Mining Impact-Targeted Activity Patterns in Imbalanced Data

Longbing Cao, *Senior Member*, *IEEE*, Yanchang Zhao, *Member*, *IEEE*, and Chengqi Zhang, *Senior Member*, *IEEE*

Abstract—Impact-targeted activities are rare but they may have a significant impact on the society. For example, isolated terrorism activities may lead to a disastrous event, threatening the national security. Similar issues can also be seen in many other areas. Therefore, it is important to identify such particular activities before they lead to having a significant impact to the world. However, it is challenging to mine impact-targeted activity patterns due to their imbalanced structure. This paper develops techniques for discovering such activity patterns. First, the complexities of mining imbalanced impact-targeted activities are analyzed. We then discuss strategies for constructing impact-targeted activity sequences. Algorithms are developed to mine frequent positive-impact-oriented $(P \rightarrow T)$ and negative-impact-oriented $(P \rightarrow \bar{T})$ activity patterns, sequential impact-contrasted activity patterns (P is frequently associated with both patterns $P \rightarrow T$ and $P \rightarrow \bar{T}$ in separated data sets), and sequential impact-reversed activity patterns (both $P \rightarrow T$ and $PQ \rightarrow \bar{T}$ are frequent). Activity impact modeling is also studied to quantify the pattern impact on business outcomes. Social security debt-related activity data is used to test the proposed approaches. The outcomes show that they are promising for information and security informatics (ISI) applications to identify impact-targeted activity patterns in imbalanced data.

Index Terms—Clustering, classification, association rules, data mining.

1 Introduction

In the emerging research on *information and security informatics* (ISI) [25], [26], [9], [10], [13], activity [5], [39] and event analysis [16], [11], [27], [30], [35], [24] have been the key research objects. *Impact-targeted activities* specifically refer to those activities associated with or leading to a specific impact of interest to the business world. The impact can be an event, a disaster, a government-customer debt, or any other interesting entities. For instance, a series of dispersed and isolated terrorism activities may finally result in a disastrous event [21], [23], [27]. In the social security network [6], [7], [39], [5], [40], a large volume of isolated fraudulent and criminal customer activities can result in a large amount of government-customer debt. For example, in the 5.4-billion government-customer activity transactions per year in Australia, the government social security agency Centrelink accumulates around one billion of customer debt from the delivery of a total of 64 billion payments to 6.5 million eligible customers in the financial year 2004-2005 [7]. Similar problems can be widely seen from other emerging areas such as distributed criminal activities, well-organized separated activities or events threatening the national security and homeland security, and self-organized computer network crimes [9], [12], [28]. Activities in traditional fields such as taxation, insurance services, telecommunication network malfunction, drug disease associations, customer contact centers, and healthcare services may also result in an impact on related organizations or business objectives.

Therefore, it is important to specifically analyze such impact-targeted activities to find out knowledge about what activity patterns are associated with certain types of the impact of interest to specific domain targets and what activity patterns are more likely to lead to the targeted impact. As a result, the findings may support related decision making by providing deep knowledge about the dynamics of impact-targeted activities, the causes of activities leading to certain types of impact, and possible solutions for preventing and minimizing the impact of activities on the society or business outcomes. For instance, in analyzing activities in the social security network, we identify those activities or activity sequences that are more likely to lead to government-customer debt. The resulting evidence and predictors can thus inform relevant officers of the risk of certain ongoing actions or activity sequences resulting in debt. As a result, a potential occurrence of debt can be prevented or minimized. Business decision making and processes, as well as governmental service and policy objectives, can thus be improved and enhanced.

However, impact-targeted activities present some special complexities, which cannot be well handled by existing information processing technologies, for instance, traditional event detection, event and process mining [11], [31], [16], and sequence analysis [18] in both ISI and data mining areas [19]. This is due to the following characteristics of impact-targeted activities. First, impact-targeted activities specifically focus on those activities that have resulted or will result in an impact on business situations. This is normally not concerned with traditional data mining such as sequence or event mining. Second, impact-targeted activities consist of only a very small fraction of the whole activity population. They are normally rare and dispersed in a large activity and customer populations. Nevertheless, it is them that lead to significant effects or even disasters to the society or related business. For instance, only around 4 percent of

• *The authors are with the Department of Software Engineering, University of Technology, Sydney, PO Box 123 Broadway, New Sotuh Wales, Australia 2007. E-mail: {lbcao, yczhao, chengqi}@it.uts.edu.au.*

Manuscript received 8 May 2006; revised 11 June 2007; accepted 19 June 2007; published online 12 July 2007.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDESI-0238-0506. Digital Object Identifier no. 10.1109/TKDE.2007.190635.

social security activities and 38 percent of entitled Australians are associated with the one billion dollars of government-customer debt in the Centrelink service delivery per year [7]. This makes impact-targeted activities deeply enclosed in imbalanced data [8], [15], [20], [38].

The imbalanced data structure differentiates impact-targeted activity data from traditional data. Although there are some methods previously proposed on mining imbalanced data [8], [15], [20], [38] and emerging patterns [14], [37], they are not applicable to the impact-targeted activity pattern mining, which discovers not only sequential patterns on imbalanced data but also impact-reversed patterns. Moreover, the impact-targeted activities are usually distributed across multiple data sources, for instance, debt data, customer circumstance data, and income data in the Centrelink case [6]. Therefore, it is necessary to mine impact-targeted activities by linking relevant data sources. This requires special techniques for constructing activity sequences that are useful for further pattern and impact analysis. The above characteristics distinguish impact-targeted activity data from traditional event data and therefore create a big challenge, that is, to mine impact-targeted activity patterns in imbalanced activity data.

This paper aims to develop effective methodologies, algorithms, and interestingness measures for discovering frequent and sequential impact-targeted activity patterns in the above imbalanced data. The main contributions are given as follows: First, a practical strategy by involving domain knowledge is proposed for constructing impact-targeted activity sequences. We then present effective algorithms for mining both positive (say, $P \rightarrow T$, where T refers to the targeted impact) and negative (for example, $P \rightarrow \bar{T}$, where $\bar{T}$ refers to the nontarget impact) frequent *impact-oriented activity patterns* in imbalanced data. In particular, we study novel algorithms for mining two types of interesting sequential activity patterns, called *impact-contrasted activity patterns* and *impact-reversed activity patterns*. Impact-contrasted activity patterns discover the significant difference resulted in by contrast patterns in two separated classes. For impact-reversed activity patterns, the addition of a derivative activity sequence on top of an underlying frequent impact-oriented activity pattern may lead to the reversal of impact. For instance, it is found that frequent activity pattern P leads to a target T: $P \rightarrow T$, whereas after adding another activity or activity sequence Q on top of P, the derivative pattern PQ is also a frequent one but more likely results in a nontarget $\bar{T}$: $PQ \rightarrow \bar{T}$. Impact-reversed activity pattern mining has the potential of finding out those significant activities that may trigger the reversal of activity impact.

In practice, *impact-oriented*, *impact-contrasted*, and *impact-reversed* activity patterns are all of interest to business people. For instance, in the social security network analysis, impact-oriented activity pattern mining can dig out those government-customer contacts that are likely associated with government-customer debt, whereas impact-reversed activity patterns may be used to detect those key actions that can prevent debt or optimize business processes if they are combined with those activity sequences at a high risk of leading to debt. Finally, we also develop business interestingness measures for quantifying the impact of activities on debt risk from developing business interestingness perspective [3].

We have deployed our approaches through analyzing activity patterns leading to debt and nondebt in the Australian debt-related social security activity data [5], [39], [40]. The activity records of Centrelink officers and customers are combined with the activity operator's circumstance information and government-customer debt outcomes to construct activity-impact sequences for mining those action types, conducted by either officers or customers, that are more likely associated with debt or nondebt. For evaluating the actionability [3] of debt-targeted activity pattern mining, we develop novel technical interestingness and business interestingness metrics to measure both the technical and business performance [3] of identified activity patterns leading to debt. The outcomes of this development are of interest to Centrelink for them to deeply understand and optimize government-customer contacts, prevent fraudulent and criminal activities leading to debt, improve the governmental social security service process, and strengthen the governmental policy objectives and payment security services. In this paper, we generalize the developed approaches so that the findings are of general interest to ISI and KDD and are also free of privacy and political embarrassment that may be caused in disclosing the Centrelink activity data. In fact, the proposed approaches can also be used for analyzing activity or event patterns in other impact-targeted areas such as counterterrorism, crimes, and other social network analyses.

The notations used throughout this paper are given in Table 1.

The remainder of this paper is organized as follows: Section 2 presents an example of activity data and its characteristics, as well as the challenges of activity data to existing KDD- and ISI-related methodologies and techniques. Section 3 discusses activity preprocessing. In Section 4, impact-oriented, impact-contrasted, and impact-reversed activity pattern mining are developed. Experiments and performance evaluation are demonstrated in Section 5. Section 6 introduces related work and further research issues. We conclude this paper in Section 7.

2 Imbalanced Impact-Targeted Activity Data

Impact-targeted activity data is widely seen in many areas such as counterterrorism. In the social security network, everyday government-customer contacts accumulate large quantities of activities. Some of these activities are found to be related to or trigger government-customer debt. However, statistical analysis has shown that debt-related activities are just a very small portion of the whole activity population. This section takes social security debt-associated activity data as an example and discusses its imbalanced characteristics, which differentiate activity data from traditional data.

