December 2001

Event Detection for Supporting Environmental Scanning: An Information Extraction-based Approach

Chih-Ping Wei
National Sun Yat-Sen University

Yen-Hsien Lee
National Sun Yat-Sen University

Event Detection for Supporting Environmental Scanning: An Information Extraction-based Approach

Chih-Ping Wei and Yen-Hsien Lee
Department of Information Management
National Sun Yat-Sen University
Kaohsiung, Taiwan, R.O.C.

Abstract

Environmental scanning, the acquisition and use of the information about events, trends, and relationships in an organization's external environment, permits an organization to adapt to its environment and to develop effective responses to secure or improve the organization's position in the future. Event detection technique that identifies the onset of new events from streams of news stories would facilitate the process of organization's environmental scanning. However, traditional feature-based event detection techniques cannot capture the genuine properties of an event contained in a news story and cannot support event categorization and news stories filtering. In this study, we developed an information extraction-based event detection (NEED) technique that combines information extraction and text categorization techniques to address the problems inherent to traditional feature-based event detection techniques. Using a traditional feature-based event detection technique (INCR) as benchmarks, the empirical evaluation results showed that the proposed NEED technique improved the effectiveness of event detection measured by miss and false alarm rates.

Keywords: event detection, information extraction, text categorization, environmental scanning

1. Introduction

As an organization's environment becomes more complex and dynamic, uncertainty faced by the organization increases. Environmental scanning is the first link in the chain of perceptions and actions that permit an organization to adapt to its environment and subsequently to develop effective responses to secure or improve their position in the future (Choo, 1999). As defined by Choo (1999), environmental scanning refers to "the acquisition and use of information about events, trends and relationships in an organization's external environment, the knowledge of which would assist management in planning the organization's future course of action." Empirical research results suggest that environmental scanning is linked with improved organizational performance (Ahituv et al., 1998).

However, the advance of information technology and the proliferation of Internet have made the amount of scanning information exploded. The increases in scope and complexity of business environments also make the interval between scanning efforts needed shorten. As a result, environmental scanning becomes more difficult to handle and has been a burden to managers. Thus, an information system to facilitate organizational scanning of environments is essential. Specifically, the system needs to support detecting the onset of new events from news documents and tracking subsequent news stories that discuss an event of interest.

Event detection is to identify the onset of new events from streams of news stories (Allan et al., 1998; Yang et al., 1999). Traditional event detection techniques usually adopted the

general feature co-occurrence approach. It identifies whether a news story contains an unseen event by comparing the similarity of features between the news story and past news stories. Because news stories discussing the same event tend to be temporally proximate, a combined measure of lexical similarity and temporal proximity as a criterion for event detection was often employed (Yang et al., 1999). Moreover, since a time gap between bursts of topically similar stories is often an indication of different events, the incorporation of a time window for event scoping was commonly adopted (Yang et al., 1999).

Nevertheless, traditional feature-based event detection techniques incur several problems. First, there exist vocabulary discrepancies between reporters even when they describe the same event. For example, some may use "merger" or "purchase" to describe a business merger event, while others may use "acquisition" for the same event. Moreover, two news stories discussing the same event may be oriented from different angles, resulting in differences in features. Secondly, two news stories for different events may contain very similar feature sets since the events belong to the same event topic. For example, in the event topic of computer virus, the features "virus," "computer," "worm," and "inflection" may appear in every virus news story. In this case, these news stories will be similar, even though they discuss two different computer viruses. Finally, it would be essential to not only detecting whether a news story contains an unseen event, but also classifying the news story into an appropriate event topic. With the event categorization, filtering of news stories that are not of interest to a specific user can easily be supported. However, traditional event detection techniques are not capable of supporting event categorization.

To overcome the problems inherent to traditional event detection techniques, understanding of news stories is necessary. It can be achieved by classifying a news story into an appropriate event topic and subsequently extracting information on the event properties associated to the target event topic. Two news stories discuss different events if they are assigned to different topics or some of the event property values are different, regardless whether the features in the two news stories are similar. On the other hand, two news stories are assumed to describe the same event if they belong to the same event topic and their event property values are the same or similar. Thus, the first and second problems of traditional feature-based event detection techniques can be solved by performing event detection based on event property values embedded in news stories rather than features appearing in news stories. Since the event topics can serve as the categories for classifying or filtering new stories, the third problem inherent to traditional event detection techniques can be overcome.

Motivating by the need for improving the event detection accuracy and supporting event categorization, the goal of this research is to develop an event detection technique based on the information extraction approach, called iNformation Extraction-based Event Detection (NEED) technique. The proposed technique will empirically be evaluated, using a traditional event detection technique as benchmarks. The rest of the paper is organized as follows. Section 2 reviews literatures relevant to this research. The development of the iNformation Extraction-based Event Detection (NEED) technique will be depicted in Section 3. An empirical evaluation using news stories from a news website will be conducted and summarized in Section 4. Finally, the contributions of this study as well as future research directions will be summarized in Section 5.

2. Literature Review

2.1 Event Detection

The objective of event detection is to identify stories in several continuous news streams that pertain to new or previously unidentified events (Yang et al., 1999). Event detection is subdivided into two forms: retrospective detection and online detection. The former entails the discovery of previously unidentified events in a chronologically ordered accumulation of documents (stories), and the latter strives to identify the onset of new events from live news feeds in real-time. Both forms of detection intentionally lack advance knowledge of novel events, but do have access to unlabelled historical news stories for use as contrast sets.

