Mining the Change of Events in Environmental Scanning for Decision Support

Mengjung Shih[1], Duenren Liu[1], Churnjung Liau[2], Chinhui Lai[1]

[1]Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan, China
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan, China
[1]{mj_shih, dlliu, chlai}@iim.nctu.edu.tw, [2]liaucj@iis.sinica.edu.tw

ABSTRACT

An organization's environment is increasingly complex. Business demand on environmental scanning has significantly increased in recent years due to an attempt to assist management in planning an organization's strategies and responses. The conventional technique for environmental scanning is event detection from text documents such as news stories. Event detection methods recognize events while they neglect to discover the changes of events. This work develops an event change detection (ECD) approach that combines association rule mining and change mining techniques. Detecting changes of events aids managers in making fast responses to the change of external environments. Association rule mining is employed to discover the subject behaviors of events from news stories. Changes of events are identified by comparing the subject behaviors of events from different time periods. The discovered event changes can provide effective decision support for decision makers to capture environmental information in a timely manner and make adequate decisions.

Keywords: Environmental Scanning, Change Mining, Association Rule Mining, Event Tracking, Event Detection

1. INTRODUCTION

With the rapid growth of Internet, the business environment today has become more complex and changeable. Environmental scanning has received considerable attention in business in recent years due to its benefits of assisting management in planning the organization's future course of action [4].

News is the best information source for businesses to get environment information (event) because of its comprehensive content. Event detection is the major research field in news, and it plays an important role for supporting environmental scanning [14]. Event detection identifies streams of news stories that discuss new events [1][2][16][17]. Organization managers can be aware of new events in their environment via event detection technologies. Event detection methods are concerned only with recognizing new events and neglect to discover the changes between these news stories. But in fact, changes of environment occur not only through new events but also in existing events.

The purpose of environmental scanning is not only to identify changes of the content of news stories that describe the same or different events, but also to aid business managers to be sensitive to the external environment resulting in a fast response to the environment. Detecting event changes is a critical work for businesses.

An event change is the change of event trends in two time periods. For example, the trend of telecommunications industries in 2001 was to provide educational services to customers, while the trend of telecommunications industries in 2002 was to provide recreational services. The change of service types is named event change. These trends are discovered from news stories of events- "Services provided in telecommunications industries (telecom services)" in 2001 and 2002 respectively.

To capture the changes of events, we must first determine the subject behaviors of an event. A subject behavior of an event is an incident that is described in most news stories, and it can be characterized by the relationships of 4W properties (When, Who, Where and What). The relationships between 4Ws may change with time, but these changes will be important information for decision making for business managers. For instance, if telecommunication companies were reported to provide recreational services (What property) in most news stories of the event "telecom services", "telecom services" can be viewed as a subject behavior of the event. If subject behaviors of different time periods are different, event change occurs.

In this research, association rule mining is employed to identify the subject behaviors of events from news stories.

Information needs usually vary with time, situation and people. Business managers may require different levels of information in developing different strategies. A decision maker may not only need to know the business operations of individual companies but also the operations of industries. To meet all possible information needs in environmental scanning, it is necessary to construct a concept hierarchy according to the content of news stories.

Motivated by the needs for capturing event changes, the goal of this research is to develop an event change identification (ECI) technique. The proposed technique combines the change mining approach and concept hierarchy to provide useful environment changes to enhance environmental scanning.

The remainder of this paper is organized as follows. Section 2 reviews literatures relevant to this research, including environmental scanning, event tracking and detection, association rule mining and change mining technologies. Section 3 introduces our event changes identification technique. In Section 4, we describe methods to detect event changes. Finally, the contributions of this study are summarized in Section 5.

2. LITERATURE REVIEW

This section reviews literatures relevant to this research, including environmental scanning, event tracking and detection, association rule mining, and changes mining technologies.

2.1. Environmental Scanning

To adapt to the environment and subsequently develop effective responses to secure or improve the business' position in the future, environmental scanning is important [4][7].

Environmental scanning acquires and uses information about a business's external environment, which is used by business managers to make decisions and plans [4]. Consequently, it is essential to scan the external business environment to improve a business's future.

2.2. Event Tracking and Detection

Event tracking and detection techniques are proposed to support people detecting new events and tracking subsequent news stories that discuss previous events.
Event tracking starts from a set of pre-classified news stories, and searches out all subsequent stories that discuss the same event [1][17]. The goal of event tracking is to locate follow-up news stories of a manager's event of interest.