2.1 An Example—Debt-Related Activity Data in Social Security Network

In the social security network, all entitled customers contact governmental departments or agencies to get government benefits. During the interaction, every single contact, for example, a circumstance change, may trigger a sequence of activities running serially or in parallel. As a result, frequent government-customer contacts generate huge quantities of activity transactions. For instance, the Australian social

TABLE 1
Notations

Notations	Notes
a_i	An activity
P, Q	Activity series or sequences
T	A target, say, a crime, a terrorism event, a fraud or a debt
$\bar{T}$	A non-target, say, non-debt
D	The whole dataset, $D = D_T \bigcup D_{\bar{T}}$
D_T	The target dataset, i.e., a subset of D that are associated with targets
$D_{\bar{T}}$	The non-target dataset, i.e., a subset of D that are not associated with targets
$\|D\|$	The size of the dataset D
$Cnt_D(P)$	The count of P in dataset D
$Supp_D(P)$	The support of P in dataset D
FP_T	Frequent patterns in the target dataset D_T
$FP_{\bar{T}}$	Frequent patterns in the non-target dataset $D_{\bar{T}}$
ICP_T	Impact-contrasted patterns in target data set
$ICP_{\bar{T}}$	Impact-contrasted patterns in non-target set
IRP_T	Impact-reversed patterns in target set
$IRP_{\bar{T}}$	Impact-reversed patterns in non-target set
$Cd_{T,\bar{T}}(P)$	The class difference of pattern P in datasets D_T and $D_{\bar{T}}$
$Cdr_{T,\bar{T}}(P)$	The class difference rate of P in datasets D_T and $D_{\bar{T}}$

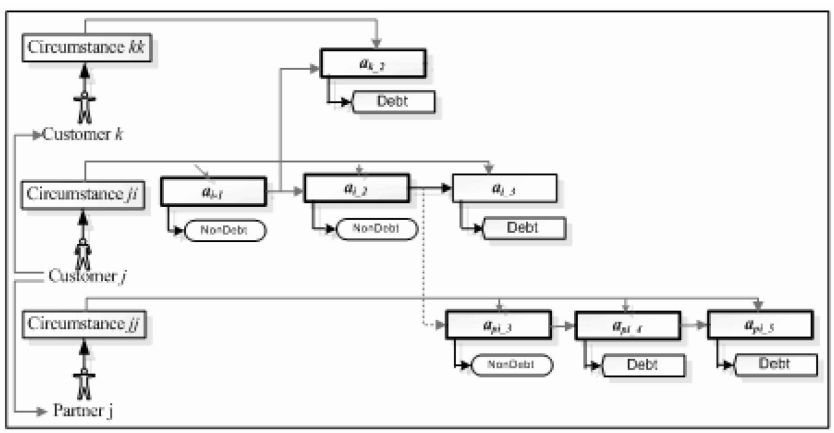


Fig. 1. Activity scenario diagram.

security governmental agency Centrelink accumulates a total of 5.4 billion activity transactions in the financial year 2004-2005 [7]. This data involves 6.5 million, namely, one-third, of Australians, 2.8 million new claims, 1.5 million debts (excluding debts of Family Tax Benefits and Child Care Allowances), 6,600 home visit reviews, 33 million phone calls, and 40 million Internet accesses. On an average day, there are 14 million electronic customer records and 12 million electronic customer transactions. These transactions consist of information about actions and follow-ups taken by either officers or customers, as well as indicators of activities associated with government debt. In fact, we can identify, extract, and generalize many activity scenarios on top of the above activity transactions. For example, Fig. 1 illustrates an activity scenario diagram consisting of partially related customers and their actions and an associated debt occurrence triggered by the scenario of customer circumstance change.

However, statistical analysis has disclosed that debt-associated activities only consist of a very small proportion of the whole activity observations. Therefore, it is necessary to scrutinize the characteristics of impact-targeted activity data. The findings may greatly inform further impact-targeted activity pattern mining.

This example can actually be expanded to many other areas, in particular, ISI fields. For instance, in the terrorism activity network, terrorist members coordinate with their respective leaders and partners, each of them following specified preparation or terrorism action procedures or tasks. Among these procedures or tasks, some of them may generate an impact to the society immediately. Some others accumulate and finally lead to a disastrous event. Similar things may happen in the financial criminal network, for example, exchange market manipulation. Many separated

TABLE 2
Top-Three Frequent Two-Item Patterns for Nondebt-Oriented Activity Sequences

Activity (P)	Consequent	$Supp_D(P)$	$Supp_D(\bar{T})$	$Supp_D(P \rightarrow \bar{T})$	Confidence	Lift
a_6	$\bar{T}$	0.26127	0.96364	0.24793	0.94894	0.98475
a_{13}	$\bar{T}$	0.20445	0.96364	0.19572	0.9573	0.99342
a_5	$\bar{T}$	0.18543	0.96364	0.17155	0.92515	0.96006

TABLE 3
Top-Three Frequent Two-Item Patterns for Debt-Oriented Activity Sequences

Activity (P)	Consequent	$Supp_D(P)$	$Supp_D(T)$	$Supp_D(P \rightarrow T)$	Confidence	Lift
a_4	T	0.14903	0.03636	0.01623	0.1089	2.99505
a_1	T	0.06259	0.03636	0.01469	0.2347	6.4549
a_5	T	0.18543	0.03636	0.01388	0.07485	2.05858

TABLE 4
Frequent Three-Item Activity Patterns in Imbalanced Activity Data

Patterns (P)	Consequent	$Supp_D(P)$	$Supp_D(T)$	$Supp_D(P \rightarrow T)$	Confidence	Lift
a_4, a_1	T	0.01755	0.03636	0.01232	0.70199	19.3066
a_6, a_6	T	0.11442	0.03636	0.0061	0.05331	1.46617
a_5, a_1	T	0.0135	0.03636	0.00562	0.4163	11.44939
a_4, a_4	T	0.06486	0.03636	0.00527	0.08125	2.2346
a_5, a_5	T	0.06817	0.03636	0.00507	0.07437	2.04538

traders from different regions take associated actions in the market to manipulate the price of an instrument. As a result, the price goes up or down dramatically as they hope.

2.2 Complexities in Mining Impact-Targeted Activities

Impact-targeted activities specifically refer to those that are associated with particular business outcomes. For instance, as shown in Fig. 1, a series of government-customer contacts $\{a_{pi\_3}, a_{pi\_4}\}$ lead to government-customer debt. In practice, impact-targeted activities only consist of a very small fraction of the whole activity records. We call impact-targeted activities *target data set* or *target activity set*, whereas the remainder is the *nontarget data set*. As a result, the impact-targeted activity data presents an *imbalanced class distribution*. The imbalanced class distribution of the impact-targeted activity data distresses existing data mining approaches. For instance, the *support* and *confidence* of target activities are much lower than nontarget activities when mining the whole activity data.

Furthermore, the imbalance of impact-targeted activities is also embodied in the aspect of an *imbalanced item set distribution*. Some activities are rarely repeated (called *unrepeated activities* if they only occur once or at a very low frequency), whereas others are frequently repeated such as activities a_{14}, a_{15}, and a_{16} in the Centrelink activity data [6], [5]. In business situations, both rarely repeated and heavily repeated activity types may be associated with a targeted impact. For instance, a series of activities may be conducted by officers or customers prior to the occurrence or confirmation of raising a debt. In addition, the frequency of activities goes up and down in terms of different time frames. As a result, some activity sequences are much longer than others. These scenarios lead to an *imbalanced item set distribution*. An imbalanced item set distribution may hide the occurrence of those rare but significant item sets if no strategies are taken to deal with the imbalance of item sets.

In a subset of Centrelink activity sequences, there are 16,540 (3.6 percent) activity sequences related to debts, whereas 438,394 (96.4 percent) sequences are associated with nondebt. The sizes of the two classes are highly skewed. Tables 2 and 3 show the top-three frequent two-item patterns for the nondebt-oriented activity data set (with $\bar{T}$ being the consequent) and the debt-oriented set (with T being the consequent), respectively. The frequency of the most frequent patterns in the debt-oriented activity sequences is only 0.016, which is much lower than that of the nondebt-oriented activity patterns. Table 4 shows the outcomes of the activity-debt association pattern mining in the above imbalanced activity data.

Such imbalanced data makes it difficult to identify useful rules due to the following reasons: First, for those debt-oriented activity patterns, the support is too low. Among all 249 three-item patterns with a support that is $\geq$ 0.01, only six (2.4 percent) activity patterns likely lead to debt (for example, "$a_4, a_1 \rightarrow T$"). In other k-item $(k > 3)$ activity pattern mining, we have not found any debt-oriented activity patterns with a support greater than 0.005. On the other hand, in some cases, although debt-oriented activity patterns may be found if the support threshold is set to be very low, the number of candidate patterns and the required space explode dramatically, which makes the mining impracticable. In addition, the overwhelming number of nondebt-oriented activity sequences identified in the imbalanced data makes the

interestingness measure *lift* very subtly to noise. A little noise can bring up a big change of the lift of the debt-oriented activity patterns. For example, the rule "$a_4, a_1 \rightarrow T$" in Table 4 has a lift of 19. However, it is not convincing that the rule can lift the confidence by 19 times. Second, for those nondebt-oriented patterns whose support is high, the high confidence actually is meaningless. For example, for the rule "$a_{13} \rightarrow \bar{T}$" in Table 2, with $support = 0.19572$ and $confidence = 0.9573$, its *lift* is only 0.99342, which shows that there is no lift. Since the right support (that is, the support of "$\bar{T}$") is 0.96, the lift of nondebt-oriented patterns cannot be greater than $1/0.96 = 1.04$. Therefore, it is difficult to find interesting nondebt-oriented activity patterns.

The above analysis indicates that an imbalanced class distribution and an imbalanced item set distribution greatly block the emergence of rare but significant impact-targeted patterns. Therefore, it is very essential to develop effective impact-targeted activity pattern mining methods to deal with the imbalanced activity data.

