# Activity Mining: Challenges and Prospects*

Longbing Cao

Faculty of Information Technology, University of Technology Sydney, Australia
lbcao@it.uts.edu.au

**Abstract.** Activity data accumulated in real life, e.g. in terrorist activities and fraudulent customer contacts, presents special structural and semantic complexities. However, it may lead to or be associated with significant business impacts. For instance, a series of terrorist activities may trigger a disaster to the society, large amounts of fraudulent activities in social security program may result in huge government customer debt. Mining such data challenges the existing KDD research in aspects such as unbalanced data distribution and impact-targeted pattern mining. This paper investigates the characteristics and challenges of activity data, and the methodologies and tasks of activity mining. Activity mining aims to discover impact-targeted activity patterns in huge volumes of unbalanced activity transactions. Activity patterns identified can prevent disastrous events or improve business decision making and processes. We illustrate issues and prospects in mining governmental customer contacts.

**Keywords:** Activity data, activity mining, impact-targeted mining, unbalanced data.

## 1 Introduction

Activities can be widely seen in many areas, and may lead to or be associated with impact to the world. For instance, terrorists undertake a series of terrorist activities which finally lead to a disaster to our society [8]. In social security network, a large proportion of separated fraudulent activities can result in huge volumes of governmental customer debt [3]. In addition, activity data may be found in business world with frequent customer contacts [10], business intervention and events, and business outcome oriented processes, as well as event data [12], national and homeland security activities [8] and criminal activities [7]. Such activities are recorded and accumulated in relevant enterprise activity transactional files. Activity data hides rich information about the relations between activities, between activities and operators, and about the impacts of activities or activity sequences on business outcomes. Activity data may enclose unexpected and interesting knowledge about the optimum decision making and processes which may result in low risk of negative impact. Therefore, it is significant to study activity patterns and impact-targeted activity behavior.

* This work is sponsored by Australian Research Council Discovery Grant (DP0667060), China Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01), and UTS ECRG and Chancellor grants.

X. Li, O.R. Zaiane, and Z. Li (Eds.): ADMA 2006, LNAI 4093, pp. 582 – 593, 2006.

Activity data embodies organizational, information and application constraints, and impact-targeted multi-dimensional complexities which combine those from temporal, spatial, syntactic and semantic perspectives. As a result, activity data presents special structure and semantic complexities. For instance, variant characteristics such as sequential, concurrent and causal relationships may exist between activities. Activity data usually presents unbalanced distribution. As a result, many existing techniques cannot be used directly, which rarely cares for the impact of mined objects.

Therefore, new data mining methodology and techniques need to be developed to preprocess and explore activity data. This leads to *activity mining*. This paper discusses the challenges and prospects in building up effective methodologies and techniques to mine interesting activity patterns. *Activity mining* aims to discover rare but significant impact-targeted activity patterns in unbalanced activity data, such as frequent activity patterns, sequential activity patterns, impact-oriented activity patterns, impact-contrasted activity patterns, and impact-reversed activity patterns. The identified activity patterns may inform risk-based decision making in terms of predicting and preventing the happenings of targeted activity impact, maintaining business goals, and optimizing business rules and processes, etc.

The remainder of this paper is organized as follows. Section 2 presents the scenario and characteristics of activity transactional data and its challenges to the existing KDD. Section 3 discusses possible activity mining methodologies. Activity mining tasks are discussed in Section 4. Finally, Section 5 concludes this paper.

## 2 Activity Data

This section introduces an example and the characteristics of activity data, and further build up an activity model representing and defining activity data. Challenges of activity data on the existing KDD are discussed further.

### 2.1 An Example

Here we illustrate the activities in social security network. In the process of delivering Government social security service to the population, large volumes of customers contact governmental service agencies [3]. For instance, over 6 million Australians access Centrelink's services at one point in time. Every single contact, e.g., a circumstance change, may trigger a sequence of activities running serially or in parallel. Among them, some are associated with fraudulent actions and result in government customer debt. For example, Table 1 lists an excerpt of activity transactions [2] relevant to a scenario of changing customer address in Centrelink. When a Newstart (NSA) benefit recipient $i$ reports his/her circumstance $C_{i\_1}$ to Centrelink, an officer conducts activity $A_{i\_2}$ and $A_{i\_3}$ to check and update $i$'s entitlement and details. In parallel, the officer also conduct activity $A_{pi\_3}$ and consequently activities $A_{pi\_4}$ and $A_{pi\_5}$ to inspect $i$'s partner $j$'s details and possible debts. Concurrently, customer $i$'s task $T_{i\_1.1}$ triggers $A_{k\_2}$ on $i$'s NSA co-tenant $k$. $A_{k\_2}$ further triggers $A_{k\_3}$, $A_{k\_4}$ and $A_{k\_5}$ on $k$ to reassess and update $k$'s rent details and possible debts.