Event detection is the major research field in news, and it plays an important role for supporting environmental scanning [14]. The goal of event detection is to identify streams of news stories that discuss new events [1][2][16][17]. Organization managers can become aware of new events in their environment via event detection technologies. But changes of environment occur not only during new events but also during existing events. Business administrators need to be notified before an event takes place, not just when it happens.

2.3. Association Rule Mining

Association rule mining searches for interesting association or correlation relationships among items in a large dataset [6]. There are two measures to represent whether a mined rule meets the criteria of regularity: support and confidence [6][15].
This work proposes an event change identification technique to detect event changes. We apply association rule mining on news data to detect the regulation of the news.

The format between news data and transaction data is very different. Transaction data is structured, and the scope of attributes and values is often fixed. In contrast to transaction data, news data is unstructured, and the scope of attributes and values often changes. For example, the protagonist of the same event may change.
Therefore, mining association rules from news data needs to address these difficulties. In this study, we used event property selection and extraction and concept hierarchy to solve these problems.

2.4. Change Mining

The objective of change mining is to discover changes of data from different time periods. The researches of change mining can be classified into two main groups:
1. Decision Tree Model: This method first constructs decision trees from different datasets. It obtains differences via comparisons with two decision trees [10]. Through graphic presentation, the user can discover the differentiae and understand the data.

2. Association Rule: This method obtains changes via comparisons with association rules mined from a different dataset [11][12]. The types of changes defined by [5][8][9][12] are emerging patterns (the emerging pattern concept captures significant changes and differences between datasets.), unexpected changes (can be found in newly discovered association rules with consequence/conditional parts different from previous rules), added rules (is a newly appearing rule in the present which could not be found in the past) and perished rules (is a rule that can only be found in the past). Association rule change mining is used to discover changes in customer behavior.

In this study, we adopted the association rule change mining technique to discover changes of events. As mentioned above, transaction data and news data are quite different. We modified the conventional change mining approach, and used event property and concept hierarchy to improve the quality of change detection.

3. EVENT CHANGE DETECTION (ECD) TECHNIQUE

In a dynamic environment, the state of events is constantly changing and evolving. It is important for business managers to capture environmental changes to adjust their business strategies.

3.1. Overview

The proposed event change detection (ECD) technique comprises two main processes: learning and detection. From a set of news stories, the learning process (as shown in Figure 1) seeks to identify event properties, concept hierarchy of property, and association rules representing the subject behaviors of event. The purpose

of the detection process is to identify event change, according to the association rules discovered in the learning process (as illustrated in Figure 2).

3.2. Learning process

The learning process of the ECI technique is to identify which event in the news story should belong and induce event association rules from a set of news stories. The learning process consists of four steps, including news fetcher, event identification, property extraction and association rule matching.

The news fetcher component of the ECD technique is to fetch news stories from online news providers.

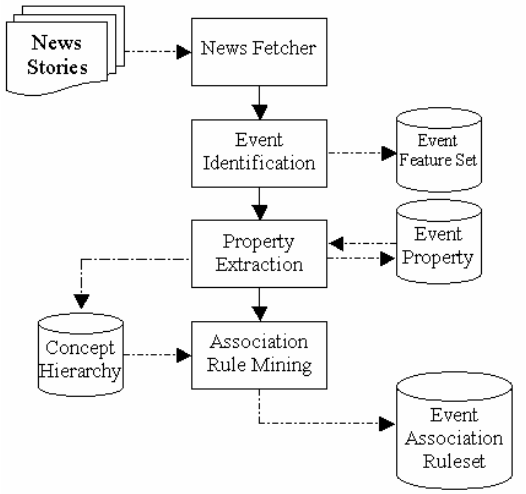

Figure 1. Learning process of identifying event change

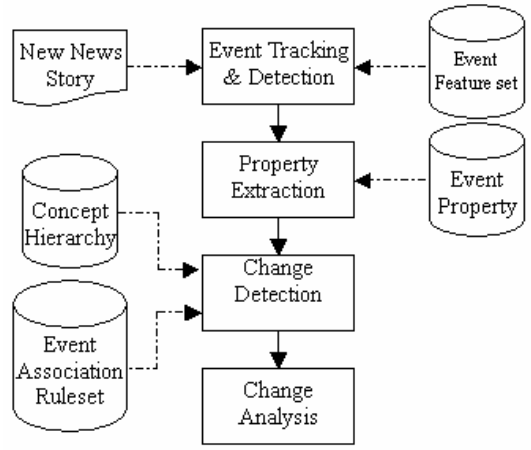

Figure 2. Detection process of identifying event change

3.2.1. Event Identification

The objective of event identification is to determine the set of features that will be used for representing individual news stories within pre-classified training news stories. The search for news categorization patterns uses feature extraction and feature selection. Feature extraction extracts a set of terms based on the news stories. For producing a set of features, we used feature extraction with the Chinese Dictionary [3] which contains one hundred and sixty thousand terms to parse each training news story to produce a list of terms commonly referred to as features.

