Mining Predisposing and Co-incident Factor by Using the Combination of Many Techniques

Suwimon Kooptiwoot

School of Information Technologies, University of Sydney
University of Sydney, NSW, 2006, Australia
suwimon@it.usyd.edu.au

Abstract. In this work we propose a new algorithm from the combination of many existing ideas consist of the reference event as proposed in [1], the event detection technique proposed in [2], the large fraction proposed in [3], the causal inference proposed in [4, 5]. We use all of these ideas to build up our algorithms to mine the predisposing factor and the co-incident factor of the reference event of interest. We apply our algorithm with OSS (Open Source Software) data set and show the result. We also test our algorithm with four synthetic data sets include noise up to 50 %. The results show that our algorithms can work well and tolerate to the noise data.

1 Introduction

Temporal mining is a data mining include time attribute in consideration. Time series data is the data set which include time attribute in the data. There are so many work and many methods and algorithms done in temporal mining. All are useful for mining the knowledge from time series data. We want to use the temporal mining techniques to mine the predisposing factor and the co-incident factor that make the number of the Download attribute change significantly and the rate of the number of the Download attribute change significantly in OSS data set.

2 The Problem

We get an OSS data set from http://sourceforge.net which is the world's largest Open Source software development website. There are 1,097,341 records, 41,540 projects in this data set. This data set consists of seventeen attributes include time attribute. The time attribute of each record in this data set is monthly. Each project in this data set is software. There are so many activities there. We are interested in thirteen attributes which indicate the number of the activities in this data set. The data of these thirteen attributes are all numeric. The value of the Download attribute is the number of the Download attribute. So the Download attribute is the indicator showing how popular the software is and show the successful of the development of the software. We are interested in the significant change of the number of the Download attribute.

Then we employ the idea of the event detection technique proposed by [2] to detect the event of the Download attribute. The event of our interest is the direction of the significant change of the data which can be up or down.

We want to find the predisposing factor and the co-incident factor of the Download event. We employ the same idea about the reference event as proposed in [1] which is the fixed event of interest and want to find the other events related to the reference event. So we call the Download attribute as the reference attribute and call the event of the Download attribute as the reference event.

The predisposing factor of the reference event can possible be the cause of the reference event or the cause of the other event which is the cause of the reference event, in somehow. And the co-incident factor of the reference event can possible be the effect of the reference event or the effect of the other event which is the effect of the reference event in somehow or be the event happening at the same time as the reference event happens or can be the result from the same cause of the reference event or just be the result from the other event which happens at the same time of the reference event happens. To make this concept clear, see the example as follow

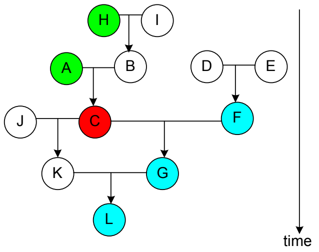

Figure1: the relationships among the events over time

If we have the event *A, B, C, D, E, F, G, H, I, J, K, L* and the relationships among them as shown below

$$H + I \rightarrow B$$
$$A + B \rightarrow C$$
$$D + E \rightarrow F$$
$$C + F \rightarrow G$$
$$J + C \rightarrow K$$
$$K + G \rightarrow L$$

That is *H* and *I* give B; *A* and *B* give *C*; *D* and *E* give *F*; *C* and *F* give *G; J* and *C* give *K; K* and *G* give *L*. But in our data set consists of only *A, C, F, G, H, L*. And the reference event is *C*. We can see that *H* and *A* happen before C, we may say that A is the cause of C and/or H is the cause of *C*. But in the real relationship as shown above, we know that *H* is not the cause of *C* directly. Or it is not because *A* and *H* give *C*. So we call *A* and *H* are the predisposing factors of *C*. And we find that *F* happens at the same time as *C* happens. And *G* and *L* happen after *C*. We call *F* and *G* and *L* as the co-incident factor of *C*. We can see from the relationship that *G* is the result from *C* + *F*. *L* is the result from *G* which is the result from *C*. And F is the co-factor of *C* that gives *G*. Only *G* is the result from *C* directly. *L* is the result from *G* which is the result from *C*.