**Table 1.** Activity transactional data

| **Customer *i* = Newstart (NSA) recipient**[1] | **Partner *j***[2] | **Customer *k* = NSA recipient** |
|---|---|---|
| Circumstance $C_{i\_1}$ = Change of Address | | |
| $A_{i\_2}$ = Accelerated Client Matching System review (triggered by $T_{i\_1.1}$) | | $A_{k\_2}$ = Accelerated Client Matching System review (by $T_{i\_1.1}$ on customer *j*) |
| $T_{i\_2.1}$ = Letter to Customer | | $T_{k\_2.1}$ = Letter to Customer |
| $A_{i\_3}$ = Customer Details Update | Parallel Activity $A_{pi\_3}$ = Customer Details Update | $A_{k\_3}$ = Customer Details Update |
| $T_{i\_3.1}$ = Change Rent Details; | $T_{pi\_3.1}$ = Change Rent Details; | $T_{k\_3.1}$ = Change Rent Details; |
| $T_{i\_3.2}$ = Change Home-ownership Details. | $T_{pi\_3.2}$ = Change Home-ownership Details. | $T_{k\_3.2}$ = Change Home-ownership Details. |
| $T_{i\_3.3}$ = NSA Reassessment | $T_{pi\_3.3}$ = PPP Reassessment | $T_{k\_3.3}$ = NSA Reassessment |
| $E_{i\_3.1}$ (from $T_{i\_3.1}$ & $T_{i\_3.2}$) = NSA Reassessment | $T_{pi\_3.4}$ = FTB Reassessment | $E_{k\_3.1}$ (from $T_{k\_3.1}$ & $T_{k\_3.2}$) = NSA Reassessment |
| | $E_{pi\_3.1}$ (from $T_{pi\_3.1}$) = FTB Reassessment | $E_{k\_3.2}$ (from $T_{k\_3.1}$) = Transfer NSA rate variation data to Debt Management System |
| | $E_{pi\_3.2}$ (from $T_{pi\_3.1}$) = Transfer FTB rate variation data to Debt Management System | |
| | $A_{pi\_4}$ (from $T_{pi\_3.1}$ *of* $A_{pi\_3}$) = New Debt | $A_{k\_4}$ (from $T_{k\_3.1}$ of $A_{k\_3}$) = New Debt |
| | $T_{pi\_4.1}$ = Raise debt | $T_{k\_4.1}$ = Raise debt |
| | $E_{pi\_4.1}$ = Profiling Assessment | $E_{k\_4.1}$ = Profiling Assessment |
| | $A_{pi\_5}$ (from $T_{pi\_4.1}$ of $A_{pi\_4}$) = Debt | $A_{k\_5}$ (from $T_{k\_4.1}$ of $A_{k\_4}$) = Debt |
| | $T_{pi\_5.1}$ = Withholdings | $T_{k\_5.1}$ = Withholdings |