Following feature extraction, feature selection is initiated to condense the size of the event feature set (the size of feature set is 30). The purpose of this phase is to remove unnecessary terms from the set that were produced in the previous phase. Several feature selection methods have been proposed in the literature, including TF, TF*IDF and so on. We selected the features in this research by TF*IDF. The top 30 features with the highest feature score are selected as features for representing news stories.

TF*IDF is defined as [6]:

$$TF \times IDF(t,d) = tf(t,d) \times \log(N / n_t) \quad (1)$$

where

N is the size of the news stories in the collection;
n_t is the number of news stories where t occurs;
$tf(t,d)$ is the term frequency (TF) within a news story d;
$\log(N/ n_t)$ is the inverted document frequency (IDF).

3.2.2. Event property

Feature-based technique is one of the most popular methods to represent the content of a news story. But feature-based representation often performs poorly because of its inherent problems, such as vocabulary discrepancies between reporters, news stories for different events containing similar feature sets, and so on. To overcome the problems of the traditional feature-based technique, event property has been used to improve event detection [1][14]. To improve the accuracy of content representation, we adopted event property to identify the subject behaviors of events. Event properties are classified into four categories (4Ws) as summarized below.

1. When: Date, Time
2. Who: Person, Organization
3. Where: Location
4. What: Action, Claim, Standpoint, Statement

3.2.3. Concept hierarchy

As previously mentioned, business decision makers may require different levels of information for developing different strategies. Such different levels of information form a concept hierarchy. In this work, we adopted concept hierarchy to meet all possible information needs. Concept hierarchy is used to define the relations of concept levels between a group of attributes [13].
In the process of event change detection, we need to determine the difference between the values of attributes. Subject behaviors of two time periods are compared. These values can be regarded as the nodes of concept hierarchy.

The node difference in a concept hierarchy is:

$$h_H(A,B) = \frac{Max\left(\sum_{L_i \in P_A} WL_i, \sum_{L_j \in P_B} WL_j\right) - \sum_{L_k \in P_{comm}(A,B)} WL_k}{Max(\sum_{L_i \in P_A} WL_i, \sum_{L_j \in P_B} WL_j)} \quad (2)$$

A and B are nodes in concept hierarchy
P_A : the path from root to node A.
P_B : the path from root to node B.
$P_{comm(A, B)}$: the common path between P_A and P_B.
L_i: a link in P_A.
L_j: a link in P_B.
WL_i: the weight on the level of link *i*.
WL_j: the weight on the level of link *j*.

3.3. Event tracking and detection

The event tracking and detection process identifies which event a recent news story belongs to and whether the news story discusses a new event. To achieve this, the event tracking and detection process consists of three steps: news document representation, similarity comparison and event assignment.

News document representation:
Each news document is represented using representative features. The feature set for each event is extracted and selected from a set of labeled training news stories.

Similarity comparison:
This step is to calculate the similarity between recent news story *D* and all known events. The cosine distance is used to compute the similarity, as shown as below:

$$Sim(D,e) = \frac{\sum_{f_k \in D \cup e} t_{Dk} \times t_{ek}}{\sqrt{\sum_{f_k \in F_D} t_{ck}^2} \sqrt{\sum_{f_k \in F_e} t_{ek}^2}} \quad (3)$$

where F_D and F_e are the feature sets of document D and event *e* respectively. t_{Dk} is the weight of the representative feature f_k of *D*, and t_{ek} is the weight of the representative feature f_k of *e*.

Event assignment:
Upon obtaining the similarity scores between the latest news story D and all known events, we set a pre-specified threshold in this step, and assign an event label according to similarity scores. If the maximum similarity score between D and the known events is below the threshold, the new document is labeled as the first story of a novel event; otherwise, we assign the news document to the event where the maximum similarity score is obtained.