Most of the proposed event detection algorithms, retrospective or online, were developed based on the document clustering approach. Yang et al. (1999) implemented two clustering methods for event detection: GAC and INCR. GAC, operating in a strict retrospective detection setting, performs agglomerative clustering, producing hierarchically organized document clusters. GAC employed the conventional vector space model to represent documents and clusters. Each document is represented using a vector of weighted terms, based on the TF×IDF (within-document frequency × inverse document frequency) scheme.

For cluster representation, the normalized vector of documents in a cluster is summed and the k most significant terms called the *prototype* or *centroid* of the cluster are selected to represent the cluster. GAC is a divide-and-conquer version of a group-average clustering algorithm. Group-average clustering maximizes the average similarity between document pairs in the resulting clusters by merging clusters in a greedy, bottom-up fashion. To improve the computation efficiency and to preserve the characteristics that events tend to appear in news bursts, GAC adopted a divide-and-conquer strategy that grows clusters iteratively in a bottom-up fashion. In each iteration, the current pool of clusters is divided according to their order in time into evenly sized buckets. Subsequently, group-average clustering is applied to each bucket locally, merging smaller clusters into larger ones. Periodically, the stories within each of the top-level clusters are reclustered. Reclustering is useful when events straddle the initial temporal-bucket boundaries or when the bucketing causes undesirable groupings of stories about different events.

On the other hand, INCR, designed for both retrospective and online detection, is a single-pass incremental clustering algorithm that produces nonhierarchical clusters incrementally (Yang et al., 1999). For retrospective detection, the TF×IDF scheme was adopted to represent documents or clusters. However, to deal with the problem of continuously incoming documents that might affect term weighting and vector normalization during online detection, the incremental IDF was employed by INCR. Moreover, INCR incorporated a time penalty when calculating the similarity between a document x and any cluster c in the past. The time penalty can be a uniformly weighted time window (i.e., a time window of m documents before x is imposed) or a linear decaying-weight function (shown as below).

$$similarity(x,c) = \begin{cases} (1-\frac{i}{m}) \times similarity(x,c) & \text{if } c \text{ has any member in the time window} \\ 0 & \text{otherwise} \end{cases}$$

where i is the number of documents between x and the most recent member document in c, and m is the time window of documents before x.

For retrospective detection, INCR sequentially processes news documents. A document is absorbed by the most similar cluster in the past if the similarity between the document and cluster is larger than a pre-selected *clustering threshold* (t_c); otherwise, the document

becomes the seed of a new cluster. For online detection, the *novelty threshold* (t_n) was introduced. If the maximal similarity between the current document and any cluster in the past is no less than t_n, the document is flagged as containing an old event.

2.2 Text Categorization

Text categorization refers to the assignment of textual documents, on the basis of their contents, to one or more pre-defined categories (Apté et al., 1994; Cohen and Singer, 1999; Dumais et al., 1998; Yang and Chute, 1994). The challenging research issue of text categorization is the development of statistical or inductive learning methods for automatically discovering text categorization patterns, based on a training set of manually categorized documents. In general, automatic learning text categorization patterns encompasses three main phases (Apté et al., 1994): feature extraction and selection, representation, and induction.

The feature extraction and selection phase is undertaken to determine a set or sets of features (a universal dictionary or local dictionaries) that will be used for representing individual documents. The universal dictionary is created for all categories, while each local dictionary is created for a particular category. The text portion of the training documents is parsed to produce a list of nouns or noun phrases (called features) none of which either belongs to a pre-defined list of stop words or is a number or part of a proper name. After the feature extraction, the feature selection is initiated to reduce the number of unnecessary features. Several feature selection methods have been proposed in the literature (Dumais et al., 1998; Lewis and Ringuette, 1994; Ng et al., 1997; Schutze et al., 1995), including TF, TF×IDF, correlation coefficient, mutual information, and χ^2 metric. The top k features with the highest feature selection metric score are selected as features for representing documents.

In the representation phase, each individual document is represented in terms of features in the dictionary (universal or local) generated in the previous phase. A document is labeled to indicate its category membership and assigned a value for each feature in the dictionary, where the values can be either boolean (e.g., indicating whether or not the feature appears in the document), or numerical (e.g., frequency of occurrence in the document being processed). Different document representation methods have been proposed (Yang and Chute, 1994), including binary, TF, IDF and TF×IDF.

The induction phase is designed to automatically discover text categorization patterns that distinguish categories from one another, based on a training set of manually categorized documents. The learning strategies for automatically learning text categorization patterns can essentially be subdivided into the following types: decision tree induction (Quinlan, 1993); decision rule induction (Apté et al., 1994; Cohen and Singer, 1999); k-nearest neighbor classification (Larkey and Croft, 1996, Masand et al., 1992; Yang, 1994); neural network (Ng et al., 1997); Naïve Bayes probabilistic classification (Baker and McCallum, 1998; Larkey and Croft, 1996; Lewis and Ringuette, 1994); and regression approach (Yang and Chute, 1994). For interested readers, a more detailed summary and empirical comparisons can be found in (Yang and Liu, 1999).

2.3 Information Extraction

Information extraction is concerned with extracting relevant data from semi-structured or unstructured documents and transforming them into structured representations (Riloff and

Lehnert, 1994). Information extraction systems do not attempt in-depth understanding of text in documents. Rather, they analyze those portions of documents that contain information relevant to a pre-specified template that defines types of information to be extracted. Examples of template representation include case frames consisting of a set of slots (Riloff and Lehnert, 1994) and ontologies based on a semantic data model (Embley et al., 1998). A key element of information extraction systems is its set of extraction rules that is used to extract from each document the information relevant to a particular extraction task (Muslea, 1999). Extraction rules are typically based on a combination of syntactic (i.e., syntactic relations between words) and semantic (i.e., semantic classes of words) constraints that help identify the relevant information within a document.

