Mining Predisposing and Co-incident Factor by Using the Fact Seen in the Chemical Reaction

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 the new algorithms from the combination of many existing ideas consist of the reference event as proposed in [1], the event detection technique proposed in [2], the causal inference proposed in [4, 5] and the new idea about the character of the catalyst seen in the chemical reaction. 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 algorithms with Open Source Software (OSS) data set and show the result. We also test our algorithms 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 significant rate of the data change which can be acceleration or deceleration.

We want to find the predisposing factor and the co-incident factor of the Download events. 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 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

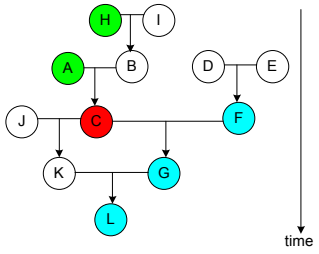

Figure 1: the relationships among the events over time

If we have the event *A, B, C, D, E, F, G, H, I, J, K, L* and the relationships among them are *H + I* → *B, A + B* → *C, D + E* → *F, C + F* → *G, J + C* → *K, K + G* → *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 event as the significant slope rate. We use the analogy of the chemical reaction to interpret the predisposing and co-incident factors. The point is the amount of the reactants, catalyst increase significantly before the reaction and then decrease significantly at the reaction process time. And the amount of the products increases significantly after the reaction. We detect two previous adjacent time points and two post adjacent time points in order to make sure that we cover all of the reactants and/or the catalysts and the products. We then judge if the number of significant changes at either of the previous time points, then we call it the predisposing factor. If it happens at either of the post time points, we call it the 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 $\Phi(t) = \alpha_i t + \beta_i$ in some intervals, we can say that in the interval, the function $a_i(t)$ has a slope of α_i where slope is a feature extracted from $a_i(t)$ in that interval

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

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

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

Definition 10: Previous time point (t_{c-1}) is the previous adjacent time point of t_c

Definition 11: Second previous time point (t_{c-2}) is the previous adjacent time point of t_{c-1}

Definition 12: Post time point (t_{c+1}) is the post adjacent time point of t_c

Definition 13: Second post time point (t_{c+2}) is the post adjacent time point of t_{c+1}

Definition 14: Slope rate (θ) is the relative slope between two adjacent time intervals

$\theta = (\alpha_{i+1} - \alpha_i) / \alpha_i$

where

α_x is the slope value at time interval $[t_{i-1}, t_i]$

α_{x+1} is the slope value at time interval $[t_i, t_{i+1}]$

Definition 15: Slope rate direction (d_θ) is the direction of θ

If $\theta > 0$, we say $d_\theta = 1$ or accelerating

If $\theta < 0$, we say $d_\theta = -1$ or decelerating

If $\theta \cong 0$, we say $d_\theta = 0$ or steady

Definition 16: A significant slope rate threshold (δII) is the significant slope rate level specified by user.

Definition 20: An event (*E2*) is detected if $\theta \geqslant \delta II$

Proposition 1: The predisposing factor of a_t denoted as $PE2a_t$ without considering d_θ is a_i

if $(({}^n a_i t_{c-1} \geqslant {}^n a_i t_c) \vee ({}^n a_i t_{c-2} \geqslant {}^n a_i t_c))$

where

${}^n a_i t_c$ is the number of *E2* of a_i at t_c

${}^n a_i t_{c-1}$ is the number of *E2* of a_i at t_{c-1}

${}^{n}a_i\ t_{c-2}$ is the number of *E2* of a_i at t_{c-2}

Proposition 2: The co-incident factor of a_t denoted as $CE2a_t$ without considering d_θ is a_i

if $(({}^{n}a_i\ t_{c+1} \geqslant {}^{n}a_i\ t_c) \vee ({}^{n}a_i\ t_{c+2} \geqslant {}^{n}a_i\ t_c))$

where

${}^{n}a_i\ t_c$ is the number of *E2* of a_i at t_c

${}^{n}a_i\ t_{c+1}$ is the number of *E2* of a_i at t_{c+1}

${}^{n}a_i\ t_{c+2}$ is the number of *E2* of a_i at t_{c+2}

Proposition 3: The predisposing factor of a_t with considering d_θ of reference's event denoted as $PE2a_t\, d_\theta a_t$ is an ordered pair $(a_i , d_\theta a_t)$ when $a_i \in \Omega$

where

$d_\theta a_t$ is slope rate direction of a_t

Proposition 3.1 : If $(({}^{ntp}a_i\ t_{c-1} \geqslant {}^{ntp}a_i\ t_c) \vee ({}^{ntp}a_i\ t_{c-2} \geqslant {}^{ntp}a_i\ t_c))$, then $PE2a_t\, d_\theta a_t \approx (a_i , 1)$

where

${}^{ntp}a_i\ t_c$ is the number of *E2* of a_i at t_c for which $d_\theta a_t$ is accelerating

${}^{ntp}a_i\ t_{c-1}$ is the number of *E2* of a_i at t_{c-1} for which $d_\theta a_t$ is accelerating