The above complexities of activity data differentiate it from traditional data sets, where data is more flat and simple, and the analysis either does not target business impact or is not in imbalanced data. Due to the imbalance essence of the impact-targeted activity data, mining impact-targeted activities are similar to find dangerous groups of needles hidden in stacks of needle pieces in a haystack. The complexities in impact-targeted activity data determine that it cannot be handled well by existing KDD methods such as process/workflow mining approaches [39], [17], [31]. This is because existing KDD methods mainly study process modeling and have nothing to do with complex activity structures and business impacts of such activities. Therefore, new methodologies, techniques, and algorithms need to be developed.

3 Activity Preprocessing

Besides common issues in data preprocessing, the impact-targeted activity data also presents some special features that need to be handled for further pattern discovery. In particular, this paper discusses strategies for improving activity quality and dealing with an imbalanced class and item set distribution. The aforementioned social security data is used for illustrating our proposed approaches.

3.1 Improving Activity Quality

The social security activity data records transactions at any time point triggered by either social security officers or customers through face-to-face contacts or technology-based interaction. Our initial exploration of the Centrelink activity data [39], [40], [5], which includes activity files, debt management files, customer circumstances files, and customer earnings files, has disclosed some major data quality issues. They consist of outlier and exceptional data, missing values, wrongly coded data, and irrelevant and redundant information. It is essential to improve the activity quality and prepare them for activity pattern mining.

To handle the above activity quality issues, we develop a series of corresponding strategies. For example, neighborhood averaging values may be used to fill in missing values. In particular, a domain-driven data mining methodology [3] is used in data quality improvement. We involve domain knowledge and constraints for the justification and final decision making of when in reality the data quality issues are true positive, true negative, false positive, or false negative. For instance, domain experts are involved to pick up those data that are irrelevant to a specific data mining goal.

3.2 Constructing Activity Sequences

As stated above, activity transactions present structural imbalance. To deal with an imbalanced class and item set distribution of activities, it is necessary to aggregate and construct or rewrite activities into activity sequences that are suitable for pattern mining. Due to the complexities of activity data, many factors must be considered in aggregating and constructing activity sequences. Multiple data sources must be involved in generating impact-targeted activity sequences. Domain knowledge and business-oriented constraints [3] play an important role in setting sliding time windows for selecting and merging activity records. Data reorganization techniques such as data partition are helpful in specifying activity sequences. Taking the social security activity sequence construction as an example, the following approach explains how we can construct both positive and negative impact-targeted activity baskets.

Impact-targeted activities present an *imbalanced class distribution*. As shown in Table 3, an imbalanced class distribution may greatly depress the *support* and *confidence* of debt-targeted activity patterns. In this case, if the impact-targeted activity patterns are mined in the whole data set, very few impact-targeted activity patterns may emerge as frequent patterns. To deal with this issue, the imbalanced activity data needs to be partitioned into suitable data sets so that impact-targeted activity patterns can easily emerge from the overwhelming nonimpact pattern set. For instance, all activities related to debt can be extracted into the target activity set. However, the extraction involves strategies like designing a sliding time window to distinguish which activities should be included into the target data set.

An imbalanced item set distribution can easily push those item sets with high frequency into a frequent impact-targeted activity pattern set. However, some of such item sets may not be of interest to business [3]. Therefore, in the activity sequence construction, some high-frequency activities may need to be aggregated. However, some activities occur very frequently, but they cannot be pruned due to their tight relationship with a targeted impact. For instance, in the Centrelink activity data, activity type a_{14} (reassess benefit) cannot be deleted from the activity value list because it has a direct and close relationship with the risk of debt. In this case, strategies for item set pruning must be discussed with the involvement of domain knowledge and domain experts [3].

In the event history analysis [11], [16], [30], for instance, drug disease association mining in health data [35], once an event happens for a person, one can remove the person's subsequent records (no longer at risk of such an event again). However, in the social security activity data, a debtor may lead to debt again. This indicates that social security debt-targeted activities are different from normal instances of event data.

To construct impact-targeted activity sequences, we set up sliding time windows to select activities happening immediately before the occurrence of a targeted impact and

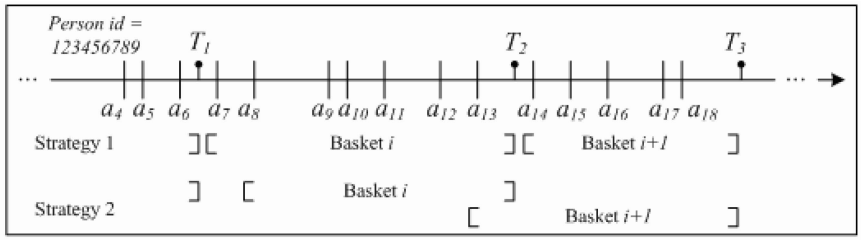


Fig. 2. Activity sequence construction.

pack them into one basket. The time window moves from the left-hand to the right-hand sides of the activity coordinate in terms of each target occurrence. This strategy is based on an assumption that activities occurred long time ago before a target is loosely associated with the target. However, different sliding strategies of time frame may be applicable for constructing activity sequences in varied application situations. In most of event analysis applications, the focus of a sliding time window is set on the first target event for each person [35], [11]. As a result, there is only one event sequence for each person. In the social security government-customer contacts, a debtor may raise one or multiple debts, and all debts are important to the government. Therefore, a basket should be corresponding to one of the debts held by a person. As a consequence, for each debt, those activities within the time window immediately before the occurrence of the debt are put in a basket. Further, the time window is moved forward to reflect the second debt for building another basket.

Furthermore, it is a strategic issue to determine the size of the sliding time window. Domain knowledge, descriptive statistics, and domain experts are most helpful for defining the window size. For instance, Fig. 2 shows two strategies, where $a_i (i = 1, \ldots, m)$ denotes an activity, and $T_j (j = 1, \ldots, n)$ is a debt that is closely associated with a series of activities. In strategy 1, all activities between the occurrences of two debts are packed into one basket. This strategy is more suitable for those applications with targeted impact happening frequently. For instance, for target T_3, its basket includes activity sequence $(a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, T_3)$. Strategy 2 fixes the length of the sliding time window and packs all activities in this window exactly before the occurrence of a debt into one basket. We use strategy 2 to select the activity sequences based on the domain problem characteristics and discussion with domain experts. With this strategy, an activity, say, activity a_{13}, may appear into two baskets, whereas some activities such as a_7 do not appear in any baskets. This is taken as acceptable to domain experts probably because an activity may be associated with two or more debts. Statistical analysis also shows that the duplicate rate of activities crossing two baskets is less than 0.5 percent, so it is not necessary to remove those duplicated activities. For the scenarios listed in Fig. 2, we can build the following positive impact-targeted activity sequences based on the above methods and sliding window strategy 2:

$$(a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, T_2), (a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, T_3).$$

For strategy 2, different methods may be used to quantify the sliding window size. One of the methods is to recode the time interval of two consecutive debts into a series of legal values. For instance, for the Centrelink debt-related activity data, all time intervals are classified into one fortnight, two fortnights, three fortnights, and so on. Statistical analysis is then used to find the more frequent time interval of two consecutive debts and the averaged debt period. An alternative method is to use the average interval between two consecutive targets (say, debt) as the time frame size.

The above methods are developed for building up positive impact-targeted baskets. On the other hand, negative impact-targeted baskets are useful for contrast analysis, which can help in the understanding of which patterns are really related or lead to the targeted events. The problem here is how we can build negative impact-targeted baskets for "normal" people leading to no impact. For the social security debt-related activity data, the idea is to build up activity baskets and sequences for those customers who have never had a debt, with the involvement of the domain expert's suggestions. Similar strategies as in constructing a positive basket can be used for defining the time window. For each nondebt basket or sequence, a virtual nontarget activity $\bar{T}$ is inserted, which is used for mining negative frequent or sequential activity patterns.

4 Activity Pattern Mining

One of the key business concerns in the activity pattern analysis is finding out which particular activity or sequence of activities directly triggers or is closely associated with the occurrence of a targeted impact. To this end, we introduce approaches to identify three types of *frequent* activity patterns: 1) *impact-oriented activity patterns* leading to either the occurrence or the disappearance of the targeted impact in the form of $P \rightarrow T$ or $P \rightarrow \bar{T}$, 2) *impact-contrasted activity patterns*, in which the same activity or activity sequence leads to both positive impact $P \rightarrow T$ in the target data set and negative impact $P \rightarrow \bar{T}$ in the nontarget data set, and 3) *sequential impact-reversed activity patterns*, in which the occurrence of an activity or activity sequence leads to the impact reversal from positive to negative, or vice versa, in separate data sets.

4.1 Mining Frequent Impact-Oriented Activity Patterns

One of the major activity mining tasks is to discover frequent activity patterns. In particular, to support social security decision making, the business aims of mining frequent debt-oriented activity patterns are to find out which activities more likely lead to debt or nondebt. Therefore, both positive and negative frequent activity patterns may be identified, as shown in Table 5. An impact-oriented activity pattern is in the form of $P \rightarrow T$, where the

TABLE 5
Positive and Negative Impact-Oriented Activity Pattern Mining

	Target occurred (T)	Target Non-occurred ($\bar{T}$)
Pattern appearing (P)	$P \rightarrow T$	$P \rightarrow \bar{T}$
Pattern not occurring ($\bar{P}$)	$\bar{P} \rightarrow T$	$\bar{P} \rightarrow \bar{T}$

left-hand side P is a sequence of activities, and the right hand of the rule is always the target T, which can be a targeted activity, event, or other types of business impact. In the social security activity analysis, the target impact refers to leading to either debt (called *positive impact*) or nondebt (called *negative impact*).