We really want to find the exact relationships among these events. Unfortunately, no one can guarantee that our data set of consideration consists of all related factors or events. We can see from the diagram or the relationship shown in the example that the relationship among the events can be complex. And if we don't have all of the related events, we cannot find all of the real relationships. This fact the researchers in data mining community know well. We don't want our users misinterpret the results from our algorithms. So what we can do with the possible incomplete data set is mining the predisposing factor and the co-incident factor of the reference event. Then the users can further consider these factors and collect more data which related to the reference event and explore more in depth by themselves on the expert ways in their specific fields.

The main idea in this part is the predisposing factor can possible be the cause of the reference event and the co-incident factor can possible be the effect of the reference event. So we employ the same idea as proposed in [4, 5] that is the cause happens before the effect. The effect happens after the cause. We call the time point when the reference event happens as the current time point. We call the time point before the current time point as the previous time point. And we call the time point after the current time point as the post time point. Then we define the predisposing factor of the reference event as the event which happens at the previous time point. And we define the co-incident factor of the reference event as the event happens at the current time point and/or the post time point.

3 Basic Definitions and Framework

We define the events as the significant slope. The method to interpret the result of the event is selecting the large fraction of the positive slope and the negative slope at each time point. If it is at the previous time point that means it is the predisposing factor. If it is at the current time point and/or the post time point that means it is a co-incident factor.

Definition1: A time series data set is a set of records *r* such that each record contains a set of attributes and a time attribute. The value of time attribute is the point of time on time scale such as month, year.

$r_j = \{ a_1, a_2, a_3, \ldots, a_m, t_j \}$

where

r_j is the j^{th} record in data set

Definition 2: There are two types of the attribute in time series data set. Attribute that depends on time is dynamic attribute (Ω) , other wise, it is static attribute (*S*).

Definition 3: Time point (t_i) is the time point on time scale.

Definition 4: Time interval is the range of time between two time points [t_1, t_2]. We may refer to the end time point of interval (t_2).

Definition 5: An attribute function is a function of time whose elements are extracted from the value of attribute i in the records, and is denoted as a function in time, $a_i(t_x)$

$a_i(t_x) = a_i \in r_j$

where

a_i attribute i;

t_x time stamp associated with this record

Definition 6: A feature is defined on a time interval $[t_1, t_2]$, if some attribute function $a_i(t)$ can be approximated to another function $\Phi(t)$ in time , for example,

$a_i(t) \approx \Phi(t)$, $\forall t \in [t_1, t_2]$

We say that Φ and its parameters are features of $a_i(t)$ in that interval $[t_1, t_2]$.

If $\Phi(t) = \propto_i t + \beta_i$ in some intervals, we can say that in the interval, the function $a_i(t)$ has a slope of $\propto_i$ where slope is a feature extracted from $a_i(t)$ in that interval

Definition 7: Slope ($\propto_i$) is the change of value of a dynamic attribute (a_i) between two adjacent time points.

$\propto_i = (a_i t_x - a_i t_{x-1}) / t_x - t_{x-1}$

where

$a_i t_x$ is the value of a_i at the time point t_x

$a_i t_{x-1}$ is the value of a_i at the time point t_{x-1}

Definition 8: Slope direction $d(\propto_i)$ is the direction of slope.

If $\propto_i > 0$, we say $d_\propto = 1$

If $\propto_i < 0$, we say $d_\propto = -1$

If $\propto_i \cong 0$, we say $d_\propto = 0$

Definition 9: A significant slope threshold (δI) is the significant slope level specified by user.

Definition 10*:* Reference attribute (a_t) is the attribute of interest. We want to find the relationship between the reference attribute and the other dynamic attributes in the data set.

Definition 11: An event ($E1$) is detected if $\propto_i \geqslant \delta I$

Definition 12: Current time point (t_c) is the time point at which reference variable's event is detected.

Definition 13: Previous time point (t_{c-1}) is the previous adjacent time point of t_c

Definition 14: Post time point (t_{c+1}) is the post adjacent time point of t_c

Proposition 1: Predisposing factor of a_t denoted as $PE1a_t$ is an ordered pair $(a_i, d_\propto)$ when $a_i \in \Omega$

If ${}^{np}a_i t_{c-1} > {}^{nn}a_i t_{c-1}$, then $PE1a_t \approx (a_i , 1)$

If ${}^{np}a_i t_{c-1} < {}^{nn}a_i t_{c-1}$, then $PE1a_t \approx (a_i , -1)$

where

${}^{np}a_i t_{c-1}$ is the number of positive slope of $E1$ of a_i at t_{c-1}

${}^{nn}a_i t_{c-1}$ is the number of negative slope of $E1$of a_i at t_{c-1}

Proposition 2: Co-incident factor of a_t denoted as $CE1a_t$ is an ordered pair $(a_i, d_\propto)$ when $a_i \in \Omega$

