ITDM

International Journal of INFORMATION TECHNOLOGY & DECISION MAKING

Volume 7
Number 2
June 2008

Activity Mining: From Activities to Actions

L. Cao, Y. Zhao, C. Zhang and H. Zhang

World Scientific
NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

International Journal of Information Technology & Decision Making
Vol. 7, No. 2 (2008) 259–273
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

ACTIVITY MINING: FROM ACTIVITIES TO ACTIONS*

LONGBING CAO[†], YANCHANG ZHAO, CHENGQI ZHANG
and HUAIFENG ZHANG

Faculty of Information Technology
University of Technology, Sydney
P.O. BOX 123, Broadway
NSW 2007, Australia
[†]*lbcao@it.uts.edu.au*

Activity data accumulated in real life, such as terrorist activities and governmental customer contacts, present special structural and semantic complexities. Activity data may lead to or be associated with significant business impacts, and result in important actions and decision making leading to business advantage. For instance, a series of terrorist activities may trigger a disaster to society, and large amounts of fraudulent activities in social security programs may result in huge government customer debt. Uncovering these activities or activity sequences can greatly evidence and/or enhance corresponding actions in business decisions. However, 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 based on case-study experience in the area of social security. Activity mining aims to discover high impact activity patterns in huge volumes of unbalanced activity transactions. Activity patterns identified can be used to prevent disastrous events or improve business decision making and processes. We illustrate the above issues and prospects in mining governmental customer contacts data to recover customer debt.

Keywords: Activity mining; impact mining; impact modeling; imbalanced data.

1. Introduction

Activities can be widely seen in many areas, and may lead to or be associated with certain impacts on the world. For instance, terrorists undertake a series of isolated terrorist activities which finally lead to a disaster in our society.[13] In social security networks, a large proportion of separated fraudulent activities can result in huge volumes of governmental customer debt.[7] In addition, activity data may be found in the business world with frequent customer contacts,[15] business intervention and events, and business outcome oriented processes, as well as event data,[17] national and homeland security activities[13] and criminal activities.[12] Such activities are

*This work is sponsored by Australian Research Council Discovery and Linkage grants (DP0667060, DP0773412, LP0775041).

recorded and accumulated in relevant enterprise activity transactional files. Activity data contains rich information about the relations between activities, between activities and operators, and about the impact of activities or activity sequences on business outcomes. Activity data may disclose unexpected and interesting knowledge about optimum decision making and processes which may result in a low risk of negative impact. Therefore, it is significant to study activity patterns and high impact activity behavior. Activity data embodies organizational, information and application constraints, and impact-oriented 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, because they rarely care for the impact of mined objects. Therefore, new data mining methodology and techniques need to be developed to preprocess and explore activity data. This promotes the research and development on *activity mining*. In this paper, we discuss 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 occurrences of certain types of activity impact, maintaining business goals, and optimizing business rules and processes. 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 Sec. 4. A case study of activity mining is presented in Sec. 5. Section 6 concludes this paper.

2. Activity Data

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

2.1. *An example*

Here, we illustrate activity data in the social security network. In the process of delivering government social security services to the public, large volumes of customers interact with governmental service agencies.[7] 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, Fig. 1 shows an excerpt of activity transactions[6]

