Identifying and Measuring Associations of Temporal Events

Hsueh-Chien Cheng* Catherine Plaisant† Ben Shneiderman‡

Department of Computer Science & Human-Computer Interaction Lab, University of Maryland

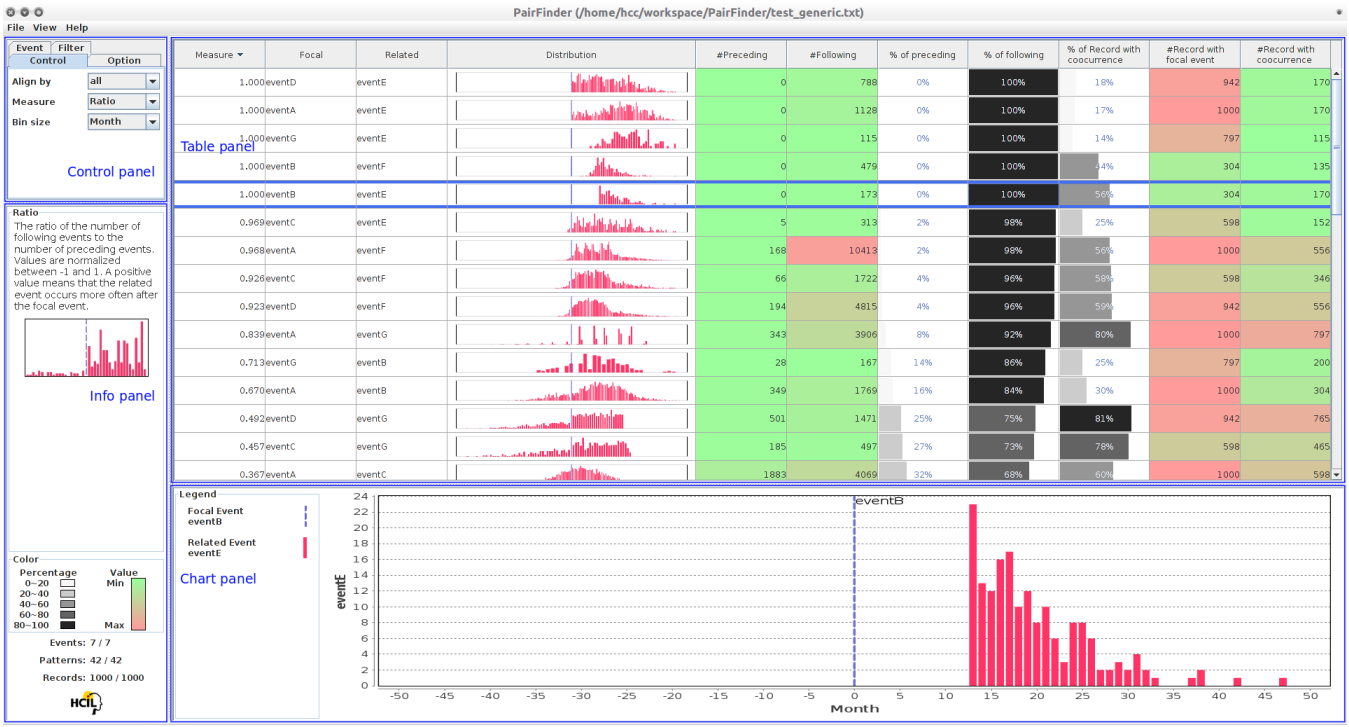


Figure 1: PairFinder is used to explore a dataset of 1000 records containing 7 types of events (A, B, C etc.). The table lists all possible pairs events. Each row analyzes one pair of events (one called "focal" and the other "related") by looking at the temporal distribution of the related events relative to the focal event. A user-selected measure here the ratio of following versus preceding events (explained in the info panel and displayed in the leftmost column) can be used to rank the pairs and bring interesting pairs to the top or bottom. Other measures are shown on the color-coded columns. The bottom panel shows a detailed histogram for the selected pair.

Abstract

Large databases of temporal records have made it possible for researchers to verify their hypotheses related to temporal event sequences. However, with the overwhelming size of data and numerous possible patterns, an important issue is what patterns should be highlighted and presented to users. We implement a visualization tool, PairFinder, to enable users to efficiently locate patterns of interest. Users can 1) see all the results of the potential event patterns and 2) use interestingness measures to rank event patterns by their interestingness. In addition, users can hide irrelevant patterns and filter records by record attributes. By looking only at the top-ranked patterns, users can easily scan large number of patterns. We demonstrate the potential of PairFinder with four case studies and summarize the patterns found in the data sets.

*e-mail: cheng@cs.umd.edu
†e-mail:plaisant@cs.umd.edu
‡e-mail:ben@cs.umd.edu

Keywords: Information Visualization, Temporal Event Sequence.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI)

1 Introduction

Increasingly rich temporal databases enable researchers to conduct detailed studies of temporal event sequences. When researchers have hypotheses of which events lead to other events, current tools, such as Lifelines2 [21], assist them to find those events, even in complex data sets.

The challenges of generating hypotheses of temporal associations among temporal events lie in the large number of possible complex patterns and the knowledge discovery task in an exploratory context. Even if only pairs of events are considered, a significant number of possible event pairs exist. In addition, temporal associations with relative timing are far more complicated than event sequences, which only consider the order among events. When conducting exploratory tasks, users do not have a particular pattern they are looking for. Existing systems which require users to input pattern structures such as database querying are not suitable

for exploratory tasks. These challenges increase the difficulty of efficient pattern discovery. New systems could be helpful in finding potential associations among events in the data stream.

As a first step toward generating hypotheses of causal relations, we limit the number of possible patterns by looking at only pairs of events. For example, an event triggers another event exactly two weeks later. This ideal case rarely exists, so search strategies that rank the strength of association among all event pairs are attractive. Using the search strategy, users can easily examine a smaller set of top-ranked patterns by ranking the associations and filtering out those with strength less than a predefined threshold. The top-ranked patterns may not be the most interesting patterns to users: they may be intuitive and common patterns users already know. Thereafter during exploratory analysis, users can scan the patterns in ascending order of their rank until they find something interesting to investigate further. Defining such measures to rank the strength of associations is challenging and different applications may prefer different measures. A few measures were included in our system to assess the interestingness of patterns from different perspectives.