4. DETECTION OF EVENT CHANGE

Event change is the change of event trends in two time periods. Those trends are discovered from news stories of the same event during different times. In order to detect event changes from a great quantity of news stories, we extended the idea of change detection methodology [13] and integrated concept hierarchy into the event change detection technique. The detection of event change is revealed in Figure 3.

Based on past research and business requirements, there are five types of possible mining changes: emerging patterns, unexpected consequent changes, unexpected condition changes, added rules, and perished rules.

4.1. Discovery of changed rule

In order to discover the five types of event changes between different time periods, we employ the rule matching method proposed by Song et al. [12] and modify it to fit event change detection by considering concept hierarchy.

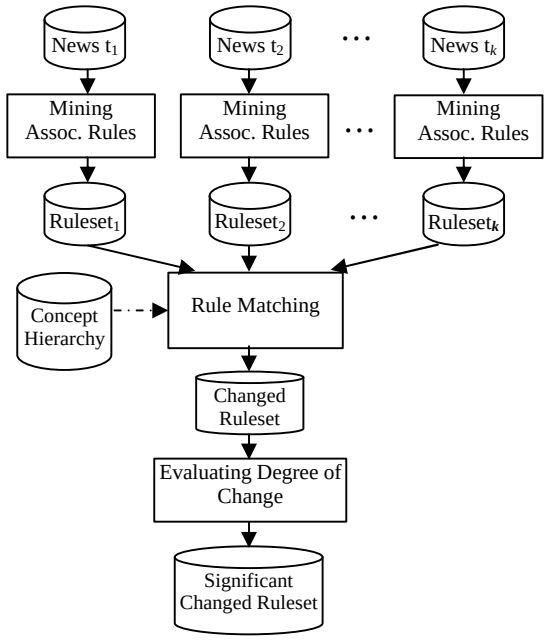

Figure 3. The process of event change detection

The rule matching method consists of three steps:

1. Calculate the maximum similarity value. For each rule in different time periods, calculate the maximum similarity value.
2. Calculate the difference measure. For each rule r_i^t in time *t*, calculate the difference measures between r_i^t , r_j^{t+k} .
3. Decide type of changes. According to the judged factors (maximum similarity value and the difference measures), this step determines the type of change for the rules.

Before we introduce the methods to calculate judged factors, some notations are defined as below.

Conditional part of rules:

p_{ij} : Degree of attribute match of conditional part

$$p_{ij} = \frac{|A_{ij}|}{\max(|X_i^t|, |X_j^{t+k}|)}$$

A_{ij} : Set of attributes common to both conditional parts of r_i^t and r_j^{t+k}

X_i^t : Set of attributes in the conditional parts of r_i^t

X_j^{t+k} : Set of attributes in the conditional parts of r_j^{t+k}

A_{ijk} : The k^{th} attribute in A_{ij} .

$v(r_i^t, A_{ijk})$: Value of the k^{th} attribute in A_{ij} of r_i^t .

$v(r_j^{t+k}, A_{ijk})$: Value of the k^{th} attribute in A_{ij} of r_j^{t+k} .

l_{ijk} : Degree of value match of the k^{th} matching attribute in A_{ij}

$$l_{ijk} = 1 - h_H\left(v(r_i^t, A_{ijk}), v(r_j^{t+k}, A_{ijk})\right)$$

Consequent part of rules:

q_{ij} : Degree of attribute match of consequent part

$$q_{ij} = \frac{|B_{ij}|}{\max(|Y_i^t|, |Y_j^{t+k}|)}$$

B_{ij} : Set of attributes common to both consequent parts of r_i^t and r_j^{t+k}

Y_i^t : Set of attributes in the consequent parts of r_i^t

Y_j^{t+k} : Set of attributes in the consequent parts of r_j^{t+k}

B_{ijm} : The m^{th} attribute in B_{ij} .

$v(r_i^t, B_{ijm})$: Value of the m^{th} attribute in B_{ij} of r_i^t .