Here, positive frequent impact-oriented activity patterns ($P \rightarrow T$, or $\bar{P} \rightarrow T$) mean that those patterns more likely lead to the occurrence of the targeted impact, say, leading to a debt, resulting from either an appeared pattern P or a disappeared pattern $\bar{P}$. On the other hand, negative frequent impact-oriented activity patterns ($P \rightarrow \bar{T}$, or $\bar{P} \rightarrow \bar{T}$) indicate that the target will not happen $\bar{T}$, say, leading to no debt.

To find useful patterns in the activity data, new interestingness metrics need to be defined to measure the frequent impact-oriented activity pattern mining. In the imbalanced target activity set, a frequent impact-oriented activity sequence leading to debt $P \rightarrow T$ if P satisfies the following conditions: 1) P is frequent in the whole data set, 2) P is far more frequent in the target data set than in the nontarget data set, and 3) P is far more frequent in the target data set than in the whole data set. To this end, we define the following interestingness measures. In the impact-oriented activity pattern mining, given an activity data set $D = D_T \bigcup D_{\bar{T}}$, a subset D_T of D consists of all activities and activity sequences that are associated with the targeted impact, whereas the subset $D_{\bar{T}}$ includes all activity sequences related to the nonoccurrence of the targeted impact. For instance, in the social security network, activity sequence pattern $P(P = (a_i, a_{i+1}, \ldots), a_i \in D, i = 0, 1, \ldots)$ is an activity pattern associated with debt $P \rightarrow T$. The count of debts (that is, the count of sequences enclosing P) resulting from P in D is $Cnt_D(P)$.

Definition 1. *The* global support *of a pattern $P \rightarrow T$ in activity set D is defined as*

$$Supp_D(P \rightarrow T) = \frac{Cnt_D(P \rightarrow T)}{|D|},$$

where $|D|$ is size of set D. If $Supp_D(P \rightarrow T)$ is larger than a given threshold, then P is a frequent activity sequence in D leading to a debt. $Supp_D(P \rightarrow T)$ reflects the global statistical significance of the rule $P \rightarrow T$ in activity set D.

Definition 2. *The* local support *of a rule $P \rightarrow T$ in the target activity set D_T is defined as*

$$Supp_{D_T}(P \rightarrow T) = \frac{Cnt_{D_T}(P \rightarrow T)}{|D_T|}.$$

Similarly, the local support of rule $P \rightarrow \bar{T}$ in the nondebt activity set $D_{\bar{T}}$ is defined as

$$Supp_{D_{\bar{T}}}(P \rightarrow \bar{T}) = \frac{Cnt_{D_{\bar{T}}}(P \rightarrow \bar{T})}{|D_{\bar{T}}|}.$$

In real-world data mining, there is a big gap [4], [29] between technical interestingness and business concerns [3]. For instance, based on the traditional measures of *support*, *confidence*, and *lift*, we may find some frequent activity patterns with a relatively high support. However, when they are presented to business analysts, these patterns are judged to be of no interest because they are common sense. Although a few other patterns with lower supports seem more interesting to domain experts, this instance shows that in business, technical measures such as the support for association rule mining do not mean everything. It is necessary to develop measures to cover interestingness concerns from both the technical and business perspectives [3], [4]. For this purpose, here, we discuss *activity impact modeling* by developing business interestingness measures to evaluate the business impact of activity patterns. In the following, *risk* and *cost* are defined to measure the impact of a pattern.

Definition 3. *The risk of a pattern, say, $P \rightarrow T$, is defined as $Risk(P \rightarrow T)$, which is the ratio of the cost $Cost(P \rightarrow T)$ associated with the particular pattern to the total cost of pattern set $TotalCost(P)$ in the data set, that is, $Risk(P \rightarrow T) = \frac{Cost(P \rightarrow T)}{TotalCost(P)}$. The* average cost *of a particular pattern $P \rightarrow T$ is defined as $AvgCost(P \rightarrow T) = \frac{Cost(P \rightarrow T)}{Cnt(P \rightarrow T)}$. Here, the* cost *is the consequence of a target such as the toll or damage of a terrorism event, the area of an epidemic, the loss from a fraud, or the amount/duration of a debt.*

In practice, *risk* and *cost* are instantiated into particular forms in terms of specific domain problems and performance measures of business expectations. For the social security debt analysis, *risk* and *cost* are instantiated into $Risk_{amt}$, $Risk_{dur}$, $AvgAmt$, and $AvgDur$, as discussed in Section 5, which stand for the debt amount risk, debt duration risk, average debt amount, and average debt duration, respectively. The above metrics can serve business performance evaluation to measure the impact of an activity pattern on debt outcome.

Based on the above methods and other existing metrics such as confidence and lift, frequent debt-oriented activity patterns are studied in both target and nontarget activity data. Module 1 of Algorithm 1 in Section 4.4 illustrates the main ideas of the frequent impact-oriented activity mining by using an a priori approach [1].

4.2 Mining Sequential Impact-Contrasted Activity Patterns

In security-related applications such as terrorism or fraud detection, potential outlaws or terrorists consist of only an extremely small part of the whole population. In such a case, the *support* will be too low to capture interesting frequent patterns or association rules. A usual method is to roughly sample the same number of records from both target and nontarget classes. However, some information may be lost after sampling, and the *support*, *confidence*, and *lift* of the discovered association rules or sequential patterns may be different from the traditional meaning, which makes it difficult to understand.

For the above problem, our idea is to mine frequent patterns separately on each class and then discover

interesting patterns by combining the results from two classes. There are two different cases that are interesting to business:

1. $Supp_{D_T}(P \rightarrow T)$ is high, but $Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})$ is low.
2. $Supp_{D_T}(P \rightarrow T)$ is low, but $Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})$ is high.

Note that it is unfair if we compare the *global supports* of the two patterns in the whole data set D because the *support* of T is much lower than that of $\bar{T}$. In each of the above cases, if there is a big contrast between two patterns, say, $Supp_{D_T}(P \rightarrow T)$ is much greater than $Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})$, it indicates that P is more likely associated with T rather than $\bar{T}$.

We use FP_T to denote those frequent item sets discovered in those impact-targeted baskets, whereas $FP_{\bar{T}}$ stands for those frequent item sets discovered in nontarget activity baskets. In most cases, those frequent item sets in FP_T but not in $FP_{\bar{T}}$ are interesting because they tell which activities or activity sequences lead to a crime, terrorism, fraud, or debt. However, in some cases, those frequent items in $FP_{\bar{T}}$ but not in FP_T may help us understand which activity sequences likely prevent the occurrence of a targeted impact. Therefore, we define *impact-contrasted patterns* ICP_T and $ICP_{\bar{T}}$ as

$$ICP_T = FP_T \setminus FP_{\bar{T}},$$
$$ICP_{\bar{T}} = FP_{\bar{T}} \setminus FP_T.$$

$ICP_{\bar{T}}$ presents frequent activity item sets, which are potentially interesting for a nontarget impact. It is useful for analyzing such applications, in which certain types of activities or activity sequences lead to the nonoccurrence of a targeted impact. After mining ICP_T, those people with a low probability of causing a business impact, say, a crime or debt, can be checked if they are identified to be associated with patterns in ICP_T. If this is the case, then they are at a higher risk of leading to the impact occurrence in the future.

To measure the interestingness of frequent impact-contrasted item sets, we define two *lifts* for the frequent item sets in ICP_T, which are different from the traditional *confidence* for association rules. The two *lifts* tell us how much the lift of the patterns is.

Definition 4. *The* class difference *of P in two data sets D_T and $D_{\bar{T}}$ is defined as:*

$$Cd_{T,\bar{T}}(P) = Supp_{D_T}(P \rightarrow T) - Supp_{D_{\bar{T}}}(P \rightarrow \bar{T}).$$

The class difference ratio *of P in D_T and $D_{\bar{T}}$ is defined as:*

$$Cdr_{T,\bar{T}}(P) = \frac{Supp_{D_T}(P \rightarrow T)}{Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})}.$$

Similarly, two new lifts, the class difference $Cd_{\bar{T},T}(P)$ *and the class difference ratio $Cdr_{\bar{T},T}(P)$, are defined for those frequent item sets in $ICT_{\bar{T}}$:*

$$Cd_{\bar{T},T}(P) = Supp_{D_{\bar{T}}}(P \rightarrow \bar{T}) - Supp_{D_T}(P \rightarrow T),$$
$$Cdr_{\bar{T},T}(P) = \frac{Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})}{Supp_{D_T}(P \rightarrow T)}.$$

The above definitions reflect not only the class difference but also the likelihood of the occurrence of impact-targeted patterns against nonimpact-targeted ones, or vice versa. For instance, if $Cd_{T,\bar{T}}(P)$ or $Cdr_{T,\bar{T}}(P)$ is larger than a given threshold, then P more frequently leads to debt than nondebt. This measure indicates the difference between the targeted class and the untargeted class. An obvious difference between them is expected for positive frequent impact-targeted activity patterns.

Definition 5. *Given* global supports $Supp_D(P \rightarrow T)$ *and* $Supp_D(P \rightarrow \bar{T})$, *the* relative risk ratio *of P is defined as* $Rrr_{T,\bar{T}}(P)$:

$$Rrr_{T,\bar{T}}(P) = \frac{Prob(T|P)}{Prob(\bar{T}|P)} = \frac{Prob(P \rightarrow T)}{Prob(P \rightarrow \bar{T})} = \frac{Supp_D(P \rightarrow T)}{Supp_D(P \rightarrow \bar{T})}.$$

If $Rrr_{T,\bar{T}}(P)$ is larger than a given threshold, then P far more frequently leads to debt than nondebt. This measure indicates the statistical difference of an activity or sequence P leading to debt or nondebt in a global manner. An obvious difference between them is expected for positive frequent impact-targeted activity patterns. In addition, if the statistical significance of P leading to T and $\bar{T}$ are compared in terms of local classes, then the *relative risk ratio* $Rrr_{T,\bar{T}}(P)$ indicates the difference of a pattern's significance in the targeted class versus the untargeted class, as defined in Definition 4.