We follow the framework proposed in [20], where the temporal data is aligned by focal events. After alignment, the other related events either precede the focal event, follow the focal event, or they occur at the same time. In general, the occurrences of events following the focal event are considered to contribute to the strength that the following event is caused by the focal event. By contrast, the occurrences of preceding event weaken the strength of that event is caused by the focal event. A simple pattern of causal relation can therefore be represented by a pair of events.

We introduce *PairFinder* to enable interactive analysis of data sets. The distribution of each pair of events is presented to users. Users can choose different time granularities (e.g. day, week, month) to categorize the occurrences of event in the relative time frame. Appropriate filtering reduces the number of visible patterns. Ordering the patterns with interestingness measures helps users locate interesting patterns more efficiently.

The rest of the paper is organized as follows: We review the literatures on related topics in sequence mining, interestingness measures, and cause-and-effect in section 2. Section 3 describes the alignment scheme. The proposed measures and the user interface of PairFinder are introduced in section 4 and section 5. We describe the evaluation conducted to demonstrate our proposed features in section 6. Finally, we conclude the work and present future directions in section 7.

2 Related Work

2.1 Sequence mining

Sequence mining refers to the category of methods for finding interesting correlations of events occurring in a sequence. For example, the event sequence "a serious side effect A is more likely to happen if the patient takes two specific drugs, B and C, in sequence" may be of interest to public health researchers.

Agrawal and Srikant, Zaki, and Pei et al. modeled the problem of sequence mining as finding the maximal length sequences above a support threshold [1, 24, 16], where sequences are ordered lists of itemsets ordered by their transaction times. Bettini et al. addressed the problem of mining frequent sequences with additional temporal constraints on multiple time granularities [2]. Campagna and Pagh proposed a graph model to mine frequent patterns of sequences [3]. Instead of common sequences, uncommon sequences with high prediction power are useful in the context of failure prediction. Weiss and Hirsh proposed a genetic algorithm to predict rare events with sequential pattern [23].

Most of the work considered the order of events instead of the relative time differences among events. We address a problem of different patterns, for example, "a serious side effect A is more likely to happen *after 3 days* if the patient takes two specific drugs, B and C, in sequence". The temporal constraints introduced in [2] require users to define the event structure in advance, which is not suitable for exploratory tasks.

2.2 Interestingness measure

The number of rules produced by a data mining algorithm may far exceed the capability of a human to examine manually. Preselecting a smaller set of rules by interestingness measures helps users assess the result more efficiently. However, choosing an appropriate measure is non-trivial since different measures define interestingness from different perspectives. The optimal measure to apply is application-dependent and relies on the domain knowledge of experts.

Besides using objective measures, which are calculated based on the data and the pattern itself, many recent studies designed systems with user interaction. Klemettinen et al. suggested that users may specify the template of interesting and uninteresting association rules [10]. They also designed a simple user interface to construct templates and select rules. Sahar proposed a system where the association rules marked by users as not interesting are used to eliminate other possible uninteresting rules [17]. Tan et al. proposed a method to find appropriate measures where users are asked to rank the interestingness of a small set of association rules [19]. Couturier et al. introduced an interactive matrix visualization with a fisheye view to help users choose interesting association rules [6]. Lenca et al. applied a multiple criteria decision aid approach to generate the ranking of measures with respect to different scenarios [11].

The existing measures and systems in the literature were mostly designed for association rules, which are different to the patterns we address. Nonetheless, user interaction has been applied successfully in selecting appropriate interestingness measures. An effective human interface may provide the right cognitive support to improve the selection process.

2.3 Cause-and-effect

Explaining why things happen has long been a compelling topic. A systematic way of finding causal relations is essential to efficient discovery of such relations. Experimental studies can be conducted to understand the causal structure of complex systems. However they may not be feasible for ethical, cost, or technical reasons. Observational data is therefore useful for researchers to apply computational methods and infer causal relations [8, 15]. Many recent learning algorithms [18, 4, 5] were developed to learn a causal Bayesian network, which is a graph model of causal relationship. However, introducing temporal representations in a Bayesian network is non-trivial and may not be well suited to exploratory tasks.

Norén et al. introduced an alignment scheme of events and an observed-to-expected ratio [13] as an indication of interestingness to discover adverse drug reactions [12]. A static visualization was created with the ratio and the expected/observed number of events over time side-by-side to show the association between pairs of events. We use a similar alignment scheme but design an interactive graphic user interface to manipulate the data set.

Effective visualizations of the causal relationship were also studied in the literature. An animation was created to visualize causal relations in [7]. In [9], various semantics of causal relations were visualized through both static and animated displays.

3 Alignment Scheme

Denote the record set by R and the event set by E. A record $r_j \in R$ is composed of sequences of timestamped event, which are denoted by pairs of event type $e_i \in E$ and timestamps t_i.

$$< e_i, t_i >$$

After aligning by a specific event occurrence $< e_i, t_i >$, the timestamps of the rest of the event occurrences in record r_j is subtracted by t_i to create a relative time frame. We use *focal event* and *related event* to refer to the event by which the records are aligned (e.g. e_i) and the event with which the relative time frame is calculated. Event occurrences that precede $< e_i, t_i >$ have negative timestamps, while those that follow $< e_i, t_i >$ have positive timestamps. To align the records by *all* the occurrences of the focal event, all the focal event occurrences are used to create alignments.

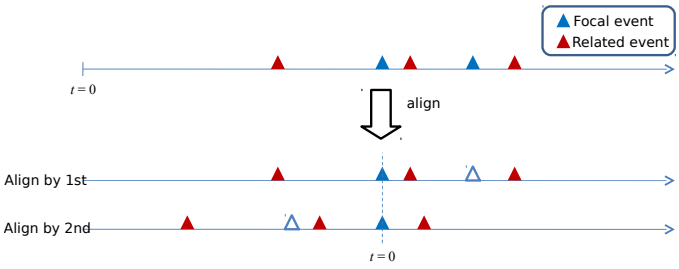


Figure 2: The occurrences of the related event aligned by all the occurrences of the focal event. Note that there are two occurrences of the focal event, therefore both the alignments of the first and second occurrences are calculated.