If (($^{np}a_i\, t_c > {}^{nn}a_i\, t_c$) $\vee$ ($^{np}a_i\, t_{c+1} > {}^{nn}a_i\, t_{c+1}$)) , then $CE1a_t \approx (a_i , 1)$

If (($^{np}a_i\, t_c < {}^{nn}a_i\, t_c$) $\vee$ ($^{np}a_i\, t_{c+1} < {}^{nn}a_i\, t_{c+1}$)) , then $CE1a_t \approx (a_i , -1)$

where

$^{np}a_i\, t_c$ is the number of positive slope of $E1$ of a_i at t_c

$^{nn}a_i\, t_c$ is the number of negative slope of $E1$ of a_i at t_c

$^{np}a_i\, t_{c+1}$ is the number of positive slope of $E1$ of a_i at t_{c+1}

$^{nn}a_i\, t_{c+1}$ is the number of negative slope of $E1$ of a_i at t_{c+1}

4 Algorithm

Input: The data set which consists of numerical dynamic attributes. Sort this data set to ascending order by time, a_t, δI

Output: $^{np}a_i\, t_{c-1}$, $^{nn}a_i\, t_{c-1}$, $^{np}a_i\, t_c$, $^{nn}a_i\, t_c$, $^{np}a_i\, t_{c+1}$, $^{nn}a_i\, t_{c+1}$, $PE1a_t$, $CE1a_t$

Method:

```
For all $a_i$
    For all time interval $[t_x , t_{x+1}]$
        Calculate $\alpha_i$
            For $a_t$
                If $\alpha_t \geqslant \delta I$
                    Set that time point as $t_c$
        Group record of 3 time points $t_{c-1}\ t_c\ t_{c+1}$
Count $^{np}a_i\, t_{c-1}$, $^{nn}a_i\, t_{c-1}$ , $^{np}a_i\, t_c$, $^{nn}a_i\, t_c$ , $^{np}a_i\, t_{c+1}$, $^{nn}a_i\, t_{c+1}$
// interpret the result
If $^{np}a_i\, t_{c-1} > {}^{nn}a_i\, t_{c-1}$ , then $PE1a_t \approx (a_i , 1)$
If $^{np}a_i\, t_{c-1} < {}^{nn}a_i\, t_{c-1}$ , then $PE1a_t \approx (a_i , -1)$
If $^{np}a_i\, t_c > {}^{nn}a_i\, t_c$ , then $CE1a_t \approx (a_i , 1)$
If $^{np}a_i\, t_c < {}^{nn}a_i\, t_c$ , then $CE1a_t \approx (a_i , -1)$
If $^{np}a_i\, t_{c+1} > {}^{nn}a_i\, t_{c+1}$ , then $CE1a_t \approx (a_i , 1)$
If $^{np}a_i\, t_{c+1} < {}^{nn}a_i\, t_{c+1}$ , then $CE1a_t \approx (a_i , -1)$
```

We employ the method proposed by [3] that is using the large fraction to judge the data change direction of the attribute of consideration. Using the combination of the ideas mentioned above, we can find the predisposing factor and the co-incident factor of the reference event of interest.

We don't use the threshold to find the event of the other attributes because of the idea of the degree of important [6]. For example, the effects of the different kind of chilly on the food are different. Only small amount of the very hot chilly make our food very hot. Very much of sweet chilly make our food not so spicy. We see that the same amount of the different kind of chilly make the different level of the spicy of our food. The very hot chilly has the degree of important on our food higher than the sweet chilly. So we do not specify the threshold of the event of the other attributes to be considered as the predisposing factor or the co-incident factor. We illustrate how this algorithm work, for example, the graph of the data is shown as follow

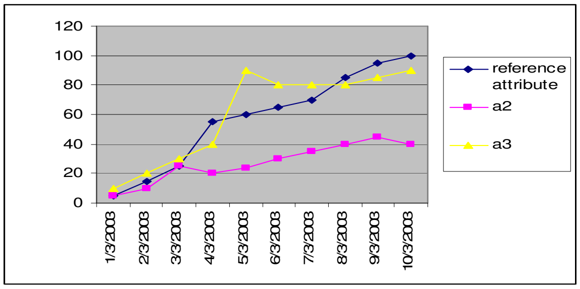

Figure 2. The data in the graph

We calculate slope value showing how much the data change and the direction of the data change per time unit.