The extraction rules can be manually coded or generated from training examples by using inductive learning techniques. Several information extraction learning system have been proposed in the literature. For example, WHISK adopted the top-down induction approach for learning extraction rules. WHISK begins with an empty rule and then extends the rule by adding terms. Terms are added to a rule one at a time until the errors are reduced to zero or a pre-pruning criterion has been satisfied. The process is repeated until a set of rules has been generated that cover all possible extractions from the training, at which time post-pruning is conducted to remove insignificant rules to prevent from overfitting. For interested readers, a survey of different information extraction learning systems can be found in (Muslea, 1999; Soderland, 1999; Eikvil, 1999).

Given a template and extraction rules relevant to a particular extraction task, several steps are often required before extracting relevant information from a target document, including syntactic analysis, semantic tagging, and discourse analysis (Muslea, 1999; Soderland, 1999; Eikvil, 1999; Cowie and Lehnert, 1996). Syntactic analysis allows preliminary recognition of phrasal units in sentences and parses the target document into a syntactic parse structure by using a syntactic grammar. Semantic tagging applies semantic interpretation rules or semantic dictionaries for recognizing the semantic classes of phrasal units, including company names, places, people's names, currencies, etc. Discourse analysis makes inference across sentence boundaries, involving co-reference resolution that refers to the problem of knowing when a new noun phrase refers back to a previously encountered referent. Subsequently, matching extraction rules on the pre-processed document is conducted for desired information extraction.

3. Information Extraction-based Event Detection (NEED) Technique

To overcome the disadvantages of traditional feature-based event detection techniques, a new event detection technique, called iNformation Extraction-based Event Detection (NEED) technique, was proposed. As mentioned, the proposed approach employs the information extraction method and the text categorization technique as a basis for event detection. The use of the information extraction turns event detection from feature-based into event property-based. This shift has the potential to improving the event detection accuracy and facilitating subsequent event tracking. On the other hand, the use of text categorization is to facilitate information extraction at the event topic level and support event categorization and filtering. Accordingly, the NEED technique comprises two main processes: learning and detection. From a set of news stories with known event topics (called training news stories), the learning process is to induce event categorization patterns for each event topic. When a new news story arrives, the detection process is applied to identify to which event topic the news story should belong and whether the news story discusses a new event.

3.1 Learning Process

The learning process is to induce event categorization patterns from a set of manually categorized news documents. Same as the text categorization mentioned previously, the learning process (as shown in Figure 1) consists of three steps including feature extraction and selection, document representation, and induction.

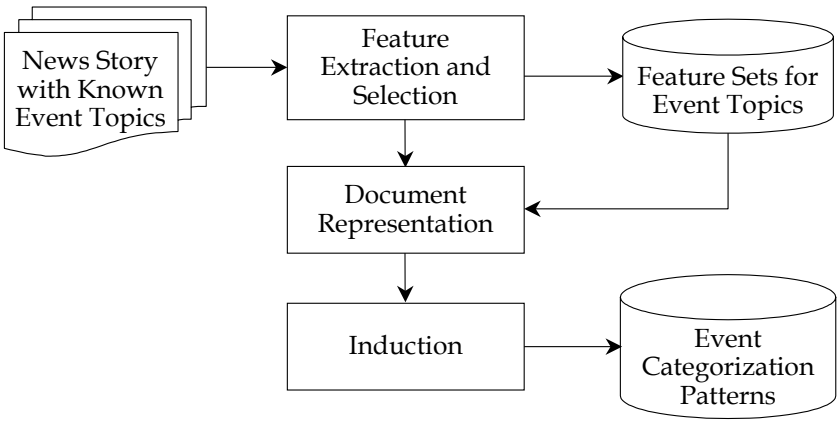

Figure 1: Learning Process of NEED Technique

Feature Extraction and Selection: A set of nouns and noun phrases from training news documents is first extracted. In this study, a rule-based part of speech tagger proposed by Brill (1994) was adopted for syntactically tagging each word in the news documents. For extracting noun phrases from syntactically tagged documents, a noun phrase parser proposed by Voutilainen (1993) was implemented. Subsequently, representative features for each event topic will be selected based on some feature selection metric. In this study, local dictionaries were constructed based on the correlation coefficient or TF×IDF feature selection method.

Document Representation: Each news story is then represented using the features set for the event topic to which the news story belongs. The binary and TF schemes were adopted as alternative document representation methods in this study.

Induction: This step is to induce for each event topic the event categorization patterns that will be used to categorize future news stories into appropriate event topics by the detection process. As mentioned in Section 2, several text categorization approaches have been proposed in the literature. The decision tree induction and decision rule induction approaches were adopted as the induction techniques in this study. Specifically, we incorporated C4.5 (Quinlan, 1993) and CN2 (Clark and Boswell, 1991) as alternative induction algorithms in the learning process of the NEED technique.

3.2 Detection Process

The detection process is to identify to which event topic a newly arrived news story will belong and to detect whether the news story discusses a new event. To achieve this, the detection process consists of three steps: event topic reasoning, event extraction, and similarity comparison, as shown in Figure 2.