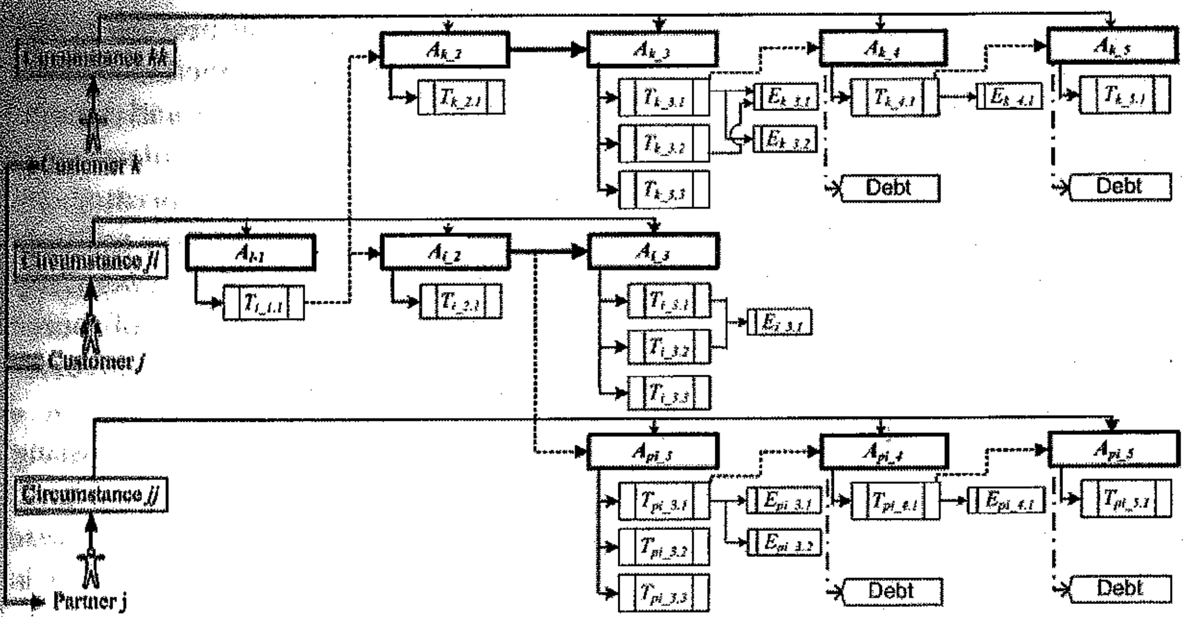


Fig. 1. Activity scenario diagram.

relevant to a scenario of changing customer address. When a benefit recipient i reports his/her circumstance C_{i_1}, 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 conducts 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 cotenant 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.

In this example, an activity may further trigger one or 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 $T_{i\_3.1}$ and $T_{i\_3.2}$). With respect to the 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 understood, and therefore predicted and prevented.

2.2. *Activity model*

The term "Activity" is an informative entity with both business and technical meanings. In business situations, an activity is a business unit of work. It corresponds to one or multiple activity operators conducting certain business arrangements forming a workflow or process. It directly or indirectly satisfies certain organizational constraints and business rules. Technically, an activity refers to one or 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 situations. Moreover, activity transactions embed much richer information about the 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 centers) 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).

Based on the understanding of the structural and semantic relationships existing in activity transactions, we generate an abstract *activity model* as shown in Fig. 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 acts on one or multiple *users* U. It may directly or indirectly generate *impact* I on business outcomes such as leading to debt or costs. In particular, activities and activity sequences 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 at some later time. During its evolution period, it may be triggered, restarted, held, frozen, deleted or amended for whatever reason.

2.3. *Challenges to KDD*

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 which lead to or are associated with certain impacts on business situations, say fraudulent social security activities resulting in government customer debt. Therefore, *activity mining aims to*

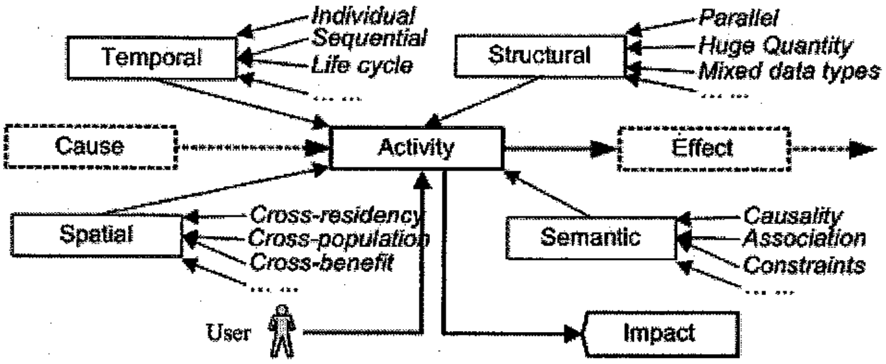


Fig. 2. Activity model.

discover specific activities of high or low risk associated with business impact, which we call impact-targeted activity pattern mining. However, the existing KDD research rarely deals with impact-targeted activity pattern mining.

- Impact-oriented activities are usually a very small portion of the whole activity population. This leads to an *unbalanced class distribution*[20] of activity data, which means positive impact-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. For example, there are some customer contact activities which run routinely or occur frequently in a short time period.
- When analyzing impact-targeted activities, *positive* and *negative* impact-oriented 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.
- When 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.[2] 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 characteristics crossing benefits, residencies and regions. For instance, an activity may trigger one or more serial or parallel tasks and certain corresponding events, generating complex action sequences.