$v(r_j^{t+k}, B_{ijm})$: Value of the m^{th} attribute in B_{ij} of r_j^{t+k} .

f_{ijm} : Degree of value match of the m^{th} matching attribute in B_{ij}

$$f_{ijm} = 1 - h_H\left(v(r_i^t, B_{ijm}), v(r_j^{t+k}, B_{ijm})\right)$$

The similarity measure S_{ij} between r_i^t and r_j^{t+k} is calculated using the following formula: ($0 \le S_{ij} \le 1$)

$$S_{ij} = \begin{cases} \dfrac{p_{ij} \times \sum_{k=1}^{|A_{ij}|} l_{ijk}}{|A_{ij}|} \times \dfrac{q_{ij} \times \sum_{m=1}^{|B_{ij}|} f_{ijm}}{|B_{ij}|}, & \text{if } |A_{ij}| \neq 0 \text{ and } |B_{ij}| \neq 0 \\ 0, otherwise \end{cases} \quad (4)$$

Where $\dfrac{p_{ij} \times \sum_{k=1}^{|A_{ij}|} l_{ijk}}{|A_{ij}|}$ represents a similarity of conditional part, and $\dfrac{q_{ij} \times \sum_{m=1}^{|B_{ij}|} f_{ijm}}{|B_{ij}|}$ represents a similarity of consequent part between r_i^t and r_j^{t+k} . If the conditional parts and consequence parts r_i and r_j^{t+k} are the same, S_{ij} equals 1.

Accordingly, the maximum similarity value of r_i^t is as below: $V_i = \max(S_{i1}, S_{i2}, \cdots, S_{i|R_{t+k}|})$

Accordingly, the maximum similarity value of r_j^{t+k} is as below: $V_j = \max(S_{1j}, S_{2j}, \cdots, S_{|R_t|j})$

The second judged factor, difference measure ∂_{ij} between r_i^t and r_j^{t+k} is given by ($-1 \le \partial_{ij} \le 1$, $|\partial_{ij}| \le 1$)

$$\partial_{ij} = \begin{cases} \dfrac{p_{ij} \times \sum_{k=1}^{|A_{ij}|} l_{ijk}}{|A_{ij}|}, & \text{if } |A_{ij}| \neq 0 \text{ and } |B_{ij}| = 0 \\ \dfrac{p_{ij} \times \sum_{k=1}^{|A_{ij}|} l_{ijm}}{|A_{ij}|} - \dfrac{q_{ij} \times \sum_{m=1}^{|B_{ij}|} f_{ijm}}{|B_{ij}|}, & \text{if } |A_{ij}| \neq 0 \text{ and } |B_{ij}| \neq 0 \\ -\dfrac{q_{ij} \times \sum_{m=1}^{|B_{ij}|} f_{ijm}}{|B_{ij}|}, & \text{if } |A_{ij}| = 0 \text{ and } |B_{ij}| \neq 0 \end{cases} \quad (5)$$

If attributes of conditional parts and consequent parts between the two rules are different, it makes no sense to compare the degree of difference, i.e., $|A_{ij}| = 0$ and $|B_{ij}| = 0$, because these two rules are completely different.

Although r_j^{t+k} is determined to be an unexpected change with regard to r_i^t by ∂_{ij} , it may be an emerging pattern of other rule. An emerging pattern may be misjudged to be an unexpected change. To improve the accuracy of judging change type, the third judged factor, modified difference measure $\partial_{ij}^{'}$ is denoted as:

$$\partial_{ij}^{'} = |\partial_{ij}| - j_{ij}, \text{where } j_{ij} = \begin{cases} 1, & \text{if } \max(V_i, V_j) \ge q_1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

The parameters q_1, q_2, and q_3, (event change thresholds) are thresholds to determine the types of change. The five types of change can be classified according to the three kinds of judged factors and the three predefined thresholds: q_1 for emerging patterns, q_2 for unexpected consequent and unexpected condition changes, and q_3 for added and perished rules. In general, $q_1 > q_2 > q_3$, and q_1 approaches to 1. Table 1 shows the value of measure for each type of change.

Table 1. Value of measure for each type of change

Type of Change	Value of measure to classify
Emerging Pattern	$S_{ij} > q_1$
Unexpected Consequent	$\partial_{ij} > 0, \partial_{ij}^{'} \ge q_2$
Unexpected Condition	$\partial_{ij} < 0, \partial_{ij}^{'} \ge q_2$
Added Rule	$V_j < q_3$
Perished Rule	$V_i < q_3$

4.2. Evaluating degree of event changes

Changes of environment are varied. Managers need to focus on essential changes. To achieve this goal, it is important to evaluate the degree of change, and rank changed rule by degree. Table 2 shows a simple formulation of change degree evaluation.