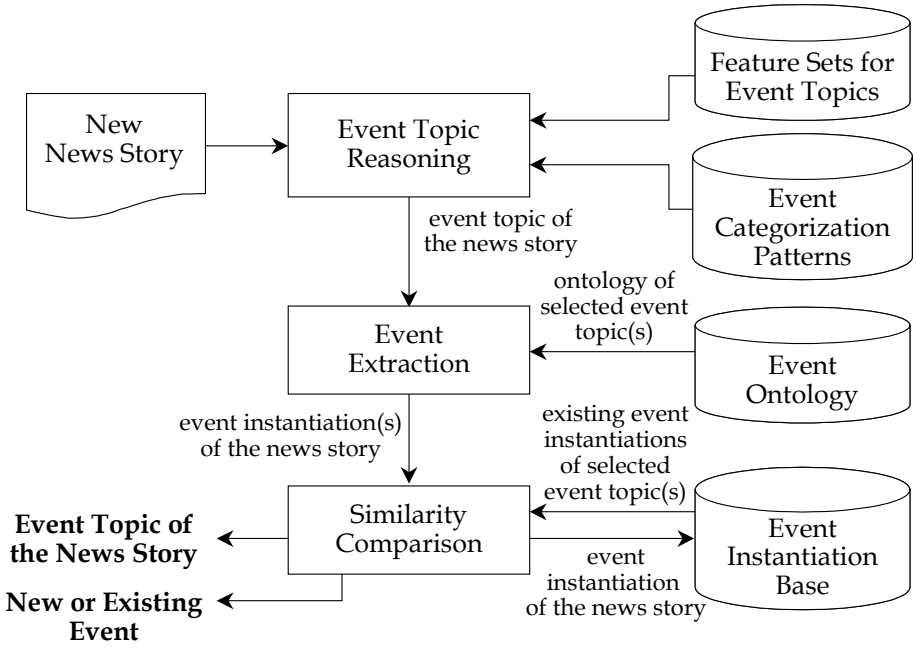

Figure 2: Detection Process of NEED Technique

Event Topic Reasoning: Based on the event categorization patterns induced previously in the learning process, the event topic reasoning step is to categorize the new news story into an appropriate event topic. Each new news story should first be represented according to the feature set for each event topic. In this study, the binary or the within document frequency (TF) method was employed. Accordingly, the reasoning with the event categorization patterns is performed. We assumed that a news story could belong to at most one event topic. Since the decision on whether the news story is classified into each event topic is made independently, two special cases arise; that is, the news story may be classified into more than one event topic or cannot be classified into any event topic. In the first special case, the conflict resolution is needed. Conflicts are resolved in this study by comparing the net support ratio. For each event topic with positive decision, the net support ratio is the number of training examples that satisfies the condition(s) of the fired rule(s) minus the number of training examples that satisfies the condition(s) but does not satisfy the decision of the fired rule(s), divided by the total number of training examples. For example, suppose a news story n_i can be classified into the event topic A, based on 20 training examples. Assume that 7 training examples satisfy the condition(s) of the fired rule(s) but 3 of them have a decision contradicting to the fired rule(s) for the event topic A. Thus, the net support ratio for the event topic A is (7-3)/20 = 0.2.

In the second case when the new news story cannot be classified into any event topic, the detection process could stop further processing and suggest that this news story is not belonging to any known event topics. Alternatively, the detection process could consider the event topic for this news story as undecided. In this view, this news story will be processed in every event topic at the subsequent steps and the decision on whether this news story contains an unseen event will be made across all event topics (to be explained in more detailed later).

Event Extraction: After the new news story is classified into an event topic, the event extraction step is to create an event instantiation (called target event instantiation) by extracting event property values from the news story based on the template and information extraction rules (called the event ontology) of the event topic. The ontology-based

information extraction system proposed by Embley et al. (1998) was adopted in this study for event extraction. If the event topic for the news story is undecided in the previous step, an event instantiation for the news story will be created for every event topic.

Similarity Comparison: The target event instantiation will then be compared with existing event instantiations of the same event topic. Within the event topic, the similarity is measured between the target event instantiation and each known event consisting of a set of existing event instantiations. The similarity function proposed for the NEED technique is defined as follows. For each slot in the template of the selected event topic, the target event instantiation is compared with every event instantiation belonging to the same event. The maximal similarity is obtained as the contribution of the slot to the overall similarity between the target news story and the event.

$$sim(TI_i, IS_j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} \underset{I_k \in IS_j}{MAX}(sim(TI_i^{(f)}, I_k^{(f)}))}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

where TI_i is the target event instantiation of the new news story,
IS_j is a set of existing event instantiations belonging to the event j,
I_k is an event instantiation in IS_j,
p is the number of slots in the template of the event topic j,
$\delta_{ij}^{(f)} = 1$ if both the fth slot in TI_i and the fth slot of some I_k in IS_j contain non-missing values; otherwise, $\delta_{ij}^{(f)} = 0$, and
$sim(TI_i^{(f)}, I_k^{(f)})$ is the similarity between the fth slot of TI_i and that of I_k.

If the fth slot is a numeric attribute, then $sim(TI_i^{(f)}, I_k^{(f)}) = 1$ when $TI_i^{(f)} = I_k^{(f)}$; otherwise, it is 0. On the other hand, if the type of the fth slot is of character strings and $SS_{ik}^{(f)}$ is the maximal matching substring between $TI_i^{(f)}$ and $I_k^{(f)}$, then $sim(TI_i^{(f)}, I_k^{(f)}) = \frac{\left|SS_{ik}^{(f)}\right|}{\max\left(\left|TI_i^{(f)}\right|, \left|I_k^{(f)}\right|\right)}$. If the target event instantiation or all event instantiations of the event contain missing values in a slot, this slot will be ignored from similarity calculation. If all slots are ignored (i.e., $\sum_{f=1}^{p} \delta_{ij}^{(f)} = 0$), the similarity between the target event instantiation and the event is set to 0.