The algorithm for mining impact-contrasted patterns is presented in Module 2 in Section 4.4.

4.3 Mining Sequential Impact-Reversed Activity Patterns

In some business situations, the occurrence of some specific activities or activity sequences may dramatically reverse the impact of an underlying activity series or sequences from positive to negative, or vice versa. Such derivative activities play important roles in leading to the reversal of the resulting impact. For instance, assume an underlying frequent impact-targeted pattern 1 as follows:

1. Pattern 1: $P \rightarrow T$. This means that activity sequence P leads to a target.
   Then, a derivative impact-targeted activity pattern 2 of pattern 1 is also frequent.
2. Pattern 2: $PQ \rightarrow \bar{T}$. This means that activity sequence P leads to a nontarget if P is followed by another activity or activity sequence Q.

In this case, we call pattern 1 the *underlying pattern* and pattern 2 the *derivative pattern*. Patterns 1 and 2 consist of a *contrast pattern pair*. The occurrence of Q directly results in the reversal of the impact of activity sequences. We call such activity patterns as *sequential impact-reversed activity patterns*. Another scenario of the impact-reversed activity pattern mining is the reversal from the negative impact-targeted activity pattern $P \rightarrow \bar{T}$ to the positive impact $PQ \rightarrow T$ after joining with a trigger activity or activity sequence Q.

In the reversal from pattern 1, which leads to debt, to pattern 2, which results in nondebt, the activity or activity sequence Q plays an important role. This phenomenon is of

great interest to business. For instance, it can be used for improving business processes or recommending activity sequences for avoiding activities or government-customer contacts that may lead to or be associated with debts.

Definition 6. *To measure the significance of Q leading to impact reversal from positive to negative, or vice versa, the metric* conditional impact ratio (Cir) *is developed as follows:*

$$\begin{aligned} &Cir(Q\bar{T}|P) \\ &= \frac{Prob(Q\bar{T}|P)}{Prob(Q|P) \times Prob(\bar{T}|P)} \\ &= \frac{Prob(PQ \rightarrow \bar{T})/Prob(P)}{(Prob(PQ)/Prob(P)) \times (Prob(P \rightarrow \bar{T})/Prob(P))} \\ &= \frac{Prob(PQ \rightarrow \bar{T})/Prob(PQ)}{Prob(P \rightarrow \bar{T})/Prob(P)}. \end{aligned}$$

Cir measures the statistical probability of activity sequence Q leading to nondebt, given that pattern P happens in activity set D. It indicates the impact of $Prob(PQ \rightarrow \bar{T})$ and $Prob(P \rightarrow T)$ on the significance of Q contributing to the impact transfer from an aspect to its opposite. If Cir is larger than a given threshold, then it is Q that makes a significant contribution to the change of pattern impact from $P \rightarrow T$ to $PQ \rightarrow \bar{T}$. Basically, the bigger Cir is, the more likely that Q triggers the impact reversal of the underlying pattern.

Definition 7. *Another metric for measuring the difference led by the occurrence of Q in the above scenario is the* conditional Piatetsky-Shapiro's *(P-S; [28])* ratio Cps*, which is defined as follows:*

$$\begin{aligned} Cps(Q\bar{T}|P) &= Prob(Q\bar{T}|P) - Prob(Q|P) \times Prob(\bar{T}|P), \\ &= \frac{Prob(PQ \rightarrow \bar{T})}{Prob(P)} - \frac{Prob(PQ)}{Prob(P)} \times \frac{Prob(P \rightarrow \bar{T})}{Prob(P)}. \end{aligned}$$

From another perspective, Cps measures the statistical or proportional significance of activity sequence Q leading to the impact reversal.

We present the algorithm for mining impact-reversed activity patterns in Module 3 of Algorithm 1 in Section 4.4.

4.4 Impact-Targeted Activity Pattern Mining Algorithm

On the basis of the above approaches, the following algorithm is designed for mining three kinds of impact-targeted activity patterns. The algorithm consists of three modules (see Fig. 3):

1. Module 1. Mining frequent impact-oriented activity patterns in the target and nontarget data sets, respectively, by using an a priori method to find those frequent activity patterns leading to either debt or nondebt.
2. Module 2. Mining frequent impact-contrasted activity patterns in the target and nontarget data sets to find contrast patterns and common frequent patterns:
   a. *Contrast patterns*. Here, two patterns are identified from the target and nontarget data sets, but there exists a significant difference of concerned interestingness measures of the identified two patterns.
   b. *Common frequent patterns*. These are common frequent patterns with the same activity or activity sequence but result in opposite business impacts.
3. Module 3. Mining impact-reversed activity patterns in the target and nontarget data sets to identify those derivative patterns that lead to the impact reversal of the underlying patterns. The idea is explained as follows: For each target (or nontarget) k item set P in FP_T (or $FP_{\bar{T}}$), we check the nontarget (or target) $(k+1)$ item set P^* in $FP_{\bar{T}}$ (or FP_T), where the first k activities in P^* are the same as those in P. Then, we compute Cir and Cps.

5 Experimental Results

The above approaches have been tested on the Centrelink debt-related activity data. The data involves four data sources, which are activity files recording activity details, debt files logging debt details, customer files enclosing customer profiles, and earnings files storing earnings details. To analyze the relationship between activity and debt, the data from the activity files and debt files are extracted. The activity data for us to test the proposed approaches is the Centrelink activity data from 1 January 2006 to 31 March 2006. We extract activity data including 15,932,832 activity records recording government-customer contacts with 495,891 customers, which lead to 30,546 debts in the first three months of 2006.

There are 35 attributes in the activity table, from which four attributes are selected: *person id*, *activity code*, *activity start date*, and *time sequence number*. *Person id* is the unique identifier of Centrelink customers, *activity code* is the type code of an activity, and *activity start date* and *time sequence number* are the start date and time of an activity, respectively. There are 29 attributes in the debt table, from which five fields are selected: *person id*, *debt id*, *debt amount*, *debt start date*, and *debt end date*. Based on *debt start date* and *debt end date*, a new field *debt duration* is generated to stand for how many days a debt period is.

In Australia, most customers are paid fortnightly. Statistical and empirical analysis indicates that debts happen more frequently to those frequent debtors in and continue for two fortnights. Therefore, we set the time interval of a debt-targeted time window as two fortnights. For those customers with debts happening between 1 February 2006 and 31 March 2006, the activity sequences are built by putting all activities in one month immediately before a debt occurrence. For those debts in the first month of 2006, because the activity data that is available for them is less than one month, these debts are ignored. The activities that we use for building nondebt baskets and sequences are those activities from 15 January to 15 February for those customers having no debts in the first three months of 2006. The reason that we select those activities after 16 January is that there are much fewer activities per day in the first half of January than in other periods. For those activities in March, it remains unknown whether they will lead to a debt or not in the following month, so they are not considered in building nondebt baskets and sequences. The date of the virtual nondebt event in a nondebt activity sequence is set to the latest date in the sequence. The time of debt or

ALGORITHM 1: Mining impact-targeted activity patterns
INPUT: target dataset D_T, and non-target dataset $D_{\bar{T}}$, and thresholds T_{sup}, T_{diff}, T_{ratio}, T_{cir}, T_{cps}
OUTPUT: contrast patterns ICP_T and $ICP_{\bar{T}}$, common frequent patterns CFP, impact-reversed activity patterns IRP_T and $IRP_{\bar{T}}$.

/*Module 1: Mining frequent target-oriented activity patterns using Apriori */
$FP_T = Apriori(D_T)$; /* mining frequent activity patterns in target dataset D_T */
$FP_{\bar{T}} = Apriori(D_{\bar{T}})$; /* mining frequent activity patterns in non-target dataset $D_{\bar{T}}$ */
$Supp_T = support(FP_T, D_T)$; /* $Supp_T$ is the local support in target dataset */
$Supp_{\bar{T}} = support(FP_{\bar{T}}, D_{\bar{T}})$; /* $Supp_{\bar{T}}$ is the local support in non-target dataset */

/* Module 2: Mining frequent impact-contrasted activity patterns including contrast patterns and common frequent patterns */
$CFP = \{P | P \in FP_T \bigcap FP_{\bar{T}},\ Supp_T(P) \geq T_{sup},\ Supp_{\bar{T}}(P) \geq T_{sup}\}$;
FOR each activity or activity sequence P frequent both in FP_T and $FP_{\bar{T}}$
  $Cd_{T,\bar{T}}(P) = Supp_T(P) - Supp_{\bar{T}}(P)$;
  $Cdr_{T,\bar{T}}(P) = Supp_T(P)/Supp_{\bar{T}}(P)$;
  $Cd_{\bar{T},T}(P) = Supp_{\bar{T}}(P) - Supp_T(P)$;
  $Cdr_{\bar{T},T}(P) = Supp_{\bar{T}}(P)/Supp_T(P)$;
ENDFOR
$ICP_T = \{P | P \in FP_T,\ Cd_{T,\bar{T}}(P) \geq T_{diff}, Cd_{T,\bar{T}}(P) \geq T_{ratio}\}$;
$ICP_{\bar{T}} = \{P | P \in FP_{\bar{T}},\ Cd_{\bar{T},T}(P) \geq T_{diff}, Cd_{\bar{T},T}(P) \geq T_{ratio}\}$;