An aggregated alignment can be constructed by putting together all the alignments calculated for the records. A histogram can be created for a pair of focal event and related event to summarize the frequency of the related event occurrences by grouping them with respect to the relative time differences.

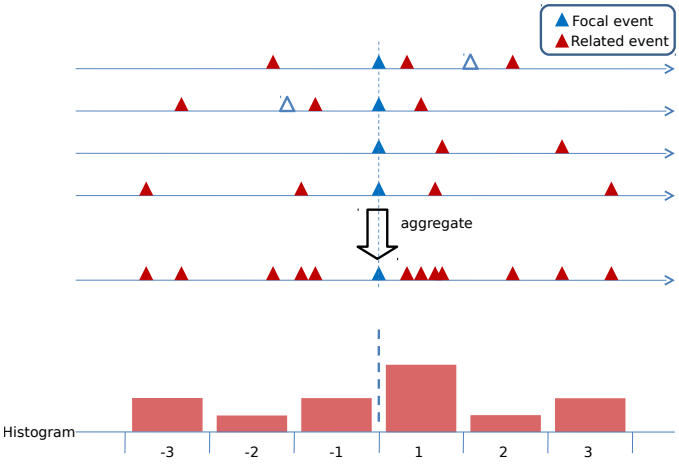


Figure 3: The aggregated alignment of the alignments calculated from three records. A histogram with equal bin size is created with the aggregated alignment.

4 Interestingness Measure

With the histogram of occurrences of the related event, different measures can be applied depending on the properties of the pattern which users intend to find. Ranking candidate patterns with a measure provides an easy approach for users to conduct exploratory analysis. Patterns which better fit the required property will show at the top of the list, therefore reduce the number of irrelevant patterns users must scan before reaching the interesting ones. In the following we introduce the measures included in PairFinder.

4.1 Occurrence ratio

Intuitively, the total number of occurrences of the related event which precede and follow the focal event is a useful indicator to determine the association between the two events. Denoted by n^- and n^+ the number of occurrences of the related event preceding and following the focal event, the value of the measure is calculated as:

$$\begin{cases} 2(\frac{max\{n^-, n^+\}}{n^- + n^+}) - 1, & n^+ > n^- \\ -[2(\frac{max\{n^-, n^+\}}{n^- + n^+}) - 1], & n^- > n^+ \\ 0, & n^- = n^+ \end{cases} \quad (1)$$

The sign of the value indicates either more occurrences come before or after the alignment. Note that the value is normalized into $[-1, 1]$.

4.2 Peak ratio

Sometimes the total number of occurrences does not carry as much information as the location of the peaks in the histogram. For example an evenly distributed histogram with roughly the same number of preceding and following occurrences may have all peaks following the focal event. Peak ratio calculates the ratio of the number of peaks which precede and follow the focal event by substituting the number of occurrences in Eqn. 1 by the number of peaks. The peak detection algorithm we used is similar to that described in [14].

4.3 Periodicity

Periodic patterns, such as "event B happens every 30 days after event A", are interesting but unlikely to be detected using merely the two aforementioned ratio measures. The value of the periodicity measure is the period, which is the reciprocal of the frequency with maximum amplitude after applying a fast Fourier transform. If no periodicity is detected (e.g. the period is infinity), the value of the measure is infinity.

4.4 Standard deviation

A well distributed histogram which looks similar to a uniform distribution has lower standard deviation. This measure is useful in finding the histograms which look different to a uniform distribution.

5 User Interface

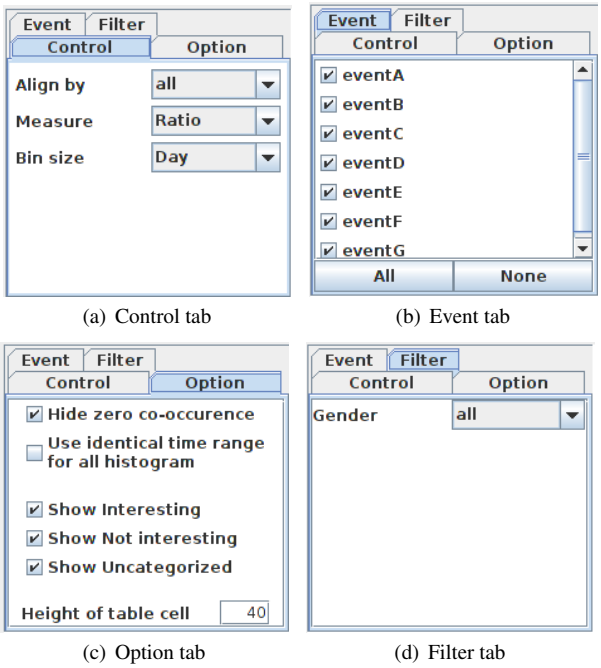
5.1 Overview

As shown in Fig. 1, the interface of PairFinder is separated into four panels: control panel, info panel, table panel, and chart panel. Users control PairFinder mainly through the widgets in control panel. The contents in the table panel change accordingly after users interact with PairFinder. Users can click on the table to select a pair of events and see the detailed histogram in the chart panel. The info panel shows the description of the measure applied and the information about the data set.

5.2 Control panel

5.2.1 Control tab

Each record in the data set is aligned by all occurrences of the focal event by default. Depending on the structure of patterns users intend to discover, they can choose to align by the k-th occurrences. Users need to choose an appropriate bin size to create meaningful visualization for interpretation. The histogram will be too sparse when the bin size is too large, and will be too dense otherwise. The measures in section 4 provide viable ways to scan large number of candidate patterns by sorting the event pairs with respect to measure values.



(a) Control tab (b) Event tab

(c) Option tab (d) Filter tab

Figure 4: Four tabs in control panel. Control tab controls how the histogram is created and the measure used. The event tab enables users to show/hide the events. The option tab shows additional options for flexibility. The filter tab shows the record attributes with which users can filter and use a subset of the records in the original data set.

5.2.2 Event tab

Users can show/hide an event by clicking on the checkbox with the event name in the right (Fig. 4(b)). Hiding irrelevant events reduces the number of event pairs showing in the table panel. Note that a pair of events will not show if more than one of the events is hidden.