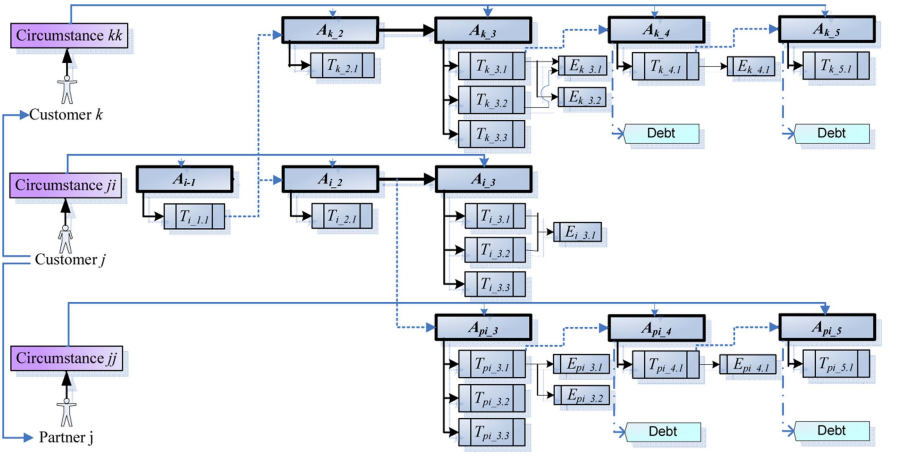


**Fig. 1.** Activity scenario diagram

[1] **Note:** In this case Rent Assistance will be paid as part of the partner's FTB in Centrelink business.

[2] = Parenting Payment – Partnered (PPP) & Family Tax Benefit (FTB) recipient.

In this example, an activity may further trigger one to many tasks. A task may trigger another activity on the same or different customer or some follow-up events (e.g., $E_{i\_3.1}$ from tasks $T_{i\_3.1}$ & $T_{i\_3.2}$). With respect to time frame, parallel or serial activities may run dependently or independently. For instance, parallel activity $A_{pi\_3}$ depends on $A_{i\_3}$ while parallel activity $A_{pi\_4}$ and concurrent activity $A_{k\_5}$ run independently even though the activities on customer $i$ are completed. Another interesting point is that some activities may generate impacts on business outcomes such as raising debts (e.g., $A_{pi\_5}$ and $A_{k\_5}$). Such debt-oriented activities are worthy of further identification so that debt can be better prevented and predicted.

## 2.2 Activity Model

The term "Activity" is an informative entity embracing both business and technical meanings. In business situations, an activity is a business unit of work. It corresponds to one to many activity operators to conduct certain business arrangement forming a workflow or process. It directly or indirectly satisfies certain organizational constraints and business rules. Technically, an activity refers to one to several transactions recording information related to a business unit of work. Therefore, an activity may undertake certain business actions, embody business processes, and trigger some impact on business. Moreover, activity transactions embed much richer information about business environment, causes, effects and dynamics of activities and potential impact on business, as well as hidden information about the dynamics and impact of activities on debts and activity operator circumstances. In general, an activity records information about *who* (maybe multiple operators) processes *what types* of activities (say change of address) from *where* (say customer service centres) and for *what reasons* (say the action of receipt of source documents) at *what time* (date and time), as well as resulting in *what outcomes* (say raising or recovering debt).

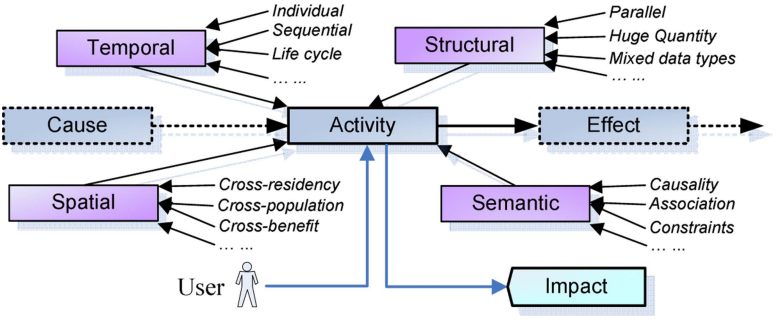


**Fig. 2.** Activity model

Based on the understanding of the structural and semantic relationships existing in activity transactions, we generate an abstract *activity model* as shown in Figure 2. An activity is a multi-element entity $A = (C, E, U, I, F)$, where $C$ and $E$ are *cause* triggering and *effect* triggered by the activity A, respectively. An activity either is operated by or act on one or multiple *users* $U$. It may directly or indirectly generate

*impact I* on business outcomes such as leading to debt or costing. In particular, an activity and its sequence present complex features in terms of *temporal*, *spatial*, *structural* and *semantic* dimensions. For each of the four dimensions, activity presents various observations which make activity mining very complicated. For instance, each activity has a life cycle starting from registration by a user and ending via completion on the same day or some time later on. During its evolution period, it may be triggered, restarted, held, frozen, deleted or amended for some reason.

### 2.3 Challenges

Activity data proposes the following challenges to existing KDD approaches.