/* Module 3: Mining impact-reversed activity patterns for activities or activity sequences leading to the reversal of impact */
FOR $k = 1$ to (the maximum length of itemset - 1)
  FOR each frequent k-itemset P in FP_T
    FOR each frequent $(k+1)$-itemset P^* in $FP_{\bar{T}}$
      IF the first k items in P^* are the same as the k items in P
        $Q = P^* \backslash P$;
        $Cir(Q\bar{T}|P) = \frac{Prob(PQ \rightarrow \bar{T})/Prob(PQ)}{Prob(P \rightarrow \bar{T})/Prob(P)}$;
        $Cps(Q\bar{T}|P) = \frac{Prob(PQ \rightarrow \bar{T})}{Prob(P)} - \frac{Prob(PQ)}{Prob(P)} \times \frac{Prob(P \rightarrow \bar{T})}{Prob(P)}$;
        $IRP_T = IRP_T \bigcup \{(P, Q, Cir, Cps)\}$;
      ENDIF
    ENDFOR
  ENDFOR
ENDFOR
$IRP_T = \{S | S \in IRP_T,\ Cir(S) \geq T_{cir}\ and\ Cps(S) \geq T_{cps}\}$;

/* The code for generating $IRP_{\bar{T}}$ is similar to the above for IRP_T, so it is ignored here to save space. */

Fig. 3. Mining impact-targeted activity patterns in the imbalanced activity data.

nondebt events are set to "23:59:59," which makes them the last events in the sequences.

Based on the above activity construction approach, debt-oriented activity sequences are generated. There are 454,934 sequences: 16,540 (3.6 percent) activity sequences associated with debts and 438,394 (96.4 percent) sequences with nondebts. Labels T and $\bar{T}$ denote the occurrence of debt and nondebt, respectively, and code a_i represents an activity.

Table 6 shows some typical frequent activity patterns discovered from the above imbalanced activity data set. The first three rules $a_1, a_2 \rightarrow T$, $a_3, a_1 \rightarrow T$, and $a_1, a_4 \rightarrow T$ have high *confidence* and *lift* but low *support* (caused by class imbalance). They are interesting to business because their *confidences* and *lifts* are high, and their *supports* and *AvgAmts* are not too low. The third rule $a_1, a_4 \rightarrow T$ is the most interesting because its $risk_{amt}$ is as high as 0.424, which means that this patterns accounts for 42.4 percent of the total amount of debts. The fourth rule $a_1 \rightarrow T$ has a high lift of 6.5, and the appearance of a_2, a_3, or a_4 can triple its *lift*, as shown by the first three rules. The values of $risk_{dur}$ show that rules $a_6 \rightarrow T$, $a_5 \rightarrow T$, and $a_7 \rightarrow T$ are associated with a longer average duration of debts, so they are also interesting to business, although their *confidences* and *lifts* do not necessarily indicate so in the traditional view of association rule mining.

Table 7 presents the contrast sequential patterns discovered in the target and nontarget data sets, respectively. $Supp_{D_T}(P)$ and $Supp_{D_{\bar{T}}}(P)$ denote the *local supports* of a pattern leading to debt in the target data set and nondebt in the nontarget data set, respectively. $Cd_{T,\bar{T}}(P)$ and $Cdr_{T,\bar{T}}(P)$ are, respectively, the difference and ratio between $Supp_{D_T}(P)$ and $Supp_{D_{\bar{T}}}(P)$. $Cd_{\bar{T},T}(P)$ and $Cdr_{\bar{T},T}(P)$ are, respectively, the difference and ratio between $Supp_{D_{\bar{T}}}(P)$ and $Supp_{D_T}(P)$. In the table, pattern "a_{14}, a_{14}, a_4" has a $Cdr_{T,\bar{T}}(P)$ of 4.04, which means that the pattern is three times more likely to lead to debt than nondebt. For the above pattern, $AvgAmt$ is 21,761 cents, and $AvgDur$ is 2.9 days, which shows that the *average debt amount* associated with the pattern is medium and that the *average debt duration* is relatively short. Its $risk_{amt}$ shows that it appears before 41.5 percent of all debts. The above measures suggest that this pattern is interesting to business. According to $AvgAmt$ and $AvgDur$, the debt related to the second pattern a_8 is of both high average amount (26,789 cents) and long duration (9.9 days). Its $Cdr_{T,\bar{T}}(P)$ shows that it is thrice likely associated with debt than nondebt; therefore, the pattern is also interesting. By comparing the first pattern a_4 with the seventh one a_{11}, we

TABLE 6
Frequent Debt-Targeted Activity Patterns Discovered in Imbalanced Activity Set

Patterns $P \rightarrow T$	$Supp_D(P)$	$Supp_D(T)$	$Supp_D(P \rightarrow T)$	Confidence	Lift	$AvgAmt$ (cents)	$AvgDur$ (days)	$risk_{amt}$	$risk_{dur}$
$a_1, a_2 \rightarrow T$	0.0015	0.0364	0.0011	0.7040	19.4	22074	1.7	0.034	0.007
$a_3, a_1 \rightarrow T$	0.0018	0.0364	0.0011	0.6222	17.1	22872	1.8	0.037	0.008
$a_1, a_4 \rightarrow T$	0.0200	0.0364	0.0125	0.6229	17.1	23784	1.2	0.424	0.058
$a_1 \rightarrow T$	0.0626	0.0364	0.0147	0.2347	6.5	23281	2.0	0.490	0.111
$a_6 \rightarrow T$	0.2613	0.0364	0.0133	0.0511	1.4	18947	7.2	0.362	0.370
$a_4 \rightarrow T$	0.1490	0.0364	0.0162	0.1089	3.0	21749	3.2	0.505	0.203
$a_5 \rightarrow T$	0.1854	0.0364	0.0139	0.0755	2.1	18290	6.2	0.363	0.334
$a_7 \rightarrow T$	0.1605	0.0364	0.0113	0.0706	1.9	19090	6.8	0.310	0.300

TABLE 7
Contrast Sequential Patterns Identified in Target and Nontarget Data Sets

Patterns (P)	$Supp_{D_T}(P)$	$Supp_{D_{\bar{T}}}(P)$	$Cd_{T,\bar{T}}(P)$	$Cdr_{T,\bar{T}}(P)$	$Cd_{\bar{T},T}(P)$	$Cdr_{\bar{T},T}(P)$	$AvgAmt$ (cents)	$AvgDur$ (days)	$risk_{amt}$	$risk_{dur}$
a_4	0.446	0.138	0.309	3.24	-0.309	0.31	21749	3.2	0.505	0.203
a_8	0.176	0.060	0.117	2.97	-0.117	0.34	26789	9.9	0.246	0.245
a_5	0.382	0.178	0.204	2.15	-0.204	0.47	18290	6.2	0.363	0.334
a_7	0.312	0.154	0.157	2.02	-0.157	0.50	19090	6.8	0.310	0.300
a_{11}	0.084	0.055	0.029	1.54	-0.029	0.65	23459	10.1	0.102	0.119
a_6	0.367	0.257	0.110	1.43	-0.110	0.70	18947	7.2	0.362	0.370
a_{14}, a_4	0.428	0.119	0.309	3.60	-0.309	0.28	21872	3.1	0.487	0.184
a_4, a_{14}	0.335	0.107	0.227	3.12	-0.227	0.32	21574	3.5	0.376	0.163
a_4, a_{15}	0.255	0.092	0.163	2.78	-0.163	0.36	21127	3.9	0.280	0.141
a_{14}, a_{14}, a_4	0.367	0.091	0.276	4.04	-0.276	0.25	21761	2.9	0.415	0.151
a_{14}, a_4, a_{15}	0.241	0.077	0.164	3.13	-0.164	0.32	21371	3.7	0.268	0.125

can see that although a_4 is more likely associated with debt, a_{11} is more likely associated with debts with a longer duration. Therefore, both patterns are interesting from different perspectives.

Table 8 is similar to Table 7, but Table 8 demonstrates the common frequent sequential patterns in both target and nontarget data sets, which do not necessarily have a big difference between $Supp_{D_T}(P)$ and $Supp_{D_{\bar{T}}}(P)$. From the business perspective, if a pattern has a high support in both the debt and nondebt data sets, it means that, probably, there are some other activities missing here but having a high impact on the occurrence of debts. Although most activity patterns are more associated with debt, pattern "a_{16}, a_{16}" is more likely associated with nondebt.

Table 9 shows an excerpt of the impact-reversed sequential activity patterns and pairs of contrast patterns. One is *underlying pattern* $P \rightarrow Impact\ 1$, whereas the other is *derivative pattern* $PQ \rightarrow Impact\ 2$, where $Impact\ 1$ is opposite to $Impact\ 2$, and Q is a derivative activity or sequence. $Supp_{D_T}(P \rightarrow T)$ and $Supp_{D_{\bar{T}}}(PQ \rightarrow \bar{T})$ are, respectively, the local supports of the underlying patterns and reverse patterns. For example, the first row shows that the local support of $a_{14} \rightarrow \bar{T}$ in the nondebt data is 0.684, and the local support of $a_{14}, a_4 \rightarrow T$ in the debt data is 0.428. Cir stands for *conditional impact ratio*, which shows the impact of the derived activity on *Impact 2* when underlying pattern happens. Cps denoted *conditional P-S ratio*. Both Cir and Cps show how much the impact is reversed by the derivative activity Q. The first row shows that the appearance of a_4 tends to change the impact from $\bar{T}$ to T when a_{14} happens first. The following analysis will help us better understand the meaning of the pair of patterns $a_{14} \rightarrow \bar{T}$ and $a_{14}, a_4 \rightarrow T$. The local supports of $a_{14} \rightarrow T$ and $a_{14} \rightarrow \bar{T}$ are, respectively, 0.903 and 0.684 (see Table 8), so the ratio of the two values is $0.903/0.684 = 1.3$. The local supports of $a_{14}, a_4 \rightarrow T$ and $a_{14}, a_4 \rightarrow \bar{T}$ are 0.428 and 0.119, respectively (see Table 7), so the ratio of the two values are $0.428/0.119 = 3.6$. The above two ratios indicate that when a_{14} happens first, the appearance of a_4 drives a_{14} likely toward debtable. This kind of pattern pairs helps us know what effect an additional activity will have on the impact of the patterns.