The complexities of activity data differentiate it from normal data sets such as those in event,[9] process[1] and workflow[10] mining, where data is much flatter and simpler. Those mainly study process modeling and have 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*

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

In the business world, activities are driven by or associated with business rules.[6] For instance, the activity sequences triggered by changing address (see Fig. 1) present an interesting internal structure. Activity A_{i-1} triggers A_{i-2} and A_{k-2} in

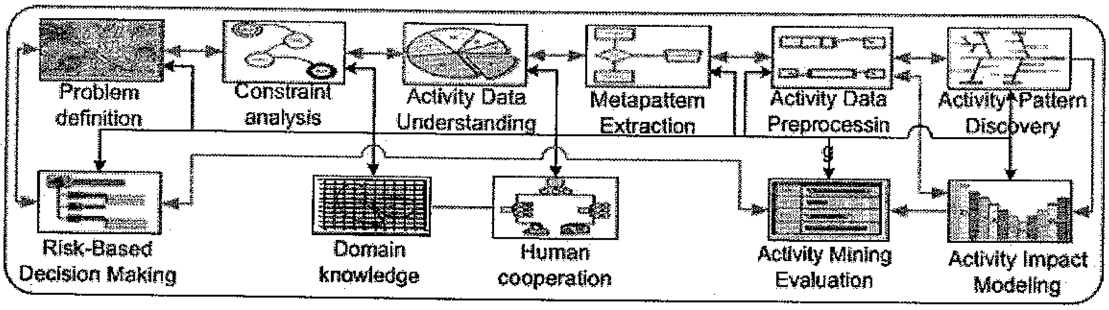

Problem definition
Constraint analysis
Activity Data Understanding
Metapattern Extraction
Activity Data Preprocessin
Activity Pattern Discovery
Risk-Based Decision Making
Domain knowledge
Human cooperation
Activity Mining Evaluation
Activity Impact Modeling

Fig. 3. An activity transaction mining framework.

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 meta-patterns may exist in activity transactions, which are helpful for further understanding and supervision of activity pattern mining. For instance, fundamental activity meta-patterns such as *serial* $(x \rightarrow y)$, *parallel* $(x||y)$, *cyclic* $(x \rightarrow x$ or $x \rightarrow z \rightarrow x)$ and *causal* $(x \Rightarrow z)$ may exist between activities x, y and z. These meta-patterns, if identified, can be used to guide further activity pattern learning. Temporal logic-based ontology specifications can be developed to represent and transform activity meta-patterns.

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 impact on business outcomes. Further work is needed 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 applications. It is worth noting that it may be necessary to move back and forth among some of the above steps. Additionally, domain experts and specific knowledge are essential for iterative refinement and improving the working 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 traditional data mining which usually only focuses on some aspects of the problem, e.g. process mining focuses on business events and workflow analysis. Driven by business rules and impact, we can highlight some key aspects and undertake activity mining in terms of *activity-centric analysis, impact-centric analysis* and *customer-centric analysis.*

Activity-centric analysis: Activity-centric analysis focuses on analyzing activity patterns, namely the relationships between activities. For instance, what activities frequently occur together? What activities usually occur before/after a specific activity? Activity centric debt modeling can be conducted in the following aspects: (1) pattern analyses of activities which have or have not led to positive/negative outcomes, (2) activity process modeling, and (3) activity monitoring.

Impact-centric analysis: Impact-centric analysis attempts to analyze the impact of activities and activity sequences on business, as well as optimize activities and processes to reduce the negative impact of activities/processes on business situations. For example, what activities will change the expected outcome? The major research includes: (1) analyzing the impacts of a type of notifiable events or a class of relevant activity sequence against debt outcomes, (2) risk/cost modeling of activities which may or may not lead to debts, and (3) activity/process optimization.

Customer-centric analysis: Customer-centric analysis studies the combination of customer circumstance patterns and activity patterns as well as their impact on business outcomes. For example, for a customer group with specific circumstances, what activities will lead to target and what activities will change the expected outcome? It includes (1) circumstance profiling, and (2) customer behavior analysis. We further discuss these approaches by illustrating potential business problems in governmental customer debt prevention in Sec. 4.

4. Activity Mining Tasks

The major challenges of mining activity transactions come from the following processes in mining activity transactions: (1) activity preprocessing, (2) activity-centric pattern mining, (3) impact-centric activity mining, (4) customer-centric activity mining, and (5) activity mining evaluation.

Table 1 further explains them by illustrating some relevant business problems through observing the example of governmental customer debt prevention.

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.