- Activities of interest to business needs are *impact-oriented*. Impact-oriented activities refer to those directly or indirectly lead to or are associated with certain impact on business situations, say fraudulent social security activities resulting in government customer debt. Therefore, *activity mining aims to discover specific activities of high or low risk associated with business impact*, which we call *impact-targeted activity pattern mining*. While the existing KDD research rarely deal with impact-targeted activity pattern mining.
- Impact-oriented activities are usually a very small portion of the whole activity population. For instance, fraudulent activities in Centrelink only account for 4% of all activities. This leads to an *unbalanced class distribution* [15] of activity data, which means positive target-related activity class is only a very small fraction of the whole data set. Unbalanced class distribution of activity data proposes challenges to impact-targeted activity pattern mining in aspects such as activity sequence construction, pattern mining algorithms and interestingness design and evalution.
- Among activities, some occur more often than others. This indicates an *unbalanced item distribution* of activity item set. Unbalanced item distribution also affects activity sequence construction, pattern mining algorithms and interestingness evaluation.
- In analyzing impact-targeted activities, *positive* and *negative* impact-targeted activity patterns can be considered, which correspond to positive and negative activity patterns. Other forms of activity patterns include *sequential activity patterns*, activity patterns representing the contrast of impact (called *contrast activity patterns*) and the reversal of impact (named *reverse activity patterns*), etc.
- In constructing, modeling and evaluating activity patterns, *constraints* from aspects such as targeted impact, distributed data sources and business rules must be considered. Real-world *constraint-based* activity pattern mining is more or less domain-driven [1]. Constraint-based mining and domain-driven data mining should be taken into account in mining impact-targeted activity patterns.
- Activities present *spatial-temporal* features such as sequential, parallel, iterative and cyclic aspects, as well as crossing benefits, residencies and regions. For instance, an activity may trigger one to many serial or parallel tasks and certain corresponding events, and generate complex action sequences.

The complexities of activity data differentiate it from normal data sets such as those in event [4], process [13] and workflow [5] mining, where data is much flat and

simple. Those mainly studies process modeling and has nothing to do with complex activity structure and business impacts of activities. Therefore, new methodologies, techniques and algorithms must be developed.

## 3 Activity Mining Methodologies

### 3.1 Activity Mining Framework

In developing activity mining methodologies, we first focus on understanding activity data and designing a framework for activity mining.

In business world, activities are driven by or associated with business rules [2]. For instance, the activity sequences triggered by changing address (see Figure 1) present interesting internal structure. Activity $A_{i-1}$ triggers $A_{i-2}$ and $A_{k-2}$ in parallel, while two series of activities $A_k$ and $A_{pi}$ go ahead after the completion of original activity sequence $A_i$. This example shows that there may exist meta-patterns in activity transactions, which are helpful for further understanding and supervising activity pattern mining. For instance, fundamental activity meta-patterns such as *serial* ($x$ --> $y$), *parallel* ($x \parallel y$), *cyclic* ($x$ --> $x$ or $x$-->$z$-->$x$) and *causal* ($x$ => $z$) may exist between activities $x$, $y$ and $z$. These meta-patterns if identified can supervise further activity pattern learning. Temporal logic-based ontology specifications can be developed to represent and transform activity metapatterns.

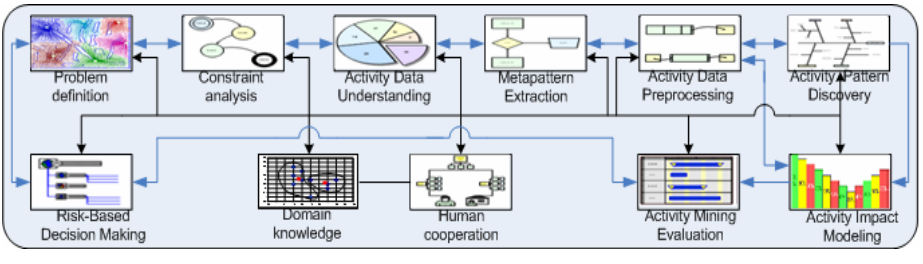


**Fig. 3.** An activity transaction mining framework

Based on the above activity data understanding, we can study a proper framework for activity mining. Figure 3 illustrates a high-level process of activity mining. It starts from understanding activity constraints, data and meta-patterns. The results are used for preprocessing activity data by developing activity preparation techniques. Then we design effective techniques and algorithms to discover interesting activity processing patterns and model activity impacts on debts. Further work is to evaluate the performance of activity analysis. Finally, we integrate the above results into an activity mining system, and deploy them into strengthening risk-based decision making in debt prevention. It is worth noting that there may be back and forth among some of the above steps. Additionally, domain analysts and knowledge are essential for iterative refinement and improving the workable capability of mining results.