Table 2. Degree of event change

Type of Change	Degree of Change
Emerging Pattern	$\dfrac{Support^{t+k}(r_j) - Support^t(r_i)}{Support^t(r_i)}$

Unexpected Change	$\frac{Support^t(r_i) - Support^{t+k}(r_i)}{Support^t(r_i)} \times Support^{t+k}(r_j)$
Perished Rule	$(1-V_i) \times Support^t(r_i)$
Added Rule	$(1-V_j) \times Support^{t+k}(r_j)$

Let $support^t(r_i)$ and $support^{t+k}(r_i)$ represent the support value of r_i at time t and $t+k$, respectively. Degree of change for emerging pattern shows the change of support value between time t and time $t+k$. Degree of change for unexpected change is the change ratio of r_i multiplied by the support value of r_j at time $t+k$. The change ratio of r_i represents the degree of unexpectedness of r_i, namely the proportion of the difference between $support^t(r_i)$ and $support^{t+k}(r_i)$ to the $support^t(r_i)$. The degree of change is larger, if the change ratio of r_i is larger and $support^{t+k}(r_j)$ is larger. Degree of change for perished (added) rule is obtained from the support value of perished (added) rule multiplied by the value of 1 minus the maximum similarity value. The degree of change is larger, if the perished (added) rule has less maximum similarity value.

5. CONCLUSION

The business environment today is more and more complicated. The ability to capture environment changes and be sensitive to the environment becomes a critical success factor for business. Research on environment scanning currently puts great emphasis on event tracking and event detection. But the main purpose of event tracking and event detection is to identify which event the news story describes. Business administrators cannot obtain environment change information until a new event occurs. It is insufficient for managers to be notified only when a new event happens.

To capture the changes of events, this research proposes an event change detection (ECD) technique. The proposed technique combines the change mining approach and concept hierarchy to detect changes to enhance environmental scanning.

REFERENCES

[1] Allen J., Papka R. and Lavrenko V., "On-line new event detection and tracking," *in Proc. of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, Melboume, Australia, pp.37-45, 1998.

[2] Brants T., Chen F. and Farahat A., "A system for event detection," *in Proc. of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, Toronto, Canada, 2003, pp.330-337.

[3] Chinese Dictionary, http://william.cswiz.org/techreport/moecdict

[4] Choo C. W., "The art of scanning the environment," *Bulletin of the American Society for information Science and Technology*, vol.25, no.3, pp.21-24, 1999.

[5] Dong G. and Li J., "Efficient mining of emerging patterns: discovering trends and differences," *in Proc. of 5th International Conference on Knowledge Discovery and Data Mining, KDD-99*, 1999, pp.43-52.

[6] Han J. and Kamber M., *Data Mining-Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.

[7] Jennings D. and Lumpkin J., "Insights between environmental scanning activities and porter's generic strategies: an empirical analysis," Journal of Management, vol.18, no.4, pp. 791-803, 1992

[8] Lanquillon C., "Information filtering in changing domains," *in Proc. of the International Joint Conference on Artificial Intelligence, IJCA99*, pp.41-48, 1999.

[9] Liu B. and Hsu W., "Post-analysis of learned rules," *in Proc. of 13th National Conference on Artificial Intelligence, AAAI-96*, pp.828-834, 1996.

[10] Liu B., Hsu W., Han H.-S. and Xia Y., "Mining Changes for Real-life Applications," *the Second International Conference on Data Warehousing and Knowledge Discovery (DaWak)*, pp.337-346, 2000.

[11] Liu B., Hau W. and Ma Y., "Discovering the Set of Fundamental Rule Changes," in *Proc. of the 7th ACM International Conference on Knowledge Discovery and Data Mining*, 2001.

[12] Song H. S., Kim J. K. and Kim S. H., "Mining the change of customer behavior in an internet shopping mall," *Expert Systems with Applications*, vol. 21, pp.157-168, 2001.

[13] Srikant R. and Agrawal R., "Mining Generalized Association Rules," *in Proc. of the 21th VLDB Conference*, Zurich, Swizerland, 1995.

[14] Wei C.-P. and Lee Y.-H. "Event detection from online news documents for supporting environmental scanning," *Decision Support Systems*, vol.36, pp. 385-401, 2004.

[15] Witten I. H. and Frank E., "Output: Knowledge Representation," in *Data Mining*, Morgan Kaufmann Publishers, San Francisco, 2000.

[16] Yang Y., Pierce T., and Carbonell J. G., "A study on retrospective and on-line event detection," in *Proc. of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, Melboume, Australia, 1998, pp.28-36.

[17] Yang Y., Carbonell J. G., R. D. Brown, T. Pierce, B. T. Archibald and X. Liu, "Learning approaches for detecting and tracking news events," *IEEE Intelligent Systems*, vol.14, pp.32-43, 1999.