Similar to traditional feature-based event detection techniques, a linear decaying-weight similarity function was employed for the NEED technique:

$$sim(TI_i, IS_j) = \begin{cases} (1 - \frac{i}{m}) \times sim(TI_i, IS_j) & \text{if } IS_j \text{ has any member in the time window} \\ 0 & \text{otherwise} \end{cases}$$

where i is the time gap measured in the number of days between TI_i and the most recent member instantiation in IS_j, and m is the time window measured in days prior to TI_i.

After the similarities between the target event instantiation and all of the known events in the selected event topic are obtained, the NEED technique labels the target news story containing a new event if the maximal similarity score between the target event instantiation and known events in the selected event topic is below a pre-specified novelty threshold (t_n); otherwise,

the target news story is labeled as containing an old event. As mentioned, if the news story whose event topic is undecided in the event topic reasoning step, its event instantiation was created in every event topic. Thus, the NEED technique labels the target news story a new event if the maximal similarity score between the news story and known events in all event topics is below the novelty threshold; otherwise, the news story is labeled as an old event. In the latter case, the news story will be assigned to the event topic where the maximal similarity score was attained.

Finally, the target event instantiation is stored in Event Instantiation Base for future event detection use. If the news story is specified as containing an old event, its event instantiation is absorbed by the event to which the new news story is associated; otherwise, it forms an event on its own.

4. Empirical Evaluation

This section reports the empirical evaluation of the proposed NEED technique, using a traditional feature-based event detection technique as performance benchmarks. News stories from November 1999 to December 1999 were collected from a news website, excite.com. Five event topics were identified and selected, including airplane crash, adjustment of interest rate, business merger, business partnership, and computer virus. 492 news stories (where 244 news from November and 248 news from December 1999) pertaining to the five event topics were manually identified. The event contained in each news story was also coded manually. For each event topic, the event ontology, as required by the NEED technique, was engineered manually in this study.

4.1 Evaluation Criteria

The effectiveness of an event detection technique is measured by the miss and false alarm rates. The miss rate is defined as the percentage of that an event detection technique fails to detect a new event, while the false alarm rate is defined as the percentage of that an event detection technique fails to detect an old event. To address the inevitable tradeoffs between miss and false alarm rates, Detection Error Tradeoff (DET) curves were employed (Yang et al., 1999; Allan et al., 1998). An event detection technique with its DET curve closer to the origin would be more desirable. In the context of supporting environmental scanning, a low miss rate may improve an organization's responsiveness to changes of its external environment and therefore can enhance the organization's adaptability to its environment. On the other hand, an improvement in the false alarm rate reduces an organization's load in filtering news stories containing known events. Because of ever-increasing complexity and dynamics of an organization's environment, responsiveness and adaptability of the organization clearly are more desirable than efficiency of environmental scanning. In this light, event detection should aim at achieving the lowest attainable miss rate while maintaining false alarm rate at a satisfactory level.

4.2 Performance Benchmarks

A traditional feature-based event detection technique was used to provide the desired effectiveness benchmarks. Specifically, the single-pass incremental clustering (INCR) for event detection proposed by Yang et al. (1999), was employed. Without loss of generality, we modified its linear time-decaying similarity function by changing the time window from the number of prior news stories to the number of days, as follows:

$$sim(x,c) = \begin{cases} (1 - \frac{i}{m}) \times sim(x,c) & \text{if } c \text{ has any member in the time window} \\ 0 & \text{otherwise} \end{cases}$$

where i is the number of days between x and the most recent document in c, and m is the time window of days before x.

4.3 Parameter Tuning Experiments for INCR

The INCR technique involves three parameters: the number of features k, time window w and novelty threshold t_n. The news stories of November 1999 were employed as the data set for parameter tuning. Specifically, the news stories from the first n days in the tuning set were used as historical news stories, while the rest of news stories in the tuning set were included as the testing set. Three different n were investigated in this study: 10, 15 and 20. To detect whether a news story in the testing set contained a new event by using the INCR technique, the news story was compared to all news stories (including historical ones) prior to the testing news story. Thus, the tuning experiment was performed three times and the overall detection effectiveness was estimated by averaging the performance across all iterations.

We investigated the number of features k ranging from 50 to infinite (k = 50, 100, 150, 200 and infinite), the time window w ranging from 7 to 30 (w = 7, 14 and 30 days), and the novelty threshold t_n ranging from 0.1 to 0.2 at 0.01 increments. At any level of w investigated, the Detection Error Tradeoff (DET) curve, in general, was getting closer to the origin as k increased from 50 to infinite (INF). On the other hand, when k = INF, the DET curve of the INCR technique moved toward to the origin as w grew from 7 to 30 (as shown in Figure 3). The increase of the novelty threshold resulted in the decrease of miss rate at the cost of false alarm rate. When w was 30 and k was infinite, the best performance was achieved at the novelty threshold of 0.19 (where the minimal Euclidean distance to the origin was attained). Thus, we selected 30 for the time window and infinite for the number of features for the INCR technique.

