PACIS 2011 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

9 July 2011

A Temporal Frequent Itemset-Based Clustering Approach For Discovering Event Episodes From News Sequence

Yen-Hsien Lee
National Chiayi University, yhlee@mail.ncyu.edu.tw

Paul Jen-Hwa Hu
University of Utah, paul.hu@business.utah.edu

Tsai-Hsin Chu
National Chiayi University, thchu@mail.ncyu.edu.tw

Tsang-Hsiang Cheng
Southern Taiwan University, cts@mail.stut.edu.tw

Hsin-Wei Chen Chen
National Chiayi University, s0961412@mail.ncyu.edu.tw

ISBN: [978-1-86435-644-1]; Full paper

A TEMPORAL FREQUENT ITEMSET-BASED CLUSTERING APPROACH FOR DISCOVERING EVENT EPISODES FROM NEWS SEQUENCE

Yen-Hsien Lee, Department of Management Information Systems, National Chiayi University, Chiayi, Taiwan, R.O.C., yhlee@mail.ncyu.edu.tw

Paul Jen-Hwa Hu, Department of Operations and Information Systems, University of Utah, Salt Lake City, UT, USA, paul.hu@business.utah.edu

Tsai-Hsin Chu, Department of E-learning Design and Management, National Chiayi University, Chiayi, Taiwan, R.O.C., thchu@mail.ncyu.edu.tw

Tsang-Hsiang Cheng, Department of Business Administration, Southern Taiwan University, Tainan, Taiwan, R.O.C., cts@mail.stut.edu.tw

Hsin-Wei Chen, Department of Management Information Systems, National Chiayi University, Chiayi, Taiwan, R.O.C., s0961412@mail.ncyu.edu.tw

Abstract

When performing environmental scanning, organizations typically deal with a numerous of events and topics about their core business, relevant technique standards, competitors, and market, where each event or topic to monitor or track generally is associated with many news documents. To reduce information overload and information fatigues when monitoring or tracking such events, it is essential to develop an effective event episode discovery mechanism for organizing all news documents pertaining to an event of interest. In this study, we propose the time-adjoining frequent itemset-based event-episode discovery (TAFIED) technique. Based on the frequent itemset-based hierarchical clustering (FIHC) approach, our proposed TAFIED further considers the temporal characteristic of news articles, including the burst, novelty, and temporal proximity of features in an event episode, when discovering event episodes from the sequence of news articles pertaining to a specific event. Using the traditional feature-based HAC, HAC with a time-decaying function (HAC+TD), and FIHC techniques as performance benchmarks, our empirical evaluation results suggest that the proposed TAFIED technique outperforms all evaluation benchmarks in cluster recall and cluster precision.

Keywords: Event Episode Discovery, Retrospective Event Detection, Event Evolution, Temporal Frequent Itemset-based Clustering.

1 INTRODUCTION

The importance of discovering event episode from online news articles increases significantly as digitalization continues to transform all aspects of business. Event episode discovery has been shown fundamental to firms' environmental scanning as it provides a higher-level abstraction of an essential event worthy of firms' attention and monitoring (Wei & Chang 2005; Nallapati et al. 2004). As online news articles are available from expanding sources and at an accelerating pace, event episode discovery from online news articles is becoming a critical dimension of environmental scanning, hereby enabling firms to conveniently obtain and analyze information pertinent important events, trends, or global environment in a timely manner. By discovering event episodes effectively, a firm can become aware of potentially crucial events that it may overlook otherwise and thus adapt to the fast-changing business environment with agility and appropriate responses (Jennings & Lumpkin 1992; Tan et al. 1998; Liu 2004; Wei & Lee 2004; Granat 2005).

An episode straddles between event and story. In general, an event refers to something that happens in some a particular time and place, whereas a story denotes any type of document containing substantive information content with a unified event focus (Allan et al. 1998b; Yang & Chute 1994 Yang et al. 1999; Nallapati et al. 2004; Wei & Lee 2004; Wei & Chang 2005). Anchored in this lens, an episode of an event is defined as a particular stage or subevent of the focal event (Wei & Chang 2005). Typically, an event has an initiating opening episode and progresses over time; that is, subsequent related episodes are then developed, steered and influenced by key preconditions and consequences of the event. For example, initial discussions of a merger deal involving two publically traded firms and the subsequent investigation by the Federal Trade Commission about its legitimacy represent two distinct episodes (or stages) of the focal event; i.e., the merger under discussion. In this light, a news story or a subset of news stories about an event may describe a specific episode of the event and, if so, their contents should be coherent temporally and topically.