Although the above results are from the debt-related data, and the analysis of the results are focused on the debt-related patterns, in fact, the proposed methods and most of the proposed interestingness measures (say, Cir and Cps) are not confined within the debt-related domain. The $AvgAmt$ and $AvgDur$ are debt-related measures, and they can be taken as the cost or consequence of activities, which can be replaced with any domain-specific measures. For example, they can be taken as the damage or the number of deaths in a terrorism attack, or they can be set as the amount of money in a robbery.

6 Related Work and Discussions

In current ISI research [25], [26], [9], [10], [13], many researchers have made efforts on event-focused analysis, for instance, event detection and history analysis [11], [24], crime analysis [10], [23], [32], and terrorism-oriented

TABLE 8
Common Frequent Sequential Patterns in Separate Data Sets

Patterns (P)	$Supp_{D_T}(P)$	$Supp_{D_{\bar{T}}}(P)$	$Cd_{T,\bar{T}}(P)$	$Cdr_{T,\bar{T}}(P)$	$Cd_{\bar{T},T}(P)$	$Cdr_{\bar{T},T}(P)$	$AvgAmt$ (cents)	$AvgDur$ (days)	$risk_{amt}$	$risk_{dur}$
a_5	0.382	0.178	0.204	2.15	-0.204	0.47	18290	6.2	0.363	0.334
a_7	0.312	0.154	0.157	2.02	-0.157	0.50	19090	6.8	0.310	0.300
a_6	0.367	0.257	0.110	1.43	-0.110	0.70	18947	7.2	0.362	0.370
a_{14}	0.903	0.684	0.219	1.32	-0.219	0.76	19251	6.6	0.905	0.840
a_{15}	0.746	0.567	0.179	1.32	-0.179	0.76	19192	7.4	0.745	0.775
a_{16}	0.604	0.597	0.007	1.01	-0.007	0.99	17434	8.7	0.548	0.738
a_{14}, a_{15}	0.605	0.374	0.231	1.62	-0.231	0.62	19235	7.0	0.606	0.594
a_{15}, a_{15}	0.539	0.373	0.167	1.45	-0.167	0.69	18918	7.7	0.531	0.584
a_{16}, a_{14}	0.479	0.402	0.076	1.19	-0.076	0.84	16726	8.1	0.417	0.549
a_{14}, a_{16}	0.441	0.393	0.049	1.12	-0.049	0.89	17013	8.5	0.391	0.532
a_{16}, a_{16}	0.367	0.410	-0.043	0.90	0.043	1.12	14627	9.6	0.279	0.496
a_{14}, a_{14}, a_{15}	0.477	0.257	0.220	1.85	-0.220	0.54	19087	6.5	0.474	0.437
a_{14}, a_{15}, a_{14}	0.435	0.255	0.179	1.70	-0.179	0.59	18279	6.7	0.413	0.412
a_{16}, a_{14}, a_{14}	0.361	0.267	0.093	1.35	-0.093	0.74	16092	7.6	0.302	0.387
a_{16}, a_{14}, a_{16}	0.265	0.255	0.010	1.04	-0.010	0.96	14262	9.3	0.197	0.346

TABLE 9
Impact-Reversed Sequential Activity Patterns in Separate Data Sets

Underlying sequence (P)	Impact 1	Derivative activity Q	Impact 2	Cir	Cps	Local support of $P \rightarrow$ Impact 1	Local support of $PQ \rightarrow$ Impact 2
a_{14}	$\bar{T}$	a_4	T	2.5	0.013	0.684	0.428
a_{16}	$\bar{T}$	a_4	T	2.2	0.005	0.597	0.147
a_{14}	$\bar{T}$	a_5	T	2.0	0.007	0.684	0.292
a_{16}	$\bar{T}$	a_7	T	1.8	0.004	0.597	0.156
a_{14}	$\bar{T}$	a_7	T	1.7	0.005	0.684	0.243
a_{15}	$\bar{T}$	a_5	T	1.7	0.007	0.567	0.262
a_{14}, a_{14}	$\bar{T}$	a_4	T	2.3	0.016	0.474	0.367
a_{16}, a_{14}	$\bar{T}$	a_5	T	2.0	0.006	0.402	0.133
a_{14}, a_{16}	$\bar{T}$	a_5	T	2.0	0.005	0.393	0.118
a_{16}, a_{15}	$\bar{T}$	a_5	T	1.8	0.006	0.339	0.128
a_{15}, a_{14}	$\bar{T}$	a_5	T	1.7	0.007	0.381	0.179
a_{16}, a_{14}	$\bar{T}$	a_7	T	1.6	0.004	0.402	0.108
a_{14}, a_{16}, a_{14}	$\bar{T}$	a_{15}	T	1.2	0.005	0.248	0.188
a_{16}, a_{14}, a_{14}	$\bar{T}$	a_{15}	T	1.2	0.005	0.267	0.220

disaster analysis [28], [26]. Data mining is used for tackling these issues. For instance, traditional data mining approaches such as association rule mining, clustering, and classification, and advanced data mining techniques like rule + exception analysis [37] have demonstrated the advantages of data mining in detecting and analyzing particular events and activities in ISI and national and homeland security areas [25], [26], [9].

However, the existing research mainly focuses on detecting the occurrence of an event or disaster. Rare studies have been on the antecedent of events and disasters, that is, a series of activities, as well as the relationship between the antecedent and consequent, that is, which activities or activity sequences actually or more likely result in an event or a disaster. Impact-targeted activity pattern mining aims to identify those activities that highly likely lead to the targeted impact. To this end, we need to develop effective techniques and algorithms to construct activity series or sequences and discover impact-targeted activities in an imbalanced data. This task challenges existing ISI research and data mining approaches.

In reality, impact-targeted activity data can be widely seen in the business world, with frequent customer contacts, interventions, events, and process updates, which may or may not lead to business outcomes or social impacts, for instance, dispersed terrorism activities resulting in a critical disaster to the national security. Therefore, it is important to deeply understand and analyze the activity data. However, it is complicated and challenging to discover interesting impact-targeted activity patterns in an imbalanced activity data. We further list some of the problems and ideas about the research on this promising area.

In the aggregation and construction of activity sequences in an imbalanced data, the strategy used may greatly impact

the performance of mining results. It is a research- and domain-specific issue to design effective methods for a sliding time window and constructing activities into sequences:

1. Further studies may be on the impact of an imbalanced class distribution and item set distribution on sequence construction.
2. The sliding window size may greatly affect the performance of activity pattern mining. For instance, strategy 2 may lead to an extra contribution to support and confidence. In social security debt-related activity pattern mining, the size of the sliding time window is sensitive to the following factors: population imbalance (some persons raise more debts than others), debt duration and frequency (the duration of frequent debtors is much shorter than those of rarely raising debt), the window size of the whole activity data set, and the specific targeted impact (the window size for income-related debt cannot be used for other types of debt). Domain knowledge, problem definition, descriptive statistics, and domain expert's discussion may play important roles in quantifying the window size.
3. There are weaknesses of both strategies 1 and 2 in Fig. 2. For instance, if the target event seldom or never happens, say, the terrorism attack in a specific city, then activity pattern mining based on strategies 1 and 2 hardly discloses interesting patterns. Unfortunately, such an imbalanced distribution exists in many applications. To handle this problem, a flexible window size may be defined to distinguish frequent targets from rarely frequent ones. This strategy needs to categorize the targets into multiple groups in terms of frequency. For more frequently targeted activities, strategy 2 is used for activity sequence construction. For rarely targeted activities, all the activities or the activities in a longer time frame are packed into one basket.

With respect to impact-targeted activity pattern mining, there are a lot of potentials. Here, we list some as follows:

1. Effective algorithms need to be developed to mine both positive and negative impact-targeted activity patterns in an imbalanced class distribution and item set distribution.
2. More research may be on the relationship between activities and targeted impact. For instance, there may exist a causal relationship between activity sequences and targeted impact, say, activities leading to debt. A semantic relationship such as linguistic relations may be further studied to understand the relations between activities or activity sequences that often co-occur in a set of situations.
3. Contrast analysis between positive and negative impact-targeted activity pattern mining may present meaningful information to distinguish patterns of interest to business concern. Further research on contrast activity pattern mining may disclose interesting information about those activities or activity sequences that significantly reverse their impact or differentiate one from another.

We have designed some new interestingness metrics to evaluate the performance of impact-targeted activity pattern mining. It is a challenging research issue to design a reliable and effective interestingness metric.

In the above analysis, we have proposed a series of technical interestingness and business interestingness metrics, which provide multidimensional and quantitative means to measure the specific computational performance and business impact of a pattern. However, they are not business friendly to business persons due to their complex statistical or proportional meaning and interleaved transmission. To make technical findings more understandable and friendly to business persons, a global measure, called *all interestingness (all_int())*, can be developed to combine technical and business interestingness and quantify the global actionable capability $act(P)$ of a pattern [3], [4]. One solution of establishing the *all interestingness* is to integrate individual contributions from all related technical and business concerns through developing a fuzzy aggregation algorithm [2].