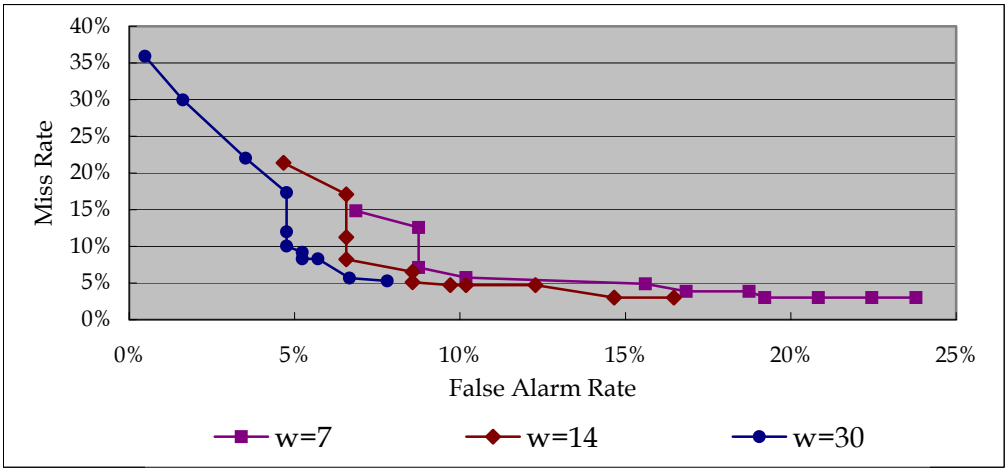

Figure 3: Detection Error Tradeoff Curves of the INCR Technique (k = Infinite)

4.4 Parameter Tuning Experiments for NEED

The NEED technique consists of two main processes: learning and detection. The learning process involved four decisions, including the feature selection method (correlation

coefficient or TF×IDF), the number of features (ranging from 50 to 200 at 50 increments, the representation method (binary or TF), and the induction algorithm (C4.5 or CN2). Similar to the tuning experiments for INCR, the news stories of November 1999 were employed as the data set for tuning. Specifically, we adopted the tenfold cross-validation technique, with which the included news stories were randomly divided into ten mutually exclusive data sets of equal size. The learning-and-testing proceeded in an iterative manner. In each learning-and-testing iteration, one data set was chosen as the testing data and the others were used for learning purpose. Thus, the overall learning performance was estimated by averaging the performance across the ten iterations.

As shown in Table 1, when CN2 was used as the induction algorithm, the TF representation outperformed the binary one in event categorization at almost any number of features investigated. When the number of features (k) was increased from 50 to 200, the average event categorization accuracy was generally decreased in any combination of feature selection and representation methods. In contrast, when C4.5 was employed (as shown in Table 2), the binary representation achieved better categorization accuracy at almost any number of features examined. The increment of the number of features generally resulted in lower categorization accuracy. In general, C4.5 appeared to outperform CN2 in event categorization accuracy. Among all experiments, the combinations of (induction algorithm = C4.5, feature selection = TF×IDF, representation = TF, k = 50) and (induction algorithm = C4.5, feature selection = correlation coefficient, representation = binary, k = 50) achieved the highest categorization accuracy. Thus, in this study, C4.5 with the TF×IDF feature selection method, the TF representation method and the number of features as 50 was adopted as the parameter setting for further experiments.

Table 1: Average Accuracy of Event Categorization (Adopting CN2 for Learning)

Feature Selection Method	**Representation Method**	**Number of Features (k)**			
		50	**100**	**150**	**200**
Correlation Coefficient	Binary	84.43%	84.02%	83.61%	81.56%
	TF	84.02%	85.25%	85.25%	85.66%
TF×IDF	Binary	82.38%	83.20%	82.79%	84.02%
	TF	85.66%	86.48%	84.43%	84.43%

Table 2: Average Accuracy of Event categorization (Adopting C4.5 for Learning)

Feature Selection Method	**Representation Method**	**Number of Features (k)**			
		50	**100**	**150**	**200**
Correlation Coefficient	Binary	88.52%	87.30%	87.30%	87.70%
	TF	85.25%	84.84%	84.84%	84.43%
TF×IDF	Binary	87.70%	88.11%	86.89%	87.30%
	TF	88.52%	88.11%	86.48%	86.48%

Once the parameter values for the learning process were determined, tuning experiments for the detection process of the NEED technique were conducted. As mentioned, the detection process involves two parameters: time window w and novelty threshold t_n. The news stories of November 1999 were employed as the data set for parameter tuning purpose. Specifically, the news stories from the first n (where n = 10, 15 or 20) days in the tuning set were used as historical news stories for event detection purpose and as training news stories for inducing event categorization patterns, while the rest of news stories in the tuning set were included as the testing set. To detect whether a news story in the testing set containing a new event, the

news story was first assigned to an event topic based on event categorization patterns induced from the historical news stories. As a result, the news story was compared to all news stories (including historical ones) that were prior to and were in the same event topic as the target testing news story. Three different n values were experiments and the overall detection effectiveness was estimated by averaging the performance across the three trials.

We investigated the time window ranging from 7 to 60 (w = 7, 14, 30, to 60). Since NEED takes into account only essential event property values rather than features in news stories during event detection, the appropriate range of the novelty threshold for NEED should be higher than that for INCR. Specifically, we investigated and the novelty threshold ranging from 0.51 to 1.0 at 0.01 increments. The DET curves of the NEED technique over different w and t_n are shown in Figure 4. As shown, the DET curves of the NEED technique shifted slightly toward the origin as the time window increased. The NEED technique arrived at the best performance when w was 60 and the novelty threshold was 0.59. We decided on the time window of 60 for the NEED technique, which appeared to achieve better performance when considering the tradeoff between miss rate and false alarm rate.