### 3.2 Activity Mining Approaches

Due to the closely coupled relationship between activities, activity users and impact on business, we need to combine them to undertake systematic analysis of activity data. This is different from normal data mining which usually only focus on some aspect of the problem, e.g., process mining focuses on business event and workflow analysis. Driven by business rules and impact, we can highlight some key aspect and undertake activity mining in terms of *activity-centric analysis*, *impact-centric analysis* and *customer-centric analysis*. We introduce them individually with regard to the example of analysing activities in governmental customer debt prevention.

*Activity-centric analysis.* Activity-centric analysis focuses on analysing activity patterns, namely the relationships between activities. For instance, we can conduct activity centric debt modelling in terms of the following aspects: (1) Pattern analyses of activities which have or haven not led to debts, (2) Activity process modelling, and (3) Activity monitoring.

*Impact-centric analysis.* Impact-centric analysis attempts to analyse the impact of activities and activity sequences on business such as debts, as well as optimizes activities and processes to reduce the negative impact of activities/processes on business. The major research includes (1) Analysing the impacts of a type of notifiable events or a class of relevant activity sequence against debt outcomes, (2) Risk/cost modelling of activities which may or may not lead to debts, and (3) Activity/process optimization.

*Customer-centric analysis.* Customer-centric analysis studies the patterns of activity operators' circumstances and circumstance changes which may or may not lead to debt in aspects such as (1) Circumstance profiling, (2) Officer behaviour analysis and (3) Customer behaviour analysis.

We further discuss these approaches by illustrating potential business problems in governmental customer debt prevention in Section 4.

## 4 Activity Mining Tasks

The major challenge of mining activity transactions come from the following processes in mining activity transactions: (1) activity preprocessing, (2) activity pattern mining, (3) activity impact modeling, and (4) activity mining evaluation.

### 4.1 Activity Preprocessing

The characteristics of activity transactional data make activity preprocessing very essential and challenging. The tasks include developing proper techniques to (1) improve data quality, (2) handle mixed data types, (3) deal with unbalanced data, (4) perform activity aggregation and sequence construction, etc.

*Unbalanced data.* As shown in Figure 4, activity data presents unbalanced class distribution (e.g., the whole set $|A|$ is divided into $|\$|$ as debt-related activity set while $|\bar{\$}|$ as non-debt set) and unbalanced item distribution. Unbalanced data mainly affect the performance and evaluation of traditional KDD approaches. Therefore, in activity preprocessing, effective methods and strategies must be considered to balance the affection of data imbalance. For balancing the impact of unbalanced class distribution,

techniques such as equal sampling in separated data sets, redefining interestingness measures such as replacing global support with local support in individual sets can be used. With respect to the imbalance of activity items, domain knowledge and domain experts must be involved to determine what strategies should be taken to balance the impact of some high proportional items. Their impact may be balanced by deleting or aggregating some items and designing interestingness measures.