Then we set the data change threshold as 15. We use this threshold to find the reference event. We find the reference event at the time 4/03. Then we mark this time point as the current time point. Next we look at the previous time point, 3/03, for the predisposing factor, we find that a2 with the positive direction and a3 with positive direction are the predisposing factor of the reference event. Then we look at the current time point and the post time point, 4/03 and 5/03, for the co-incident factor, we find that at the current time point, a2 with the negative direction and a3 with the positive direction are the co-incident factor. And at the post time point, a2 with the positive direction and a3 with the positive direction are the co-incident factor. We can summarize the result in the pattern table as shown below

Table 1. The direction of each attribute at each time point

	Previous time point	Current time point	Post time point
a2	up	down	up
a3	up	up	up

5 Experiments

We apply our methods with the OSS data set. We consider only thirteen attributes which are Download, Page views, Bugs0, Bugs1, Support0, Support1, Patches0, Patches1, Tracker0, Tracker1, Tasks0, Tasks1, CVS. We use the Download attribute as the reference attribute and consider the rest of all of the other attributes in finding the predisposing factor and the co-incident factor of the Download event for both the event type I and the event type II. For the event type I, we separate the case of the positive direction and negative direction of the significant data change of the Download attribute. For the event type II, we consider both not considering the direction and considering the direction of the significant rate of the data change of the Download attribute. We set the data change threshold of the Download attribute as 50. The results are in Table 2 and Table 3.

In case slope of the Download is positive

Table 2. The direction of the data change of each attribute at each time point

	previous	current	post
P/V	up	up	down
Bugs0	up	up	down
Bugs1	up	up	down
Support0	up	up	down
Support1	up	up	down
Patches0	up	up	down
Patches1	up	up	down
Tracker0	up	up	down
Tracker1	up	up	down
Tasks0	down	down	down
Tasks1	up	up	down
CVS	up	up	down

In case slope of the Download is negative

Table 3. The direction of the data change of each attribute at each time point

	previous	current	post
P/V	up	down	down
Bugs0	up	down	down
Bugs1	up	down	up
Support0	up	down	down
Support1	up	down	up
Patches0	up	down	up
Patches1	up	down	up
Tracker0	up	down	down
Tracker1	up	down	up
Tasks0	down	down	down
Tasks1	down	down	down
CVS	down	down	down

6 Performance

Our methods consume time to find the predisposing factor and the co-incident factor of the reference event just in $O(n)$ where n is the number of the total records. The most of time consuming is the time for calculating the slope (the data change) of every two adjacent time points of the same project which take time $O(n)$. And we have to spend time to select the reference event by using the threshold which takes time *O*(*n*). We have to spend time to group records around the reference event (at the previous time point(s), the current time point and the post time point(s)) which is *O*(*n*). And the time for counting the number of the event of the other attributes at each time point around the current time point is *O*(*n*). The time in overall process can be approximate to *O*(*n*), which is not exponential. So our methods are good enough to apply in the big real life data set.

From our applying with OSS data set, the machine we use in our experiments is PC PentiumIV 1.6 GHz, RAM 1 GB. The operating system is MS WindowsXP Professional. We implement this algorithm in Perl 5.8 on command line. The data set test has 1,097,341 records, 41, 540 projects with total 17 attributes. The number of attributes of consideration is thirteen attributes. The size of this data set is about 48 MB.

We want to see that our program consume running time in linear scale with the size of the data or not. Then we test with the different number of records in each file and run each file at a time. The result is shown and Graph 2.

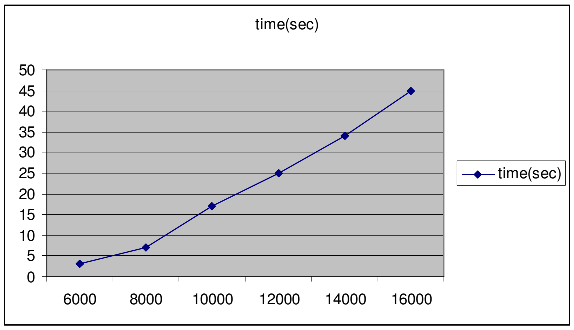

Figure 3. Running time (in seconds) and the number of records to be run at a time

From this result confirm us that our algorithms consume execution time in linear scale with the number of records.

