSIS

A. Assessment Method

After given the PPDD from process instance set, use the AHP to determine the assessment indicators and calculate the weight vector, and then get the WPPDD, the next step is to assess the certain process instance and show the result.

The formula designed according to the vector model as follows:

$$\mathrm{Score}(P_{input}, PPDG) = \mathrm{amtProportion}(P_{input}, PPDG) \cdot \sum_{E \in P_{input}} (\mathrm{Ef}(E \in PPDG) \cdot \mathrm{fProcess}(E)^2 \cdot \mathrm{Eva}(E)) \tag{3}$$

The function amtProportion (P_{input}, PPDD), in charge of global scale control, is a increasing function that returns higher value when there is more elements in P_{input}. The integer K is an adjustment factor chosen by the user.

$$\text{amtProportion (} P_{input}, \text{ PPDD) } = K*|P_{input}|/C_1 \tag{4}$$

The function Ef is to show the proportion of the P_{input} element set in the WPPDD in order to assess the frequency of the elements in the P_{input} appearing in WPPDD.

$$Ef = n/N \tag{5}$$

The function fProcess(E) is to calculate the proportion of the process having the certain elements in the whole process set. It will decrease the high frequency of certain element appearing in the process instance set. N_{pinE} stands for the amount of process having the element E;

$$fProcess(E) = 1+\log(N_{pinE}/[(N_{pinE}/C_1)+1]) \quad (6)$$

The function Eva(E) is the weight vector calculated by section IV which help the researcher to analyse the importance among different indexes.

In general, the higher result gotten by the Score function means the better process it is.

B. *Case Study*

The example describes the use of an insurance company's claims process, a total of 46138 events, 3512 process instance. Since the process is relatively simple, a total of eight basic types of Actions that carried on one Node each besides the Start Node and End Node, ten Transitions is worked out after mapping.

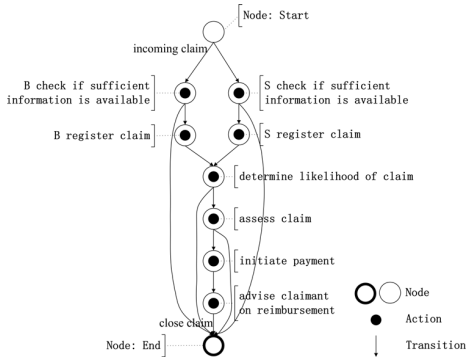

Figure 4. The example's PPDD

After calling PPDD initialization algorithm, the number each node and transition will be counted, where the noisy data omitted, shown as follow:

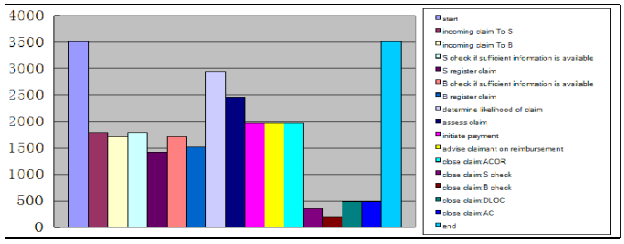

Figure 5. The distribution of each element

Here gives two test process instances:

$P_1=\{N_{start}, T_{incoming\ claim\ to\ s}, N_{s\ check}, T_{close\ claim\ s\ check}, N_{end}\}$

$P_2=\{N_{start}, T_{incoming\ claim\ to\ s}, N_{s\ check}, N_{register\ claim}, N_{determine\ likelihood\ of\ claim}, N_{assess\ claim}, N_{initiate\ payment}, N_{advise...}, T_{close\ claim\ ACOR}, N_{end}\}$

Use the weight vector generated in Section IV, combined with the proposed formula Score function, the calculation of the above process instance can be drawn as:

Score（P_1）=0.167, Score（P_2）=0.916

In the process P_1, the customer's information is incomplete, so the claims process quickly terminated; and process P_2, provide a complete and effective information that comply with the relevant requirements, so that the subsequent application for settlement of the claim, including registration, processing, payment, advice claimant on reimbursement go on wheels. The score obtained by calculating reflects both the importance of the two process instances. The latter because it involves the entire process flow, and more complete, so the higher score is calculated, while the former process because it involves fewer process elements which results in a smaller contribution, so that the smaller score obtained.

VI. CONCLUSION

Empirical results which is presented as numerical data by the formula show that it can be set as a specific target under the new process recorded in event log for effective assessment and be displayed directly for its consequence; by setting the threshold value, process screening and classification, including excavation work for the follow-up that need evaluation can be based on the results more easily which is useful for further process mining work.

ACKNOWLEDGMENT

Part of this work is supported by the National Science Foundation(61173140), the Science Foundation of Shandong Province Project (Y2008G28), and the Independent Innovation Foundation of Shandong University (2010JC010).

REFERENCES

[1] W.M.P. van der Aalst and K.M. van Hee. Workflow Management: Models, Methods, and Systems. MIT Press: Cambridge, MA, 2002.

[2] W.M.P. van der Aalst, A.J.M.M. Weijters. Process mining: a research agenda. Computes in Industry, 2004, 53:231-244.

[3] B.F. van Dongen, W.M.P. van der Aalst, A meta model for process mining data, in: J. Casto, E. Teniente (Eds.), Proceedings of the CAiSE'05 Workshops (EMOI-INTEROP Workshop), vol. 2, FEUP, Porto, Portugal, 2005, pp. 309–320.

[4] XU Yan, TAN Pei-qian.Study on Process Mining. LOGISTICS SCI TECH. Vol 29, No. 128.

[5] Moghimi F., Zheng C. A Decision-making Model to Choose Business Intelligence Platforms for Organizations. Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on.

[6] Margaret H. Dunham. Data Mining. Beijing:TSINGHUA UNIVERSITY PRESS. 2005.

[7] Zhongxin Wu. Lucene Analysis and Application. Beijing:China Machine Press. 2008.9.

[8] Qi Hu.jBPM 4 Workflow Application Development Guide. Beijing: Publishing House of Electronics Industry. 2010.

[9] Ian H. Witten & Eibe Frank. Data Mining. Beijing:China Machine Press. 2006.10.

[10] Quinlan J R. Introduction of decision trees. Machine Learning. 1986(1): 84-100.

[11] Wil M.P. van der Aalst. Process Mining. Srpinger. 2011.