5.2.3 Option tab

By default, the event pairs with zero co-occurrence (e.g. there are no records in which the focal event and related event occurred together) are hidden. As an alternative, users can choose to use an identical time range for all the histograms. An identical time range enables better comparison of different histograms since the bins are aligned. However, for histograms with a significantly smaller time range the bars can reside densely in a small region. Users can categorize a pair of events indicating whether the pair is interesting by right-clicking the corresponding row and change its category (e.g. interesting, not interesting, and uncategorized). Unchecking/checking the checkboxes in the option tab will hide/show pairs of events in that specific category.

5.2.4 Filter tab

Each record in the dataset may carry additional attributes, which are referred to as record attributes, such as gender of the patient. Users can filter the records by these attributes. For example, users can select to use only the subset of records with gender "Female" (Fig. 4(d)). Non-female records will not be included in the calculation of alignment. If more than one record attributes exists, multiple filters can be specified.

5.3 Info panel

The description of the measure used shows in the info panel. On the bottom of the info panel are the number of visible/total events, the number of visible/total pairs of events, and the number of eligible/total records. Note that users can hide/show specific events (section 5.2.2), and filter records by record attributes (section 5.2.4).

5.4 Table panel

For the row corresponding to a focal event and related event pair, the table has the following columns from left to right. By clicking at the column header, the users can order the rows in ascending/descending order with respect to the values in that column.

1. Measure
2. The name of focal event
3. The name of related event
4. The histogram calculated from the aggregated alignment
5. The number of occurrences of the related event preceding the focal event, n^-
6. The number of occurrences of the related event following the focal event, n^+
7. The percentage of the occurrences of the related event preceding the focal event, $\frac{n^-}{n^-+n^+} \times 100\%$
8. The percentage of the occurrences of the related event following the focal event, $\frac{n^+}{n^-+n^+} \times 100\%$
9. The percentage of the records with focal event which also have the related event
10. The number of records with focal event
11. The number of records with co-occurrence of related event and focal event

For the columns with real values (e.g. column 5, 6, 10, 11), the cells with maximum/minimum value in the column have background color red/green. The background of other cells is colored with linearly interpolated color depending on their values. For the columns with percentages (e.g. column 7, 8, 9), the background is a bar which occupies the corresponding percentage of the cell width. For example, a cell with value 70% shows a bar in the background with bar width equals 70% of the cell width. The bars are colored in five gray scale colors from black to white for percentages equally spaced from 100% to 0%.

5.5 Chart panel

A detailed histogram is shown in the chart panel with the legend in the left. The users can zoom-in to specific region of the histogram by dragging the mouse cursor to select the region. The enlarged histogram enables fine-grained inspection of the distribution.

5.6 Comparison among groups of records

To facilitate easy comparison of the histogram with respect to different record groups (e.g. male and female), the users can click on the menubar "view → view comparison..." and open a new window to see histograms of different groups at a time. The controlling widgets are similar to those used in the control tab (section 5.2.1) with an additional drop-down list of record attributes to select an attribute with which the data is split into groups. Identical time ranges are used for all the histograms to enable easy comparison.

Fig. 5 shows an example of the dataset split by record attribute "Attribute 1". In the group of records with "Attribute 1" equals "True", "Event A" mostly follows "Event B" with a few exceptions. In the other group which "Attribute 1" equals "False", "Event A"

Data set	No. records	No. event	No. occurrences
Website logs	9538	17	28614
Graduate student history*	1000	7	12177
Real-time strategy game logs	158	70	30046
Sport logs	199	31	87625
Medical records	6583	19	135951

*synthetic data

Table 1: The data sets used in the evaluation.

always precedes "Event B". By looking at the two histograms, the users can easily find out that the associations between the pair of events are significantly different in the two groups. Note that the number of possible values of an attributes can be more than two though in the example the attribute has boolean value.

6 Evaluation

We conducted four case studies with the data sets in Table 1. In the following we summarize the results.

6.1 Website logs

Our first case study was conducted with a member in our lab to analyze a website log. The study was conducted with an early version of PairFinder. After explaining briefly the user interface and the alignment scheme, the user operated PairFinder without assistance and was encouraged to think-aloud during the session.

The user was comfortable with PairFinder showing all the pairs of the events. It was easy to locate the target event pair through sorting the rows with event names. The histogram and the measures were easy to interpret. Among the four measures, the periodicity measure took the user more time to understand. However, the user commented on the periodicity measure to be potentially useful in finding recurring events.

During the 40-minute session, the user was able to identify some patterns of interest. For example, the events involving user model and point-of-interest come shortly after a search event.

Based on the feedback, we used bar length to encode visually the percentage of the cell in addition to color. Better descriptions of the measure were shown in the info panel. Other minor visual hints were added to the user interface.

6.2 Graduate student history

The graduate student history dataset contains 1000 synthetic records with seven events such as class sign-up, proposal, and paper submission.

A good starting point is to align the records by the first occurrence of focal event and use the periodicity measure. Only four event pairs have non-infinite periods. We easily identified the occurrences of "Class Signup" after aligned by the first occurrence of "Master Degree" have a period of 6.733 months. This is because students got a master's degree at the end of the semester and registered for classes at the beginning of the semester. Therefore, the period is about six month after alignment by the time students got master. Students registered for more classes were registered before they got a master's degree.

Another useful strategy is to align the records by the first occurrence of focal event and the use standard deviation measure. The histogram of the occurrences of "Job Interview" after aligning by the first occurrence of "Master Degree" has the fourth largest standard deviation. Most job interviews occurred shortly before and after students got master degree.

6.3 Real-time strategy game logs

In a later case study, we studied the game logs of *StarCraft*, a classic real-time strategy game which has been famous among both casual and professional players. We focused on the strategy of Protoss,

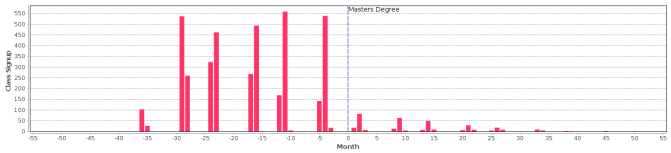


Figure 6: The histogram of "Class Signup" aligned by the first occurrence of "Master Degree". Students signed up for class roughly every 6 months. More classes were taken before students got master degrees.