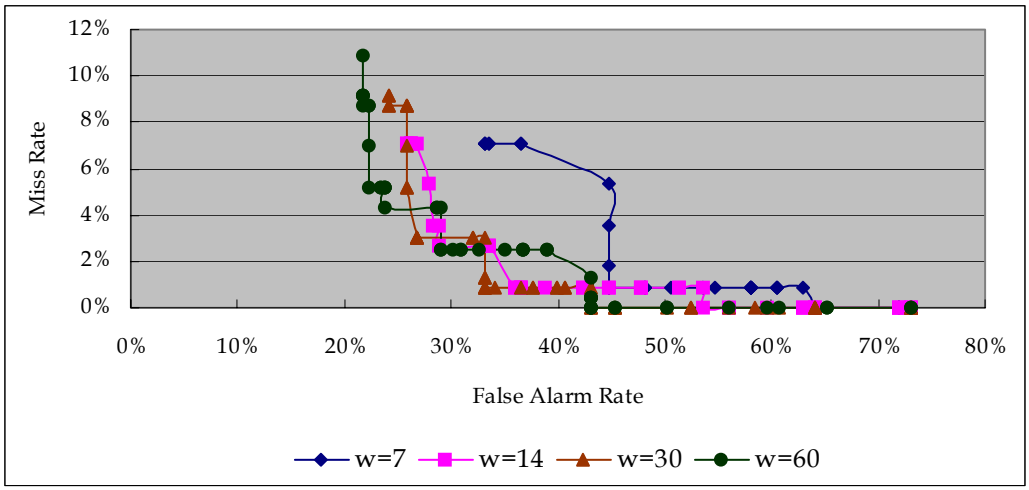

Figure 4: Detection Error Tradeoff Curves of the NEED Technique

4.5 Comparative Evaluation of Event Detection Techniques

The traditional feature-based event detection (INCR) and information extraction-based event detection (NEED) techniques were compared using the parameter values determined in the previous subsections. Similar to previous tuning experiments, the data corpus was divided into two sets: historical (including news stories in November 1999) and testing (including news stories in December 1999). Since the NEED technique requires inducing event categorization patterns, the historical data set was also used for the learning purpose. To expand the number of trials, 70% of news stories were randomly selected from the historical and the testing set, respectively, and the random selection process was repeated 30 times. In each trial, the reduced historical set was used for inducing event categorization patterns, as required by the NEED technique. To detect whether a news story in the reduced testing set contained a new event, the news story was compared to all news stories (including those in the reduced historical set) prior to the testing news story (by using the INCR technique) or compared to all news stories that were prior to the target testing news story and were in the same event topic as the target testing news story (by using the NEED technique). The overall detection effectiveness was estimated by averaging the performance across all trials.

We investigated the novelty threshold t_n for INCR ranging from 0.01 to 0.5 and that for NEED ranging from 0.51 to 1.0 at 0.01 increments. As shown in Figure 5, at almost any level of false alarm rate that was lower than 45%, the INCR technique achieved lower miss rates than the NEED technique did. However, if a low miss rate was desirable, the NEED technique outperformed its counterpart at almost any level of miss rate lower than 4%. In general, the miss rate achieved by the NEED technique was lower than 8% at any novelty threshold investigated. As mentioned, in the context of supporting environmental scanning, a low miss rate attainable by an event detection technique is more important than a low false alarm rate. Hence, the NEED technique was superior to the INCR technique.

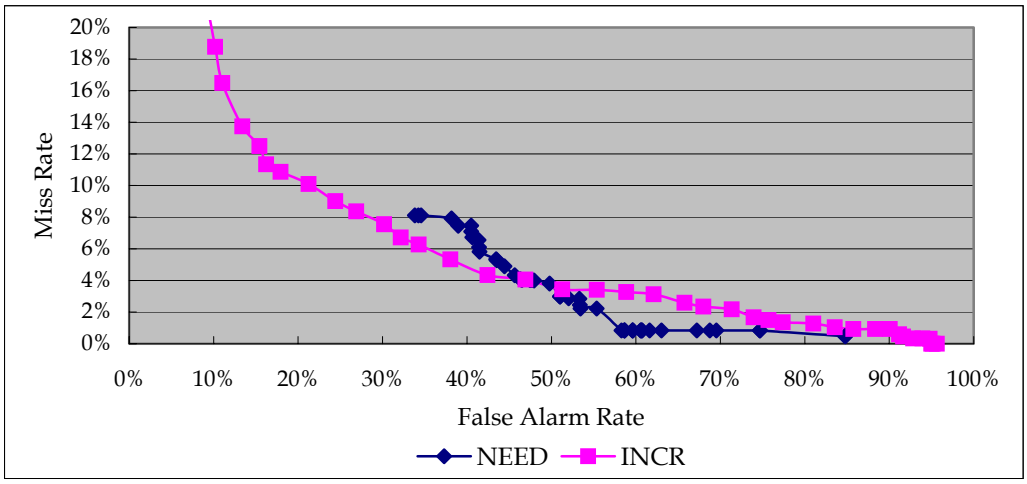

Figure 5: Detection Error Tradeoff Curves of Different Event Detection Techniques

5. Conclusion and Future Research Directions

Environmental scanning is an important process of strategic management that permits an organization to adapt to its environment and subsequently to develop effective responses to secure or improve their position in the future. Event detection that detects the onset of new events from news documents is essential to facilitating an organization's environmental scanning activity. Traditional feature-based event detection techniques detect events by comparing the similarity between features of news stories and incur several problems. For example, being a feature-based approach, it cannot capture the genuine properties of an event contained in a news story and cannot support event categorization and news stories filtering. In this study, we developed an information extraction-based event detection (NEED) technique that combines text categorization and information extraction techniques to address the problems inherent to traditional feature-based event detection techniques. Using a traditional feature-based event detection technique (i.e., INCR) as benchmarks, the empirical evaluation results showed that the proposed NEED technique improved the effectiveness of event detection measured by miss and false alarm rates.