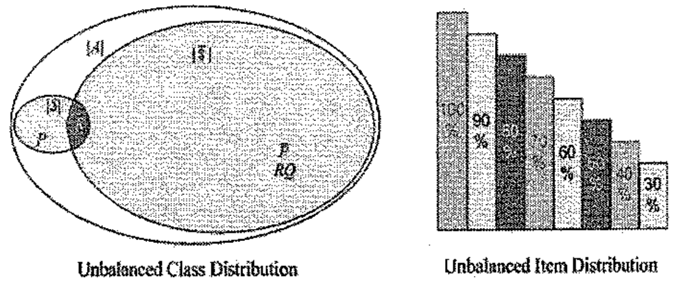
Unbalanced data: As shown in Fig. 4, activity data presents unbalanced class distribution (e.g. the whole set $|A|$ is divided into $|T|$ as debt-related activity set and $|\overline{T}|$ as non-debt set) and unbalanced item distribution. Unbalanced data mainly affect the performance and evaluation of traditional KDD approaches. Therefore, in activity preprocessing, effective methods and strategies must be considered to balance the effect of data imbalance. To balance 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 adopted to balance the impact of various high proportional items.

Table 1. Activity mining tasks.

	Activity Analysis Goals	Business Problems
Activity preprocessing	(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 meta-patterns)	Can we find relations such as $x \rightarrow y$, $x \| y$, $x \rightarrow z$ or $x \rightarrow z \rightarrow x$ and $x \Rightarrow z$ among 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 three weeks? debt?" or "if A_1 triggers A_2, A_2 triggers A_3? no-debt?"
	(6) Activity process simulation and modeling (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 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 are more likely to 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 a 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?

Table 1. (*Continued*)

	Activity Analysis Goals	Business Problems
Customer-centric analysis	(11) Operator circumstance profiling	What are the demographics of those customers which are 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 conflicts with business ones, how to assess them?



Unbalanced Class Distribution

Unbalanced Item Distribution

Fig. 4. Unbalanced activity data.

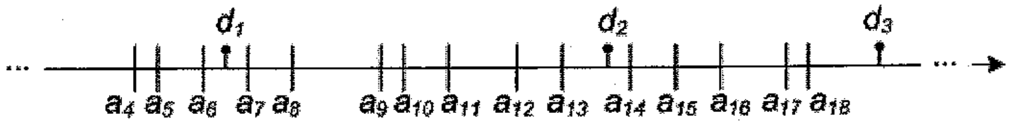


Fig. 5. Constructing activity sequences.

Their impact may be balanced by deleting or aggregating some items and designing interestingness measures.

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 various activity sequences. For instance, the activity series in Fig. 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.

4.2. *Activity pattern mining*

Impact-targeted activities are usually combined 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.

Activity-centric analysis focuses on inspecting relations between impact-related activities. This includes mining activity patterns such as frequent, sequential[11] and causal ones in constrained scenarios. To mine frequent patterns, association rule

Table 2. Impact-targeted activity pattern risk.

Risk Level	Risk Factor		Risk Group	
High Low	Activity features	Customer circumstances	Activity processing patterns	Customer behavior patterns

method can be expanded to discover temporally associated[18] activities by recording activity records and developing new measures and negative association rules. Those frequent activity patterns can be identified leading to positive or negative impact. 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[14] 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 patterns, relational casual patterns and probabilistic causal patterns can be analyzed through considering aspects such as activity forming, contiguity and interaction between activities, and spatial-temporal features of activities and debt-related activity sequences.

Impact-centric analysis seeks to discover activity patterns which are likely to change the expected business outcome. For instance, activity a_{14} is likely to lead to no-debt, but it is likely to result in debt if it is followed by a_4. That is, the activity a_4 is of high impact on outcome when a_{14} happens first.

Customer-centric activity analysis mainly investigates user decision-making behavior and profiling as well as the impact of a users or a class of users actions on related stakeholders. This identifies officer/customers 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 a logistic regression tree and a temporal decision tree can be studied by 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.

4.3. *Activity mining evaluation*

It is essential to specify proper mechanisms[16] for evaluating the workable capability of identified activity patterns and risk models. Technically, we implement *impact-centric mining* which develops interestingness measures in terms of particular activity mining methods. The negative impact prevention 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 verified and expanded, or new measures may be designed to satisfy activity mining demands. In pilot analysis, the measures of some existing KDD methods are found to be invalid when deployed in activity mining. For instance, *support* may be too low to measure frequency, and lift is sensitive to noises for unbalanced data. 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 of debts and cost-effectiveness. Additionally, measures themselves need to be evaluated in terms of interpretability and actionability.