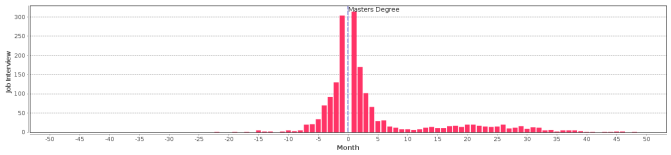


Figure 7: The histogram of "Job Interview" aligned by the first occurrence of "Master Degree". Job interviews occurred shortly before and after students got master degrees.

which is one of the three different races, namely Protoss, Terran, and Zerg. We downloaded 158 1-on-1 Protoss vs. Terran game replays from a website collecting numerous replays and converted into game logs with a ripper. The logs contain the events of players constructing buildings, train units, and research abilities and upgrades. There are 70 events in the dataset including one additional event, which we added to mark the start of the game. The Protoss strategies are classified using the method in [22]. As record attributes, each record (e.g. one Protoss player playing one game) is labeled with its strategy. In the 158 replays we used, 58 replays came from World Cyber Games, a prestigious international competition for professional gamers, and 100 replays were public games.

The main interest is to find patterns related to combat units/researches. After hiding rare events which occurred in less than 5 records and non-combat units/buildings, we were left with 42 events and 1702 visible patterns.

Many of the visible patterns are trivial associations, which are enforced by the game rules. For example, the required buildings of certain combat unit must be built before that unit can be trained. However, the time between the construction of required buildings and the training of units may be affected by the level of competition and vary among strategies.

In the following we show three interesting patterns found. We confirmed the patterns we found with expert players and summarize the possible explanations. We sometimes decided to align by the first occurrence because units/researches are available once the player has at least one required building for those units/researches. Aligning by the first occurrence is a reasonable choice over the default align-by-all.

6.3.1 Dark Templar and High Templar

The players can train Dark Templar and High Templar if they have at least one Templar Archive. Dark Templars are invisible and deal considerable damage. High Templars can only cast spells and do not attack directly.

Aligning the records by the first "Build Templar Archive", Dark Templars are trained more frequently than High Templars shortly after the Templar Archive is built. However, more High Templars are trained at a later time.

http://www.gosugamers.net/starcraft/replays/
http://lmrb.net

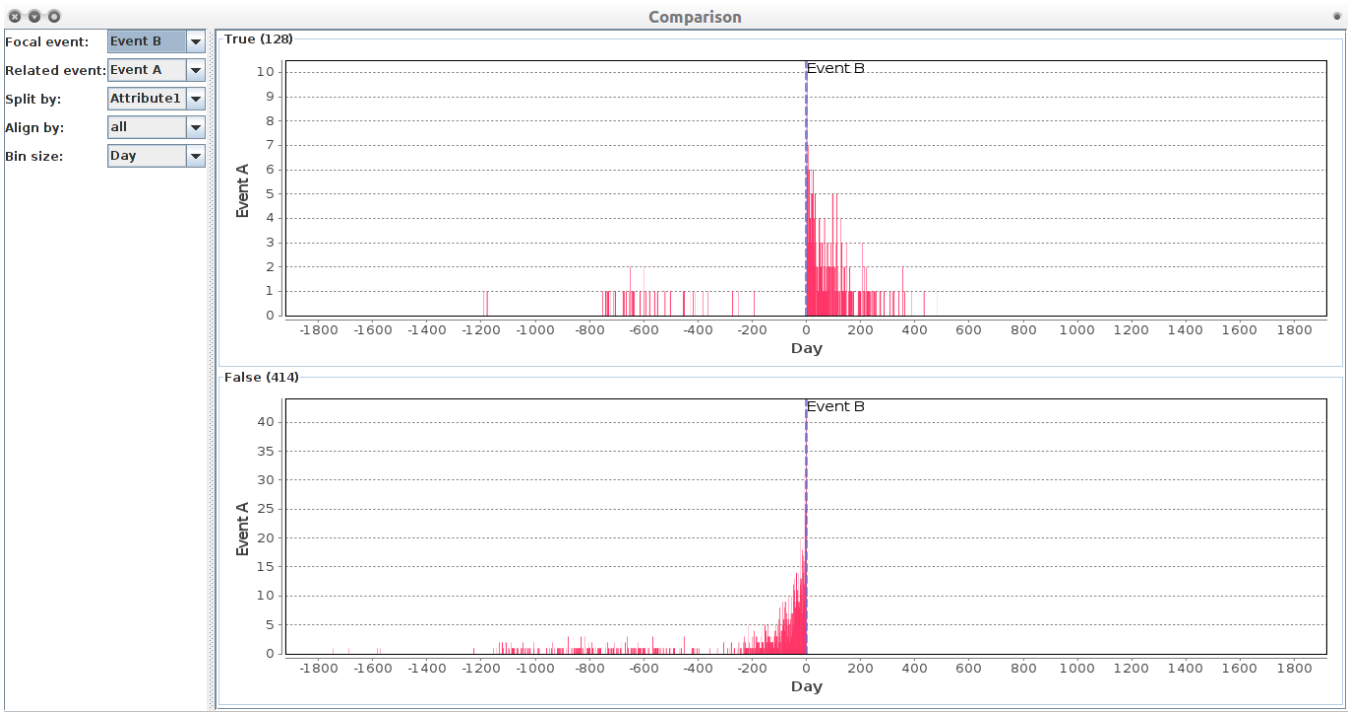
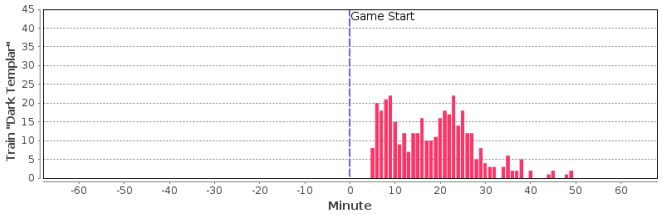
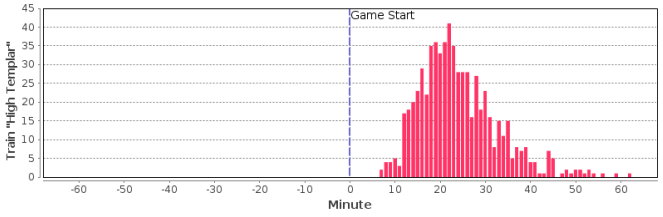


Figure 5: Compare between groups of records with different record attributes using the comparison window. After aligning by "Event B", the distributions of "Event A" are significantly different in the two groups split by "Attribute 1".