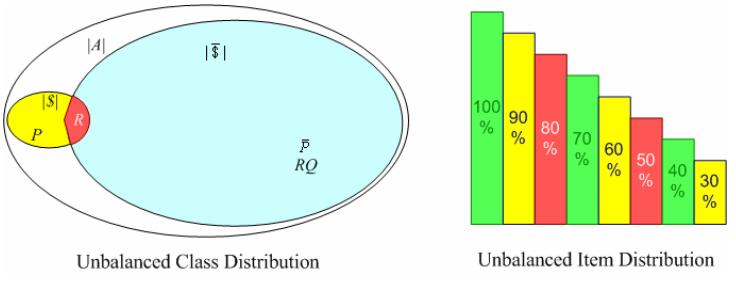


**Fig. 4.** Unbalanced activity data

*Activity sequence construction.* It is challenging to construct reliable activity sequences. The performance of activity sequences greatly affects the performance of activity modeling and evaluation. Different sliding window strategies can be used and correspondingly generate varied activity sequences. For instance, the activity series in Figure 5 could be constructed or rewritten into varied activity sequences, say the sequence for $d_2$-related activities could be $S_1$: {$a_8$, $a_9$, $a_{10}$, $a_{11}$, $a_{12}$, $a_{13}$, $d_2$}, $S_2$: {$a_7$, $a_8$, $a_9$, $a_{10}$, $a_{11}$, $a_{12}$, $a_{13}$, $d_2$}, $S_3$: { $a_{11}$, $a_{12}$, $a_{13}$, $d_2$, $a_{14}$, $a_{15}$ }, etc. The design of sliding window strategies must be based on domain problems, business rules and discussion with domain experts. $S_1$ considers a fixed window, $S_2$ may cover the whole debt period, while $S_3$ account for the further effect of $d_2$ on activities. Domain knowledge plays an important role in determining which one of the three strategies makes sense.

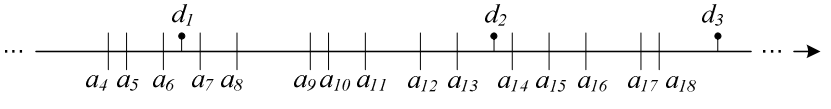


**Fig. 5.** Constructing activity sequences

## 4.2 Activity Pattern Mining

Impact-targeted activities are usually mixed with customer circumstances and business impact. Therefore, activity pattern mining is a process of mining interesting activity processing and user behavior patterns based on different focuses such as *activity-centric*, *impact-centric* and *customer-centric* analyses. As shown in Table 2, activity pattern mining aims to identify *risk factors* and *risk groups* highly or seldom related to concerned business impact by linking activity, impact and customer files together.

**Table 2.** Impact-targeted activity pattern mining

| Risk level | Risk factor | | Risk group | |
|---|---|---|---|---|
| High | Activity features | Customer circumstances | Activity processing patterns | Customer behavior patterns |
| Low | | | | |

Impact-targeted risk factors include major activity features and customer circumstances at high or low risk of leading to or being related to targeted business impact. For instance, in social security network, the activity "reassessing benefit" is found highly correlated with leading to debt. Impact-targeted risk groups target identifying activity processing patterns or customer behavior patterns more or less resulting in targeted impact. For example, frequent activity sequences are likely associated with government customer debt. In the following, we illustrate some novel impact-targeted activity patterns.

*Customer-centric activity analysis* mainly investigates user decision-making behavior and profiling as well as the impact of a user's or a class of users' actions on related stakeholders. This identifies officer/customer's demographics and profiling leading to debts, e.g., studying the impact of staff proactive actions on debt compared with passive and customer-triggered activities, or the impact of face-to-face dealings vs. technology-based contacts. We can develop classification methods for debt-related customer segmentation. Classification methods based on logistic regression tree and temporal decision tree can be studied via considering temporal factors in learning debt/no-debt, low-debt/high-debt and debt reason patterns. The results of frequent, sequential and causal activity patterns can benefit the analysis of customer demographics and circumstances leading to debts.

*Activity-centric analysis* focuses on inspecting relations between debt-related activities. This includes mining activity patterns such as frequent, sequential [6] and causal ones in constrained scenarios. To mine frequent patterns, association rule method can be expanded to discover temporally associated [14] activities by recoding activity records and developing new measures and negative association rules. Those frequent activity patterns can be identified leading to debt, no debt or debt completion. Negative associations such as "*if* activity *a* and *b* but not *c then* debt" can be studied. Further, based on the constructed sequences of activities, sequential activity patterns in or crossing activity sequences can be investigated by considering temporal relations between activities. We can test various sequence combinations based on different sliding window strategies, and incorporate the identified meta-patterns and frequent patterns into sequential activity mining. In addition, there exists certain causal relation [9] in activity sequences. Causal pattern mining aims to find and explain contiguity relations between activities and between activities and debt reasons/state changes. Determinant underlying causal pattern, relational casual pattern and probabilistic causal patterns can be analyzed by considering aspects like activity forming, contiguity and interaction between activities and spatial-temporal features of activities and debt-related activity sequences.