7 Conclusions and Future Work

Impact-targeted activity pattern mining is of great interest to both research and applications, especially ISI-related areas such as analyzing counterterrorism activities, criminal activities, and fraudulent activities. However, an impact-targeted activity data presents special structural complexities, in particular, an imbalanced class distribution and imbalanced item set distribution. Mining rare activities leading to a significant impact to the society and national security is very challenging in both ISI and data mining areas.

This paper has addressed the above interesting issue. We have proposed effective and practical techniques, algorithms, and interestingness measures for discovering rare but significant *impact-targeted activity patterns* in an imbalanced activity data. In particular, we have developed algorithms to mine *frequent impact-oriented activity patterns* leading to either targeted impact or impact disappearance, *impact-contrasted activity sequential patterns* differentiating the significance of the same activity sequence resulting in contrast impacts, and *impact-reversed sequential activity patterns*, in which the appearance of a derivative activity/sequence on the underlying pattern triggers the reversal of pattern impact. We have also developed new technical interestingness metrics and business metrics to evaluate the performance of identified impact-targeted activity patterns. The proposed approaches have been demonstrated in the Australian social security activity data. The experimental results show that the proposed approaches are promising for identifying impact-targeted activity patterns in many complex applications in the ISI field such as analyzing criminal activities and terrorism activities before they result in a disastrous event to the society.

We have discussed potential issues in this interesting and significant area in Section 6. With the Australian Research Council (ARC) Linkage Grant and Centrelink Support, we are performing impact-targeted activity prediction and developing integrated all-interestingness measures for assessing the risk of identified activity patterns.

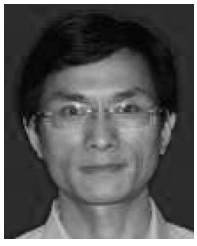
Acknowledgments

The authors would like to express our gratitude to Mr. Fernando Figueiredo for his domain knowledge and helpful comments. Moreover, they would also like to thank Dr. Jie Chen for his technical discussion and to Ms.

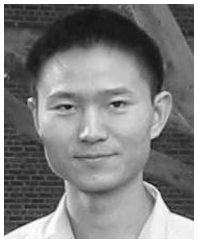
Yvonne Morrow, Mr. Peter Newbigin, Mr. Rick Schurmann, Mr. Carol Ey, and Ms. Michelle Holden at Centrelink, Australia, for their domain knowledge and support. This work is sponsored by Australian Research Council Discovery Grant (DP0773412, DP0667060) and ARC Linkage (LP0775041), and UTS internal grants.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proc. Int'l Conf. Very Large Data Bases (VLDB '94),* pp. 487-499.
[2] L. Cao and C. Zhang, "Fuzzy Genetic Algorithms for Pairs Mining," *Proc. Ninth Pacific Rim Int'l Conf. Artificial Intelligence (PRICAI '06),* pp. 711-720, 2006.
[3] L. Cao and C. Zhang, "Domain-Driven Data Mining: A Practical Methodology," *Int'l J. Data Warehousing and Mining,* vol. 2, no. 4, pp. 49-65, 2006.
[4] L. Cao and C. Zhang, "Two-Way Significance of Knowledge Actionability," *Int'l J. Business Intelligence and Data Mining,* vol. 4, 2007.
[5] L. Cao, Y. Zhao, C. Zhang, and H. Zhang, "Activity Mining: From Activities to Actions," *Int'l J. Information Technology and Decision Making,* 2007.
[6] "Integrated Activity Management Developer Guide," technical report, Centrelink, Sept. 1999.
[7] *Centrelink Annual Report 2004-05,* Centrelink, 2004.
[8] N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations Newsletter,* vol. 6, no. 1, June 2004.
[9] H. Chen, F. Wang, and D. Zeng, "Intelligence and Security Informatics for Homeland Security: Information, Communication, and Transportation," *IEEE Trans. Intelligent Transportation Systems,* vol. 5, no. 4, pp. 329-341, 2004.
[10] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime Data Mining: A General Framework and Some Examples," *Computer,* vol. 37, no. 4, pp. 50-56, Apr. 2004.
[11] J. Chen, H. He, G. Williams, and H. Jin, "Temporal Sequence Associations for Rare Events," *Proc. Eighth Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '04),* pp. 235-239, 2004.
[12] H. Chen and F. Wang, "Guest Editors' Introduction: Artificial Intelligence for Homeland Security," *IEEE Intelligent Systems,* vol. 20, no. 5, pp. 12-16, Sept./Oct. 2005.
[13] H. Chen, *Intelligence and Security Informatics for International Security: Information Sharing and Data Mining.* Springer, 2006.
[14] G. Dong and J. Li, "Efficient Mining of Emerging Patterns: Discovering Trends and Differences," *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '99),* pp. 43-52, 1999.
[15] H. Guo and H. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach," *ACM SIGKDD Explorations Newsletter,* special issue on learning from imbalanced datasets, vol. 6, no. 1, pp. 30-39, June 2004.
[16] V. Guralnik and J. Srivastava, "Event Detection from Time Series Data," *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '99),* pp. 33-42, 1999.
[17] M. Hammori, J. Herbst, and N. Kleiner, "Interactive Workflow Mining-Requirements, Concepts and Implementation," *Data and Knowledge Eng.,* vol. 56, pp. 41-63, 2006.
[18] J. Han, J. Pei, and X. Yan, "Sequential Pattern Mining by Pattern-Growth: Principles and Extensions," *Recent Advances in Data Mining and Granular Computing,* W.W. Chu and T.Y. Lin, eds. Springer, 2005.
[19] J. Han and M. Kamber, *Data Mining: Concepts and Techniques,* second ed. Morgan Kaufmann, 2006.
[20] N. Japkowicz, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies," *Proc. AAAI Workshop Learning from Imbalanced Data Sets,* 2000.
[21] P. Kantor et al., "Intelligence and Security Informatics," *Proc. Third IEEE Int'l Conf. Intelligence and Security Informatics (ISI '05),* 2005.
[22] D. Luo, L. Cao, C. Luo, and C. Zhang, "Towards Business Interestingness in Actionable Knowledge Discovery," *Proc. PAKDD Workshop Data Mining for Business '07,* 2007.
[23] J. Mena, *Investigative Data Mining for Security and Criminal Detection,* first ed. Butterworth-Heinemann, 2003.
[24] J. Mena, *Homeland Security Techniques and Technologies (Networking Series).* Charles River Media, 2004.
[25] National Strategy for Homeland Security, Office of Homeland Security, 2002.
[26] Nat'l Research Council, *Making the Nation Safer: The Role of Science and Technology in Countering Terrorism.* Nat'l Academy Press, 2002.
[27] W. Potts, *Survival Data Mining: Modeling Customer Event Histories.* Wiley and Sons, 2006.
[28] M. Sageman, *Understanding Terror Networks.* Univ. of Pennsylvania Press, 2004.
[29] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Trans. Knowledge and Data Eng.,* vol. 8, no. 6, pp. 970-974, Dec. 1996.
[30] M. Skop, *Survival Analysis and Event History Analysis.* Wiley, 2005.
[31] W.M.P. Van der Aalst et al., "Process Mining: A Research Agenda," *Computers in Industry,* vol. 53, pp. 231-244, 2004.
[32] G. Wang, H. Chen, and H. Atabakhsh, "Automatically Detecting Deceptive Criminal Identities," *Comm. ACM,* vol. 47, no. 3, pp. 71-76, 2004.
[33] F.-Y. Wang, C. Karleen, D. Zeng, and W. Mao, "Social Computing: From Social Informatics to Social Intelligence," *IEEE Intelligent Systems,* vol. 22, no. 2, pp. 79-83, Mar./Apr. 2007.
[34] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications.* Cambridge Univ. Press, 1994.
[35] G. Williams et al., "Temporal Event Mining of Linked Medical Claims Data," *Proc. Seventh Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '03),* 2003.
[36] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," *ACM Trans. Information Systems,* vol. 22, no. 3, pp. 381-405, 2004.
[37] Y. Yao, F. Wang, J. Wang, and D. Zeng, "Rule + Exception Strategies for Security Information Analysis," *IEEE Intelligent Systems,* vol. 20, no. 5, pp. 52-57, Sept./Oct. 2005.
[38] J. Zhang, E. Bloedorn, L. Rosen, and D. Venese, "Learning Rules from Highly Unbalanced Data Sets," *Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM '04),* pp. 571-574, 2004.
[39] Y. Zhao, L. Cao, J. Chen, and C. Zhang, "Centrelink Improving Income Reporting Data Exploration Report," technical report, Mar. 2006.
[40] Y. Zhao, L. Cao, Y. Morrow, Y. Ou, J. Ni, and C. Zhang, "Discovering Debtor Patterns of Centrelink Customers," *Proc. Australasian Data Mining Conf. (AusDM '06),* Nov. 2006.



Longbing Cao is a senior lecturer in the Faculty of Information Technology, University of Technology, Sydney. His research interests include data mining, multiagent technology, and agent and data mining integration. He is a senior member of the IEEE.



Yanchang Zhao is a postdoctoral research fellow in the Faculty of Information Technology, University of Technology, Sydney. His research interests include clustering, association rules, and domain-driven data mining. He is a member of the IEEE.

Chengqi Zhang is a research professor in the Faculty of Information Technology, University of Technology, Sydney. His research interests include data mining, multiagent technology, and agent and data mining integration. He is a senior member of the IEEE.