The comment from the expert players is that Dark Templars are mostly used to disturb the opponent at an early stage of the game. To confirm this comment, we aligned records by the event "Game Start", which marks the starting time of the game, to see the distribution of the time Dark Templars and High Templars are trained. As shown in Fig. 8, there are more Dark Templars in the first 11 minutes of the game, and more High Templars after.



(a) Histogram of Dark Templar



(b) Histogram of High Templar

Figure 8: The histograms of "Train High Templar" and "Train Dark Templar" aligned by the first occurrence of "Game Start". More Dark Templars were trained in the first 11 minutes after the start of the game.

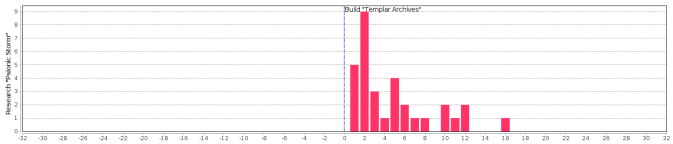
6.3.2 Templar Archive and Psionic Storm

Psionic Storm is an area effect spell which damages the units in the casted region. It is available for players to research at Templar Archive. Though the spell does a significant amount of damage, it takes fine control to cast properly on enemy units and the research cost is considerable.

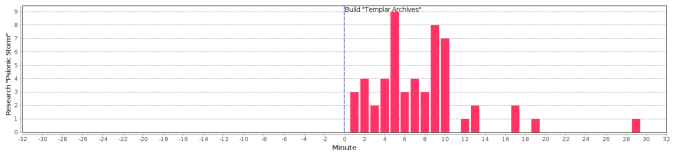
After aligning by the first occurrence of "Build Templar Archive", the histograms with related event "Research Psionic Storm" are shown in Fig. 9. After further inspection of the histograms, we found that 44% WCG players had "Research Psionic Storm" in 2 minutes following the focal event. 75% of the WCG players had "Research Psionic Storm" in 6 minutes. While only a small fraction (14%) of public game players had "Research Psionic Storm" in 2 minutes. About a half (49%) of public game players completed that research in 6 minutes.

The difference may come from the capability gap between professional players and common players. Professional players have better arrangements of their resources and thereafter schedule the game plan more precisely than common players, which may lead to a closer smaller time difference between the research and the completion of the building. Another possible reason is that Psionic Storm requires better skill to use properly during a combat, therefore common players do not research them with high priority as professional players do. One rejected hypothesis interesting enough to mention is that, instead of Psionic Storm, people may build Templar Archive only for the purpose to enable second level weapon/armor upgrade. If this were true, we expected to see the corresponding upgrade event shortly after Templar Archive was built. However, there was no evidence found in the data to support this idea.

We then split the records with record attributes of strategies using the comparison window. A small number of players, 8 in WCG and 8 in public game, playing "Fast DT" had the least percentage (18%) of having the related event within 6 minutes. The reason for this is that players playing "Fast DT" build Templar Archive for a



(a) WCG players



(b) Public game players

Figure 9: The histograms of "Research Psionic Storm" aligned by the first "Build Templar Archive" with different groups of players. WCG players and public game players had distinct choices on the time of the related event.

different purpose. They may build Templar Archive very early in order to train Dark Templar sooner, but not for Psionic Storm.

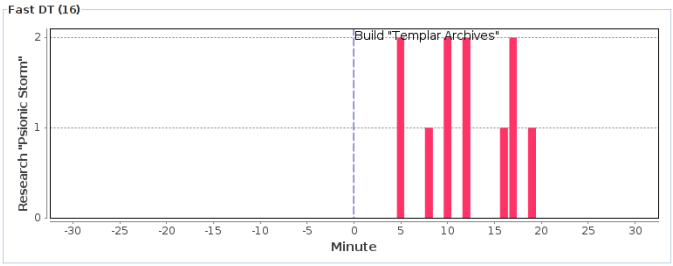


Figure 10: The histogram of "Research Psionic Storm" aligned by the first "Build Templar Archive" with the players using strategy "Fast DT". The related event mostly occurred more than 6 minutes after the focal event.

6.3.3 Shuttle and Reaver

Shuttles are flying units used for transportation of ground units. Reavers are units which deal devastating damage but they are fragile and move slowly.

Aligning the records by all "Shuttle", many "Reaver" come shortly after. The total number of events precede and follow the focal event is not significantly different (e.g. 53% and 47%). The ratio measure value of this pattern is -0.051, in which a value close to zero indicates the pattern is not interesting. However, using the peak ratio measure, the value is 1.000, which means the patterns is very interesting and the peaks of "Reaver" always comes after "Shuttle". We were able to identify this pattern ranking all the patterns with peak ratio easily.

This particular pattern comes from a practice called "Reaver Drop", where players build Shuttles to transport Reavers. Shuttles are fast moving flying units which ignore the limitation of terrain. Therefore using Shuttles with Reavers enables better flexibility. See Fig. 11 for the histogram of "Reaver" after the records are aligned by "Shuttle".

6.4 Sport logs

We downloaded 199 play-by-play summaries of NBA Basketball games of season 2011–2012 in November, 2011 from the official

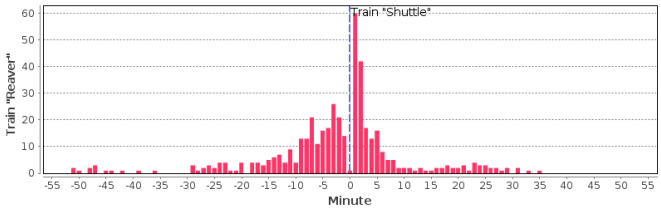


Figure 11: The histograms of "Train Reaver" aligned by all "Train Shuttle". Many Reavers were trained after the training of Shuttle in two minutes.