Further evaluation may be necessary by using significance tests, cross-validation and ensemble from both business and technical perspectives. Under certain conditions, 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 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.

5. A Case Study

We test[3,5] some of the above approaches on social security debt-related activity data. The data involves three data sources, which are *activity data* recording activity details, *debt data* logging debt details and *customer data* recording customer circumstances. To analyze the relationship between activity and debt, data from activity files and debt files are extracted. The activity data for us to test the proposed approaches is activity data from 1st January to 31st March 2006. The extracted activity data contain 1.5 million activity records relating to around half million customers.

Preprocessing: First, the activity data is preprocessed to improve data quality and construct activity sequences. There are some activities which are scheduled routinely, and they are removed from activity sequences. Some system activities occur many times during a very short period, and they are combined according to the suggestions of domain experts.

Activity sequence construction: To construct activity sequences, for each customer, the activities which happened within one month immediately before a debt occurrence are built into a debt-related activity sequence. For those customers having no-debts in the first three months of 2006, their activities from 16th January 2006 to 15th February 2006 are built into non debt-related activity sequences.

Activity-centric analysis: The activity sequences are put in two separately datasets: one for debt-related activity sequences, and the other for non debt-related activity sequences. The frequent activity sequence patterns are discovered separately on the above two datasets, and they are then combined to generate contrasting sequence patterns, such as $P \rightarrow T$ is of high support but $P \rightarrow \overline{T}$ is of low support, or vice versa.

Impact-centric analysis: Based on the patterns discovered above, the impact-centric activity patterns are then mined. An impact-centric pattern is of the following form: $P \rightarrow T$ and $P, Q \rightarrow \overline{T}$, which means: "although P is highly associated with debt, the appearance of Q after P will decrease the probability of debt occurrence". Similarly, impact-centric patterns like $P \rightarrow \overline{T}$ and $P, Q \rightarrow T$ are also discovered.

Customer-centric analysis: The demographic patterns of customers with/without debts are mined with a decision tree. Those demographic patterns are then combined with activity-centric patterns and impact-centric patterns to build comprehensive patterns of debtors/non-debtors by taking into consideration both demographic characteristics and activity patterns. The resulting patterns will tell stories for questions like: (1) what kind of customers with what activity sequence are likely/unlikely to have debts, and (2) when there is an activity pattern for a certain group of customers, the occurrence of a specific activity will significantly increase/decrease the probability of raising debts.

Evaluation: Finally, the discovered patterns are evaluated by business experts based on domain knowledge and business measures.

Table 3 illustrates an excerpt of impact-centric sequential activity patterns. In the table, the labels of activities are not real codes due to privacy reason. Each row of the table consists of a contrast pattern pair. One is *underlying pattern* taking the form of "*Underlyingsequence* $\rightarrow$ *Impact 1*," and the other is *reverse pattern* in the form of "*Underlyingsequence* + *Derivativeactivity* $\rightarrow$ *Impact 2*," where *Impact 1*

Table 3. Impact-targeted activity patterns.

Activity Pattern		Explanation
Activity-centric analysis	Positive associations/ sequences	Activity associations/sequences P related to impact $T : P \rightarrow T$
	Negative associations/ sequences	P related to non-impact $\overline{T} : P \rightarrow \overline{T}$
	Contrast associations/ sequences	P related to impact $T : P \rightarrow T$ in impact data set; P also associated with non-impact $\overline{T} : P \rightarrow \overline{T}$ in non-impact data set
Impact-centric analysis	Reverse associations/ sequences	P related to impact $T : P \rightarrow T$ in impact data set, while P, Q associated with non-impact $\overline{T} : P, Q \rightarrow \overline{T}$ in non-impact data set
Customer-centric analysis	Demographic pattern + activity pattern	$D + P \rightarrow T/\overline{T}$, where D is a demographic pattern
	Demographic pattern + impact pattern	$D + P \rightarrow T$ but $D + PQ \rightarrow \overline{T}$; or $D + P \rightarrow \overline{T}$ but $D + PQ \rightarrow T$

Table 4. Impact-targeted activity pattern example.