Some future research works related to this study should be continued. The detection effectiveness of the NEED technique would be based on accurate and complete extraction rules for each event topic. However, the manual engineering of extraction rules is often time-consuming and error-prone. Thus, a mechanism for learning extraction rules for each event topic is essential to the NEED technique. Furthermore, the experimental data set used to evaluate the NEED technique only comprised news stories across two months and of five event topics. A larger data set with more event topics for empirical evaluation of the proposed technique is desirable. Finally, the lexical and temporal similarity function was employed in

the NEED technique. However, the incorporation of domain knowledge can improve the effectiveness of the proposed technique. For example, two event property values, "IBM" and "International Business Machine" will be evaluated as two completely different values by the existing similarity function. However, with the inclusion of domain knowledge (e.g., a company name may exist in an acronym form) in the similarity function, "IBM" and "International Business Machine" can successfully be evaluated as an equivalent value.

Acknowledgments

This work was supported by National Science Council of the Republic of China under the grant NSC 89-2416-H-110-005.

References

Ahituv, N., Zif, J. and Machlin, I., "Environmental Scanning and Information Systems in Relation to Success in Introducing New Products," *Information and Management* (33), 1998, pp.201-211.

Allan, J., Papka, R. and Lavrenko, V., "On-line New Event Detection and Tracking," *Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp.37-45.

Apté, C., Damerau, F. and Weiss, S., "Automated Learning of Decision Rules for Text Categorization," *ACM Transactions on Information Systems* (12:3), 1994, pp.233-251.

Baker, L. D. and McCallum, A. K., "Distributed Clustering of Words for Text Categorization," *Proceedings of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp.96-103.

Brill, E., "Some Advances in Rule-Based Part of Speech Tagging," *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994.

Choo, C. W., "The Art of Scanning the Environment," *Bulletin of the American Society for Information Science*, 1999, pp.21-24.

Clark, P. and Boswell, R., "Rule Induction with CN2: Some Recent Improvements," *Proceedings of the 5th European Conference (EWSL '91)*, 1991, pp.151-163.

Cohen, W. W. and Singer, Y., "Context-sensitive Learning Methods for Text Categorization," *ACM Transactions on Information Systems* (17:2), April 1999, pp.141-173.

Cowie, J. and Lehnert, W., "Information Extraction," *Communications of the ACM* (39:1), January 1996, pp.80-91.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M., "Inductive Learning Algorithms and Representations for Text Categorization," *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98)*, 1998, pp.148-155.

Eikvil, L., "Information Extraction from World Wide Web: A Survey," Norwegian Computer Center, Report No. 945, July 1999.

Embley, D. W., Campbell, D. M. and Smith, R. D., "Ontology-Based Extraction and

Structuring of Information from Data-Rich Unstructured Documents," *Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management (CIKM '98)*, Bethesda, MD, 1998, pp.52-59.

Larkey, L. and Croft, W., "Combining Classifiers in Text Categorization," *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996, pp.289-297.

Lewis, D. and Ringuette, M., "A Comparison of Two Learning Algorithms for Text Categorization," *Proceedings of Symposium on Document Analysis and Information Retrieval,* 1994.

Masand, B., Linoff, G. and Waltz, D., "Classifying News Stories Using Memory Based Reasoning," *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92)*, 1992, pp.59-64.

Muslea, I., "Extraction Patterns for Information Extraction Tasks: A Survey," *Proceedings of AAAI Conference*, 1999.

Ng, H. T., Goh, W. B. and Low, K. L., "Feature Selection, Perceptron Learning, and A Usability Case Study for Text Categorization," *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, 1997, pp.67-73.

Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

Riloff, E. and Lehnert, W., "Information Extraction as A Basis for High-Precision Text Classification," *ACM Transactions on Information Systems* (12:3), July 1994, pp.296-333.

Schutze, H., Hull, D. A. and Pedersen, J. O., "A Comparison of Classifiers and Document Representations for the Routing Problem," *Proceedings of 18th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995.

Soderland, S., "Learning Information Extraction Rules for Semi-Structured and Free Text," *Machine Learning* (34), 1999, pp.233-272.

Voutilainen, A., "Nptool: A Detector of English Noun Phrases," *Proceedings of Workshop on Very Large Corpora*, Ohio, June 1993.

Yang, Y., "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1994, pp. 13-22.

Yang, Y. and Chute, C. G., "An Example-Based Mapping Method for Text Categorization and Retrieval," *ACM Transactions on Information Systems* (12:3), 1994, pp.252-277.

Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T. and Liu, X., "Learning Approaches for Detecting and Tracking News Events," *IEEE Intelligent Systems* (14:4), July/August 1999, pp.32-43.

Yang, Y., and Liu, X., "A Re-examination of Text Categorization Methods," *Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp.42-49.