NBA website. 31 events are extracted from the downloaded webpages, including 4 events marking the beginning of the first, second, third, and fourth quarter and one event marking the end of the fourth quarter. The events in overtime (e.g. after the end of the fourth quarter) were omitted. The remaining 26 events consist of the events on court split into two groups: those committed by the home team and the away team. For example, the home team made a three point jump shot or the away team got a rebound are included as events in the logs.

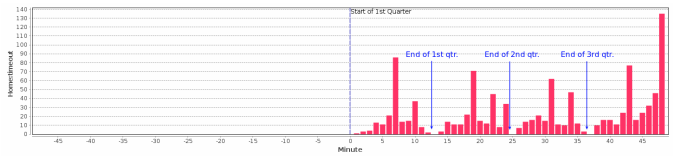
The obvious patterns are those enforced by the game rule. For example, the beginning of the second quarter occurs exactly 12 minutes after the beginning of the first quarter since each quarter is 12 minutes long. Or the home team made or missed a free throw after a foul of the away team. Many of the others are trivial such as the home team got a rebound shortly after the away team missed a shot. In the following we summarize some interesting patterns we found in the data set.

Aligned by all occurrences of events, all event pairs with occurrence ratio 1.000 and -1.000 are apparent in the order they occurred in the records (e.g. all events occurred after the start of the first quarter). With histograms showing the frequency of occurrences on the relative time frame, we were able to identify some interesting patterns. Aligned by the focal event "Start of 1st Quarter", the histograms of the occurrences of "Home:timeout" and "Away:timeout" show different patterns as in Fig. 12. The histograms clearly show the number of timeouts is related to the quarters. Timeouts were rarely called in the first and the last minutes of the first quarter (home:1/2, away:0/3). Also, timeouts were rarely called in the first and the last minutes of the third quarter (home:0/3, away:0/3). For the fourth quarter, A significant amount of timeouts were called in the last minute before the quarter ends (home:135, away:151), since the end of the fourth quarter is the end of regular game time (e.g. no overtime). Compared to the last minutes of the first and the third quarters, more timeouts were called in the last minute of the second quarter (home:34, away:42). The end of the second quarter is the end of the first half, where there is a long break before the second half starts for the team to adjust its game strategy. Therefore, the scores ending the first half are more important than that ending the first and third quarter.

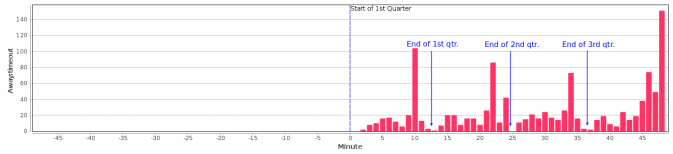
An interesting finding is that besides calling timeouts at the end of the second and fourth quarters, the home team called timeouts in the middle of each quarter (e.g. the 7th, 19th, 31st, and the 43rd minute). The away team called timeouts at a latter time (e.g. the 10th, 22nd, 34th, 46th).

Using the peak ratio, we found another interesting pattern with measure value 1.000 that many away teams called timeouts in the third and fourth minute following home timeouts. This finding is consistent with the previous finding where the peaks of home timeouts are in the middle of the quarters, while the the peaks of away timeouts are closer to the end of the quarters. As shown in Fig. 13, two peaks following the focal event. Note that we were unlikely to

http://www.nba.com



(a) Histogram of home timeout



(b) Histogram of away timeout

Figure 12: The histograms of "Home:timeout" and "Away:timeout" aligned by the first occurrence of "Start of 1st Quarter". Compared to the last minutes of the first and the third quarters, more timeouts were called in the last minutes of the ends of the second the fourth quarters. The peaks of home timeouts are in the middle of each quarter, while the peaks of away timeouts are closer to the end of the quarters.

find this pattern using occurrence ratio, since the measure value of occurrence ratio is 0.053.

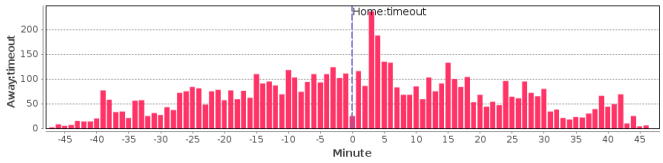


Figure 13: The histogram of "Away:timeout" aligned by "Home:timeout". The away team called timeouts more often at the third and fourth minute following the timeout of the the home team.

6.5 Medical records

The last data set are anonymized records of respiratory therapy and room transfer events. Respiratory therapy events include events such as nasal cannula, ventimask, BiPAP, and intubation. Room transfer events keep the record of when patients went to different rooms such as floor, IMC, ICU, and special rooms, which are not categorized as other rooms and may include operating rooms.

We started by aligning the records by all occurrences and used occurrence ratio measure. Most of the event pairs with measure values 1.000 and -1.000 are obvious patterns. For example, patients were admitted before they exited or patients were admitted first then sent to ICU. An interesting pattern was found with measure value -0.928 with the focal event "Intubation" and the related event "Special". In the histogram shown in Fig. 14, only a small percentage (4%) of "Special" occurred after "Intubation", which means that patients were mostly transfered into special rooms before they had intubation.

We continued to look at this pattern by aligning the records by the first occurrence of "Intubation". Patients were sent into special rooms mostly in one day before and after the first intubation. Surprisingly, more patients were sent into special rooms after the first intubation than before the first intubation (Fig. 15(a)). We then aligned the records by the third and fifth occurrences of intubation and found that the number of "Special" event following the alignment went down as we aligned by later occurrences (Fig. 15(b) and

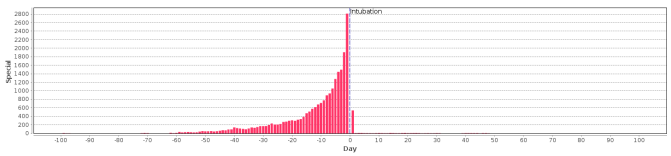
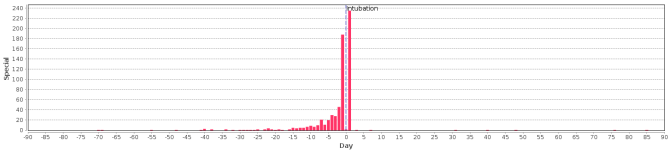
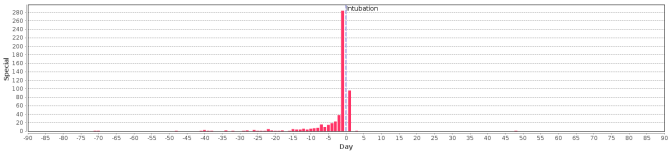


Figure 14: The histogram of "Special" aligned by all occurrences of "Intubation". Most patients were sent into special rooms before they had intubation.