7 Accuracy test with synthetic data sets

From the concept, the predisposing factor is the event happening before the reference event, and the co-incident factor is the event happening at the same time or after the reference event. Some predisposing factors can act as the catalyst in the chemical reaction. The catalyst can effect on the reaction in two ways, the first is activating the reaction in case the reaction can happen itself but slowly, the second is initializing the reaction in case the reaction cannot happen itself. The point is, not all of the reaction need the catalyst. We look at the reference event as the chemical reaction. So we synthesize data to be some reference events have catalyst involve and some not.

We synthesize 4 data sets as follow

1. Correct complete data set
2. Put 5 % of noise in the first data set
3. Put 20 % of noise in the first data set
4. Put 50 % of noise in the first data set

We set the data change threshold as 10. The result is almost all of four data sets correct, except only at the third data set with 20 % of noise, there is only one point in the result different from the others, that is, the catalyst at the current point changes to be positive slope in stead of steady.

8 Conclusion

The combination of the existing methods to be our new algorithm can be used to mine the predisposing factor and the co-incident factor of the reference event very well. As seen in our experiments, our algorithm can be applied with both the synthetic data set and the real life data set. The performance of our algorithm is also good. They consume execution time just in linear time scale and also tolerate to the noise data.

9 Discussion

The threshold is the indicator to select the event which is the significant data change of the attribute of consideration. When we use the different threshold in the detecting of the event, the results can be different. So setting the threshold of the data change has to be well justified. It can be justified by looking at the data and observing the characteristic of the attributes of interest. The users have to realize that the results they get can be different depending on the threshold setting. The threshold reflects the degree of important of the predisposing factor and the co-incident factor of the reference event to the reference event. If the degree of important of an attribute is very high, just little change of the data of that attribute can make the data of the reference attribute change very much. From this fact, if we set the threshold to be high, we will not be able to detect the event of the attribute or the factor which has the high degree of important to the reference event. On the other hand, if the degree of important of the attribute of consideration is low, that means the data change of this attribute has to be high to be able to make effect on the reference event. If we set the data change threshold of this attribute low, we will select the event of the little change of the data which actually does not effect on the reference event to be in our consideration. This make the result can be wrong. So the setting the data change or the rate of the data change threshold is very sensitive to the accuracy of the result.

References

1. Bettini, C., et al., *Discovering Frequent Event Patterns with Multiple Granularities in Time Sequences.* IEEE Transactions on Knowledge and Data Engineering, 1998. **10**(2).
2. Guralnik, V. and J. Srivastava. *Event Detection from Time Series Data*. in *KDD-99*. 1999. San Diego, CA USA.
3. Mannila, H., H. Toivonen, and A.I. Verkamo, *Discovery of frequent episodes in event sequences.* Data Mining and Knowledge Discovery, 1997. **1**(3): p. 258-289.
4. Blum, R.L., *Discovery, Confirmation and Interpretation of Causal Relationships from a Large Time-Oriented Clinical Databases: The Rx Project.* Computers and Biomedical Research, 1982. **15**(2): p. 164-187.
5. Blum, R.L., *Discovery and Representation of Causal Relationships from a Large Time-Oriented Clinical Databases: The Rx Project.* Lecture Notes in Medical Informatics, 1982. **19**.
6. Salam, M.A. *Quasi Fuzzy Paths in Semantic Networks*. in *Proceedings 10th IEEE International Conference on Fuzzy Systems*. 2001. Melbourne, Australia.
7. Freemantle, M., *Chemistry in Action*. second edition ed. 1995, Great Britain: MACMILLAN PRESS.
8. Robinson, W.R., J.D. Odom, and J. Henry F. Holtzclaw, *Essentials of General Chemistry*. Tenth edition ed. 1997, USA: Houghton Mifflin Company.
9. Snyder, C.H., *The Extraordinary Chemistry of Ordinary Things*. third edition ed. 1998, USA: John Wiley & Sons, Inc.
10. Liska, K. and L.T. Pryde, *Introductory Chemistry for Health Professionals*. 1984, USA: Macmillan Publishing Company.
11. Harrison, R.M., et al., *Introductory chemistry for the environmental sciences*. Cambridge Environmental Chemistry Series, ed. P.G.C. Camphell, J.N. Galloway, and R.M. Harrison. 1991, Cambridge: Cambridge University Press.