**Table 3.** Impact-targeted activity patterns

| Activity pattern | | Explanation |
|---|---|---|
| Frequent activity patterns | Positive associations | Activity associations $P$ related to impact \$: $P \rightarrow \$$ |
| | Negative associations | $P$ related to non-impact $\overline{\$}$: $P \rightarrow \overline{\$}$ |
| | Positive sequences | Activity sequences $P$ related to impact \$: $P \rightarrow \$$ |
| | Negative sequences | Activity sequences $P$ related to non-impact $\overline{\$}$: $P \rightarrow \overline{\$}$ |
| Contrast activity patterns | Contrast associations | $P$ related to impact \$: $P \rightarrow \$$ in impact data set; $P$ also associated with non-impact $\overline{\$}$: $P \rightarrow \overline{\$}$ in non-impact data set |
| | Contrast sequences | |
| Reverse activity patterns | Reverse associations | $P$ related to impact \$: $P \rightarrow \$$ in impact data set, while $\{P,Q\}$ associated with non-impact $\overline{\$}$: $P,Q \rightarrow \overline{\$}$ in non-impact data set |
| | Reverse sequences | |

### 4.3 Activity Mining Evaluation

It is essential to specify proper mechanisms [11] for evaluating the workable capability of identified activity patterns and risk models. Technically, we implement *impact-centric mining* which develops interestingness measures *tech_int*() in terms of particular activity mining methods. The debt preventable capability of the identified findings can also be assessed by checking the existing administrative/legal business rules and domain experts.

For technical evaluation, the existing interestingness can be testified and expanded or new measures may be designed to satisfy activity mining demand. In pilot analysis, the measures of some existing KDD methods are found invalid when deployed into activity mining, e.g., *support* may be too low to measure frequency. For newly developed activity mining methods, we can design specific measures by considering technical factors such as activity statistics, debt ratios and customer circumstance changes. On the other hand, the identified patterns and models can also be examined in terms of rigor and relevance to business factors such as business goals, significance, efficiency, risk to debts and cost-effectiveness. Additionally, measures themselves need be evaluated in terms of interpretability and actionable capability.

Further evaluation may be necessary by using significance test, cross-validation and ensemble from both business and technical perspectives. Under certain condition, it is useful to present an overall measurement of identified patterns by integrating interest from both technical and business perspectives. To this end, fuzzy set-based fuzzy aggregation and ranking may be useful to generate overall examination of the identified patterns and models. In addition, multiple measures may apply to one method. We can aggregate these various concerns to create an integrated measure for global assessment.

### 4.4 Matching List Between Activity Analysis Goals and Business Problems

The following Table 4 further explains them by illustrates some relevant business problems through observing the example of governmental customer debt prevention.

**Table 4.** Activity mining tasks

| | Activity analysis goals | Business problems |
|---|---|---|
| Activity pre-processing | (1) Activity data quality | How to identify wrongly coded activities and debts led by them? |
| | (2) Mixed data types | How to systematically analyze data mixing continuous, categorical and qualitative types? |
| | (3) Activity aggregation & sequence construction | How to aggregate sequence $A_i$ with its partnered one $A_{pi}$ into an integrated sequence in Fig. 1? |
| Activity-centric analysis | (4) Activity meta-pattern analysis (e.g., parallel, causal, cyclic metapatterns) | Can we find relations such as $x$ --> $y$, $x \parallel y$, $x$ --> $x$ or $x$-->$z$-->$x$ and $x$ => $z$ between activities $x$, $y$ and $z$? |
| | (5) Activity pattern analysis (e.g., frequent, sequential, causal patterns) | Can we find rules like "if activity $A_1$ then $A_2$ but no $A_5$ in the following 3 weeks → debt"? or "if $A_1$ triggers $A_2$, $A_2$ triggers $A_3$→ no debt"? |
| | (6) Activity process simulation and modelling (e.g., reconstruct processes) | Can we reconstruct some processes (activity series) based on activity transactions which may or may not lead to debts? |
| | (7) Activity replay and monitoring (e.g., generating recommendation or alerts) | Can we find knowledge like "If the activity $A_3$ is triggered, then generating an alert to remind the likely risk of this activity?" |
| Impact-centric analysis | (8) Activity impact analysis (e.g., the impact of an activity sequence on debt) | Is there correlation between activity types and debt types indicting what activity types more likely lead to certain types of debts? |
| | (9) Risk/cost modeling of activities (e.g., leading to debt or operational costs) | To what extent a certain activity/activity type/activity class will lead to certain type of debt? |
| | (10) Activity or process optimization | If activity $A_3$ is triggered, then recommending activity $A_6$ rather than $A_4$ then $A_5$, which will lead to low/no debts or the ending of debt? |
| Customer-centric analysis | (11) Operator circumstance profiling | What are the demographics of those customers who more likely lead to debt? |
| | (12) Officer behavior analysis | What are the impacts of those activities triggered by staff proactively compared with other passive activities and customer-triggered activities? |
| | (13) Customer behavior analysis | Whether face-to-face dealings lead to low debts compared with technology-based contacts such as by Internet/email? |
| Activity mining evaluation | (14) Technical objective & subjective measures | Are the existing technical objective measures ok when they are deployed to mine activity data? |
| | (15) Business objective & subjective measures | How to evaluate the impact of activity patterns on business? |
| | (16) Integrated evaluation of activity patterns | If technical interestingness clashes with business one, how to assess them? |