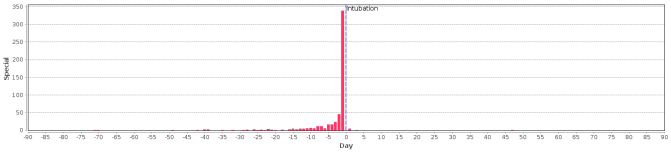
Fig. 15(c)). The percentages of the occurrences of related event following the first, third, and fifth occurrence of focal events are 36%, 17%, and 1%.



(a) Aligned by first occurrence of intubation



(b) Aligned by third occurrence of intubation



(c) Aligned by fifth occurrence of intubation

Figure 15: Histograms of the related event "Special" aligned by different occurrences of "Intubation". The proportion of occurrences of "Special" event following the alignment went down as we aligned by later occurrences.

With the previous pattern, we started comparing the histograms of "Special" with other room transfer events. The histogram of "Floor" are significantly different from other histograms, which may indicate that it is different from the other room transfer events. We found that "IMC" is the only event that has more occurrences following than preceding the first intubation in one day as in Fig. 16. The decreases of occurrences following a later occurrence of focal event in one day can also be found in IMC (50%, 40%, and 34%) and ICU(39%, 21%, and 11%).

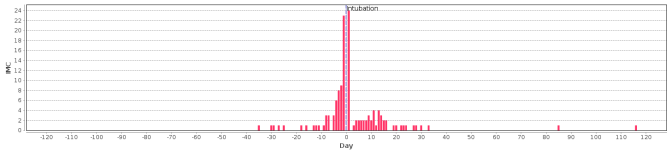


Figure 16: Histograms of the related event "IMC" aligned by the first occurrence of "Intubation". Similar to special rooms, more patients were sent into IMC after the first intubation.

7 Conclusion and Future Work

PairFinder shows the histogram of the occurrences of the related event aligned by the focal event. The histogram gives a summary about how the two events relate to one another. Since the number of candidate patterns can easily go beyond human capability, the ranking by measures provides an indication of how interesting the candidate pattern is. We performed four studies on a variety of data sets with PairFinder. With interactive inspection of the data sets, several interesting patterns between pairs of events were found.

The limitation of PairFinder is that we focus on pairs of events therefore the candidate patterns involve only the two events. We do not address the problem of finding more complex associations involving more than two events such as "event A triggers event B after 5 days if there is no event C which follows event A in one day". Complex associations have to be addressed more specifically with careful design.

The future work includes defining other measures to assess the interestingness of histograms. Also, the ability to define events hierarchically in PairFinder may be helpful to provide different abstraction of events.

Acknowledgement

We appreciate the partial support of the NIH/National Cancer Institute grant RC1-CA147489, Interactive Exploration of Temporal Patterns in Electronic Health Records.

References

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. L. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14. IEEE, IEEE Comput. Soc. Press, 1995.

[2] C. Bettini, X. S. Wang, S. Jajodia, and J. L. Lin. Discovering frequent event patterns with multiple granularities in time sequences. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):222–237, 1998.

[3] A. Campagna and R. Pagh. On finding frequent patterns in event sequences. In *Proceedings of IEEE International Conference on Data Mining*, pages 755–760. IEEE Press, 2010.

[4] J. Cheng. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90, 2002.

[5] D. M. Chickering. Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[6] O. Couturier, J. Rouillard, and V. Chevrin. An interactive approach to display large sets of association rules. In *Proceedings of the 2007 conference on Human interface: Part I*, pages 258–267, 2007.

[7] N. Elmqvist and P. Tsigas. Animated visualization of causal relations through growing 2D geometry. *Information Visualization*, 3(3):154–172, 2004.

[8] C. N. Glymour and G. F. Cooper. *Computation, Causation, and Discovery*. AAAI/MIT Press, 1999.

[9] N. Kadaba, P. Irani, and J. Leboe. Visualizing causal semantics using animations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1254–1261, 2007.

[10] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proceedings of the Third International Conference on Information and Knowledge Management*, pages 401–407, New York, New York, USA, 1994. ACM Press.

[11] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610–626, 2008.

[12] G. N. Norén, A. Bate, J. Hopstadius, K. Star, and I. R. Edwards. Temporal pattern discovery for trends and transient effects: its application to patient records. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 963–971, 2008.

[13] G. N. Norén, J. Hopstadius, and A. Bate. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. *Statistical methods in medical research*, pages 1–13, June 2011.

[14] G. K. Palshikar. Simple Algorithms for Peak Detection in Time-Series Simple Algorithms for Peak Detection in Time-Series. In *Proceedings of the 1st International Conference on Advanced Data Analysis, Business Analytics and Intelligence*, 2009.

[15] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge Univ Press, 2000.

[16] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224. Published by the IEEE Computer Society, IEEE Comput. Soc, 2001.

[17] S. Sahar. Interestingness via what is not interesting. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 332–336. ACM Press, 1999.

[18] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[19] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data mining*, volume 54 of *KDD '02*, pages 32–41. ACM Press, 2002.

[20] T. Wang, C. Plaisant, A. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 457–466. ACM, 2008.

[21] T. Wang, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Visual information seeking in multiple electronic health records: design recommendations and a process model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 46–55. ACM, 2010.

[22] B. G. Weber and M. Mateas. A data mining approach to strategy prediction. *2009 IEEE Symposium on Computational Intelligence and Games*, pages 140–147, Sept. 2009.

[23] G. M. Weiss and H. Hirsh. Learning to predict rare events in event sequences. *Artificial Intelligence*, pages 359–363, 1998.

[24] M. Zaki. Efficient enumeration of frequent sequences. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pages 68–75. ACM, 1998.