Underlying Activity	Impact 1	Derivative Activity	Impact 2	SUPP1	SUPP2
a_{14}	no-debt	a_4	debt	0.684	0.428
a_{16}	no-debt	a_4	debt	0.597	0.147
a_{16}	no-debt	a_7	debt	0.597	0.190
a_{14}	no-debt	a_7	debt	0.684	0.202

is opposite to *Impact 2*. SUPP1 is the local support of underlying patterns. For example, the first row shows that the local support of "$a_{14} \rightarrow$ *nodebt*" is 0.684. SUPP2 is the local support of reverse patterns. For the first row, it means that the local support of "$a_{14}, a_4 \rightarrow$ *debt*" is 0.428. The four pattern pairs show that both a_4 and a_7 have a high impact on debt when a_{14} or a_{16} happens first. This real-world activity mining in social security areas has identified interesting results for the government and governmental agencies to take actions to their advantage. More results and comprehensive activity mining in social security areas are available in Refs. 3 and 5

6. Conclusions and Future Work

Impact-related activity data present special structure complexities such as unbalanced class and item distribution. Mining rare but significant positive/negative impact-targeted activity patterns in unbalanced data is very challenging, but may lead to significant actions undertaken in decision-making. 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. The findings can support and enhance business actions and decision-making. Techniques coming from impact-targeted activity mining can prevent disastrous events or improve business decision making and processes.

Issues and tasks listed in Sec. 4 indicate the challenges and directions of this new data mining topic. With a large linkage project[a] support, we are developing new and practical activity mining techniques by taking social security issues as examples.

Acknowledgments

The authors thank Dr. Jie Chen for his technical discussion, as well as to Ms. Yvonne Morrow, Mr. Peter Newbigin and Mr. Rick Schurmann at Centrelink Australia for their domain knowledge and support in the project.

[a] Australian Research Council Linkage grant: Data Mining of Activity Transactions to Strengthen Debt Prevention, LP0775041, 2007–2009.

References

1. W. M. P. Van der Aalst and A. J. M. M. Weijters, Process mining: a research agenda, *Computers in Industry* **53** (2004) 231–244.
2. L. B. Cao and C. Q. Zhang, Domain-driven data mining: a practical methodology, *Int. J. Data Warehousing and Mining* **2**(4) (2006) 49–65.
3. L. B. Cao, Y. C. Zhao, F. Figueiredo, Y. Ou and D. Luo, Impact-targeted activity mining, *PAKDD2007 Industry Track* (2007).
4. L. B. Cao, Domain driven actionable knowledge discovery, *IEEE Intel. Syst.* **22**(4) (2007) 78–89.
5. L. B. Cao, Y. C. Zhao and C. Q. Zhang, Mining impact-targeted activity patterns in imbalanced data, *IEEE Trans. on Knowledge and Data Engineering* (to appear).
6. Centrelink, *Integrated Activity Management Developer Guide* (1999).
7. Centrelink, *Centrelink Annual Report 2004–05.*
8. Z. Chen, From data mining to behavior mining, *Int. J. Information Technology & Decision Making* **5**(4) (2006) 703–712.
9. V. Guralnik and J. Srivastava, Event detection from time series data, *Proc. KDD-99*, 33–42.
10. M. Hammori, J. Herbst and N. Kleiner, Interactive workflow mining requirements, concepts and implementation, *Data & Knowledge Engineering* **56** (2006) 41–63.
11. J. Han, J. Pei and X. Yan, Sequential pattern mining by pattern-growth: principles and extensions, in *Recent Advances in Data Mining and Granular Computing* (Springer Verlag, 2005).
12. J. Mena, *Investigative Data Mining for Security and Criminal Detection*, 1st edn. (Butterworth-Heinemann, 2003).
13. National Research Council, *Making the Nation Safer: The Role of Science and Technology in Countering Terrorism* (Natl Academy Press, 2002).
14. M. Pazzani, A computational theory of learning causal relationships, *Cognitive Science* **15** (1991) 401–424.
15. W. Potts, *Survival Data Mining: Modeling Customer Event Histories* (2006).
16. A. Silberschatz and A. Tuzhilin, What makes patterns interesting in knowledge discovery systems, *IEEE Trans. on Knowledge and Data Engineering* **8**(6) (1996) 970–974.
17. M. Skop, *Survival Analysis and Event History Analysis* (Michal Ŝkop, 2005).
18. G. Williams *et al.*, Temporal event mining of linked medical claims data, *Prod. PAKDD03.*
19. Q. Yang and X. D. Wu, 10 Challenging problems in data mining research, *Int. J. Information Technology & Decision Making* **5**(4) (2006) 597–604.
20. J. Zhang, E. Bloedorn, L. Rosen and D. Venese, Learning rules from highly unbalanced data sets, *Proc. ICDM2004*, 571–574.