${}^{ntp}a_i\ t_{c-2}$ is the number of *E2* of a_i at t_{c-2} for which $d_\theta a_t$ is accelerating

Proposition 3.2 : If $(({}^{ntn}a_i\ t_{c-1} \geqslant {}^{ntn}a_i\ t_c) \vee ({}^{ntn}a_i\ t_{c-2} \geqslant {}^{ntn}a_i\ t_c))$, then $PE2a_t\, d_\theta a_t \approx (a_i , -1)$

where

${}^{ntn}a_i\ t_c$ is the number of *E2* of a_i at t_c for which $d_\theta a_t$ is decelerating

${}^{ntn}a_i\ t_{c-1}$ is the number of *E2* of a_i at t_{c-1} for which $d_\theta a_t$ is decelerating

${}^{ntn}a_i\ t_{c-2}$ is the number of *E2* of a_i at t_{c-2} for which $d_\theta a_t$ is decelerating

Proposition 4: The co-incident factor of a_t with considering $d_\theta a_t$ denoted as $CE2a_t\, d_\theta a_t$ is an ordered pair $(a_i , d_\theta a_t)$ when $a_i \in \Omega$

Proposition 4.1 : If $(({}^{ntp}a_i\ t_{c+1} \geqslant {}^{ntp}a_i\ t_c) \vee ({}^{ntp}a_i\ t_{c+2} \geqslant {}^{ntp}a_i\ t_c))$, then $CE2a_t\, d_\theta a_t \approx (a_i , 1)$

where

${}^{ntp}a_i\ t_c$ is the number of *E2* of a_i at t_c for which $d_\theta a_t$ is accelerating

${}^{ntp}a_i\ t_{c+1}$ is the number of *E2* of a_i at t_{c+1} for which $d_\theta a_t$ is accelerating

${}^{ntp}a_i\ t_{c+2}$ is the number of *E2* of a_i at t_{c+2} for which $d_\theta a_t$ is accelerating

Proposition 4.2 : If $(({}^{ntn}a_i\ t_{c+1} \geqslant {}^{ntn}a_i\ t_c) \vee ({}^{ntn}a_i\ t_{c+2} \geqslant {}^{ntn}a_i\ t_c))$, then $CE2a_t\, d_\theta a_t \approx (a_i , -1)$

where

${}^{ntn}a_i\ t_c$ is the number of *E2* of a_i at t_c for which $d_\theta a_t$ is decelerating

${}^{ntn}a_i\ t_{c+1}$ is the number of *E2* of a_i at t_{c+1} for which $d_\theta a_t$ is decelerating

${}^{ntn}a_i\ t_{c+2}$ is the number of *E2* of a_i at t_{c+2} for which $d_\theta a_t$ is decelerating

4 Algorithms

We deal with the rate of the data change, and we see the fact about the catalyst in the chemical reaction, that is, the catalyst can activate the rate of the chemical reaction to make it happen faster. So we look at the character of the catalyst in the chemical reaction in [7-11]. It is not necessary to have the catalyst in the chemical reaction. Not all of the chemical reaction has the catalyst. But once the catalyst is in the chemical reaction, it can activate the rate of the chemical reaction to make it happen faster. Some events act as the catalyst. So we use the idea taken from the fact of the catalyst in the chemical reaction. That is, its amount at the time before the reaction happen is higher than its amount at the time of the reaction happen. And its amount at the time after the reaction happen is higher than its amount at the time of the reaction happen. So we compare the number of the event of the attribute of consideration at the previous time point with its own number at the current time point. And we also compare the number of the event of the attribute of consideration at the post time point with its own number at the current time point.

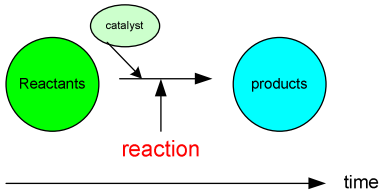

Figure 2. The chemical reaction include the catalyst

The combination of many ideas mentioned above include the new idea taken from the fact seen in the chemical reaction as explained can be used to find the predisposing factor and the co-incident factor of the reference event.

Now we present a few algorithms appropriate for the type of data we have. Each algorithm is tailored for the type of patterns we wish to explore.

4.1 Algorithm EII without considering d_θ

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

Output: ${}^n a_i\, t_{c-2}$, ${}^n a_i\, t_{c-1}$, ${}^n a_i\, t_c$, ${}^n a_i\, t_{c+1}$, ${}^n a_i\, t_{c+2}$, $PE2a_t$, $CE2a_t$

Method:

/ Basic part*

For all a_i

For all time interval $[t_x , t_{x+1}]$

Calculate α_i

For all two adjacent time intervals

Calculate θ

For a_t

If $\alpha_t \geqslant \delta II$

Set that time point as t_c

Group record of 5 time points $t_{c-2}\ t_{c-1}\ t_c\ t_{c+1}\ t_{c+2}$

**/ End of Basic part*

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 $(({}^{n}a_i\ t_{c-1} \geqslant {}^{n}a_i\ t_c) \vee ({}^{n}a_i\ t_{c-2} \geqslant {}^{n}a_i\ t_c))$, then a_i is $PE2a_t$