## 5 Conclusions and Future Work

Rare but significant impact-targeted activities differentiate from traditional mining objects in aspects such as closely targeting certain business impact. For instance, a series of dispersed terrorist activities may finally lead to a serious disaster to our society. Impact-targeted activity data presents special structure complexities such as

unbalanced class and item distribution. Mining rare but significant impact-targeted activity patterns in unbalanced data is very challenging. This paper analyzes the challenges and prospects of activity mining. We present an example to illustrate the complexities of activity data, and summarize possible impact-targeted activity pattern mining methodologies and tasks based on our practice in identifying fraudulent social security activities associated with government customer debt. In practice, activity mining can play an important role in many applications and business problems such as counter-terrorism, national and homeland security, distributed fraudulent and criminal mining. Techniques coming from impact-targeted activity mining can prevent disastrous events or improve business decision making and processes.

## Acknowledgement

Thanks are given to Dr Jie Chen, Mr Yanchang Zhao and Prof Chengqi Zhang at UTS for their technical discussion, as well as to Ms Yvonne Borrow, Mr Peter Newbigin and Mr Rick Schurmann at Centrelink Australia for their domain knowledge.

## References

1. Cao L, Zhang C. Domain-driven data mining: a practical methodology, *Int. J. of Data Warehousing and Mining*, 2006.
2. Centrelink. *Integrated activity management developer guide*, 1999.
3. Centrelink. *Centrelink annual report 2004-05*.
4. Guralnik V, Srivastava J. Event Detection from Time Series Data, *KDD-99*, 33-42.
5. Hammori M, Herbst J, Kleiner N. Interactive workflow mining—requirements, concepts and implementation, *Data & Knowledge Engineering,* 56 (2006) 41–63.
6. Han J., Pei J. and Yan X. Sequential Pattern Mining by Pattern-Growth: Principles and Extensions, in *Recent Advances in Data Mining and Granular Computing*, Springer Verlag, 2005.
7. Mena, J. *Investigative Data Mining for Security and Criminal Detection*, First Edition, Butterworth-Heinemann, 2003.
8. National Research Council, *Making the Nation Safer: The Role of Science and Technology in Countering Terrorism*, Nat'l Academy Press, 2002.
9. Pazzani M. A Computational Theory of Learning Causal Relationships, *Cognitive Science*, 15:401-424 1991.
10. Potts, W. *Survival Data Mining: Modeling Customer Event Histories*, 2006
11. Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems, *IEEE TKDE*, 8(6):970-974, 1996.
12. Skop, M. Survival analysis and event history analysis. © Michal Škop, 2005.
13. Van der Aalst W.M.P., Weijters A.J.M.M. Process mining: a research agenda, *Computers in Industry*, 53 (2004) 231–244.
14. Williams G, et al. Temporal Event Mining of Linked Medical Claims Data. *PAKDD03.*
15. Zhang, J., Bloedorn, E.; Rosen, L.; Venese, D. **Learning rules from highly unbalanced data sets,** *2004 ICDM Proceedings*, pp571 – 574.