If $(({}^{n}a_i\ t_{c+1} \geqslant {}^{n}a_i\ t_c) \vee ({}^{n}a_i\ t_{c+2} \geqslant {}^{n}a_i\ t_c))$, then a_i is $CE2a_t$

4.2 Algorithm EII with considering $d_{\theta}a_t$

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

Output: ${}^{ntp}a_i\ t_{c-2}$, ${}^{ntp}a_i\ t_{c-1}$, ${}^{ntp}a_i\ t_c$, ${}^{ntp}a_i\ t_{c+1}$, ${}^{ntp}a_i\ t_{c+2}$, ${}^{ntn}a_i\ t_{c-2}$, ${}^{ntn}a_i\ t_{c-1}$, ${}^{ntn}a_i\ t_c$, ${}^{ntn}a_i\ t_{c+1}$, ${}^{ntn}a_i\ t_{c+2}$, $PE2a_t\ d_{\theta}\ a_t$, $CE2a_t\ d_{\theta}\ a_t$

Method:

/ Basic part */*

Count ${}^{ntp}a_i\ t_{c-2}$, ${}^{ntp}a_i\ t_{c-1}$, ${}^{ntp}a_i\ t_c$, ${}^{ntp}a_i\ t_{c+1}$, ${}^{ntp}a_i\ t_{c+2}$, ${}^{ntn}a_i\ t_{c-2}$, ${}^{ntn}a_i\ t_{c-1}$, ${}^{ntn}a_i\ t_c$, ${}^{ntn}a_i\ t_{c+1}$, ${}^{ntn}a_i\ t_{c+2}$

// interpret the result

If $(({}^{ntp}a_i\ t_{c-1} \geqslant {}^{ntp}a_i\ t_c) \vee ({}^{ntp}a_i\ t_{c-2} \geqslant {}^{ntp}a_i\ t_c))$, then a_i is $PE2a_t\ d_{\theta}a_t$ in acceleration.

If $(({}^{ntn}a_i\ t_{c-1} \geqslant {}^{ntn}a_i\ t_c) \vee ({}^{ntn}a_i\ t_{c-2} \geqslant {}^{ntn}a_i\ t_c))$, then a_i is $PE2a_t\ d_{\theta}a_t$ in deceleration.

If $(({}^{ntp}a_i\ t_{c+1} \geqslant {}^{ntp}a_i\ t_c) \vee ({}^{ntp}a_i\ t_{c+2} \geqslant {}^{ntp}a_i\ t_c))$, then a_i is $CE2a_t\ d_{\theta}a_t$ in acceleration.

If $(({}^{ntn}a_i\ t_{c+1} \geqslant {}^{ntn}a_i\ t_c) \vee ({}^{ntn}a_i\ t_{c+2} \geqslant {}^{ntn}a_i\ t_c))$, then a_i is $CE2a_t\ d_{\theta}a_t$ in deceleration.

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. We consider both not considering the direction and considering the direction of the significant rate of the data change of the Download attribute. The results are

We set the rate of the data change threshold of the Download attribute and the rest of all of the other attributes as 1.5.

In case without considering the slope rate direction of the Download attribute

Predisposing Factor(s): Tasks0, Tasks1, CVS

Co-incident Factor(s): Support0, Support1, Patches0, Patches1

In case considering the slope rate direction of the Download attribute

The acceleration of the Download attribute

Predisposing Factor(s): none
Co-incident Factor(s): Bugs0, Bugs1, Support0, Support1, Patches0, Patches1, Tracker0, Tracker1

The deceleration of the Download attribute

Predisposing Factor(s): Bugs0, Bugs1, Support0, Support1, Patches0, Tracker0, Tasks0, Tasks1, CVS
Co-incident Factor(s): Support1

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) and the slope rate (the rate of 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 these algorithms 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.

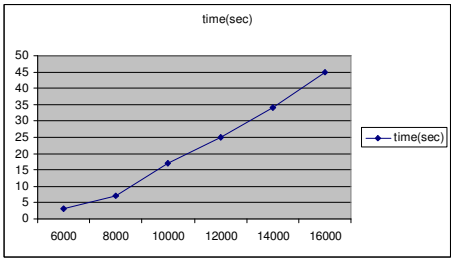

Figure 3. Running time (in seconds) and the number of records to be run at a time

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

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 rate of the data change threshold as 10. The results of all of three types of the event with four of the synthetic data sets which include noise up to 50 % are all correct at 100 %.

From the result, we see that our algorithms tolerate to the noise data.

8 Conclusion

The combination of the existing methods and the new idea from the fact seen in the chemical reaction to be our new algorithms 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 algorithms can be applied with both the synthetic data set and the real life data set. The performance of our algorithms 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 rate 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 threshold of the rate 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 or the rate of 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.