The sheer volume of online news articles available to firms makes the conventional, manual approach for event episode discovery ineffective, if feasible at all. The manual approach that often demands substantial time and processing requirements is tedious and error-prone. In turn, these stringent requirements and limitations favor the use of a system-enabled approach to automatically identify event episode from a large collection of online news articles. Specifically, event episode discovery can be supported by multi-document summarization. Existing multi-document summarization techniques (Goldstein et al. 2000; Wei et al. 2004) do not consider episodes in a sequence of documents. However, if we can effectively identify the episodes described by a sequence of documents, we then can develop an episode-based multi-document summarization technique to select important sentences from the news articles pertinent to major episodes reported by the sequence of documents.

At a nutshell, event episode discovery is similar to retrospective event detection that clusters news articles into distinct groups (or subgroups). One notable difference is that event episode discovery discovers the episodes of an event from a sequence of news articles about that event, whereas retrospective event detection identifies the underlying events, often different, from a sequence of news articles. Existing techniques for retrospective event detection follow and extend appropriate document clustering approaches by taking into consideration unique characteristics of event-based documents. For example, Nallapati et al. (2004) incorporates temporal localization (i.e., news stories that describe the same event tend to be proximate temporally) by using a time-decaying function to adjust the similarity of two stories; that is, the greater the temporal difference between two stories, the lower their similarity. Although preliminary evaluation results suggesting the temporal-based clustering approach outperform the traditional feature-based clustering approach significantly, the document-based time decaying function may not be appropriate for event episode discovery because event episodes usually are associated with different temporal patterns. As a result, lasting episodes may not be accurately depicted by the described time decaying function. Furthermore, the different episodes of an event may exhibit temporal overlaps and documents, though adjacent temporally, may pertain to different episodes. The effectiveness of the document-based time decaying function will deteriorate in either case.

To address the limitation of document-based time decaying function, we suggest the temporal characteristic of event-based news stories be considered and analyzed at a feature level. The rationale is that temporally proximate features should be representative of the underlying event episode; thus, they are more important than features further apart temporally. As a result, we can derive from the temporal proximity characteristic novel features and use them to represent a new episode to be considered for event episode discovery. Our study emphasizes the importance of temporal characteristics of features in event episodes and proposes the time-adjoining frequent itemset-based event-episode discovery (TAFIED) technique. Specifically, our proposed technique extends frequent itemset-based hierarchical clustering (FIHC) technique by considering temporal localization in its fitness measure between a clusters and a document. Since the FIHC technique uses frequent items to group documents (i.e., features appear frequently in documents), our fitness measure for evaluating the goodness of documents within a cluster may reveal the temporal proximity of features indirectly.

Event episode discovery is applicable to other applications. For example, such discovery can facilitate and improve the development of event tracking techniques. Initiated by few news stories regarding a focal event, event tracking attempts to identify subsequent news stories that describe the progression or development of that event (Allan et al. 1998; Yang et al. 1999). Events can be categorized and events of the same category (type) may have a seemingly defined progression pattern comprised of different episodes that follow a temporal or causal sequence (Wei & Chang 2005). By discovering episodes of different events of the same category, we can advance the generalization of the underlying event evolution pattern (through the identified episodes and their associations with respect to the focal event category) and thus better support event tracking (Wei & Chang 2005). For instance, we may reveal from different events pertaining to earthquake (i.e., earthquake event category) that the "the rescue actions" episode usually follows the "the casualty report" episode. If so, when tracking an earthquake event just happened, before any reports about "rescue actions" arrive, news article reporting casualty may not be associated with the earthquake under examination, regardless of the similarity between their contents. Such event evolution patterns, when properly revealed, allow firms to better anticipate and respond to the subsequent development of an event.

The reminder of this paper is organized as follows: Section 2 provides an overview of event episode discovery and reviews relevant previous research that includes the FIHC technique. In Section 3, we detail the overall design of the proposed time-adjoining frequent itemset-based event-episode discovery technique, followed by an empirical evaluation and key results in Section 4. We conclude this study with a conclusion and some future research directions in Section 5.

2 BACKGROUND OVERVIEW AND LITERATURE REVIEW

Firms increasingly reply on news stories for obtaining information about their business environment, such as customers, suppliers, and competitors (Nallapati et al. 2004). Sources of such news articles are shifting from traditional, print-based sources to various Web sites that publish in great quantity and frequency. Previous research has examined the use of online news articles to detect or track new events (Allan et al. 1998a; Allan et al. 1998b; Allan et al. 2002; Makkonen 2003; Makkonen et al. 2004; Papka 1999; Wei & Lee 2004; Yi 2005; Yang et al. 2005 Yang et al. 2002). Despite the availability of techniques for event detection or tracking, firms still need to process a large number of news articles in their environmental scanning partly because of the proliferation of online news sources. For example, a query to the Google News (http://news.google.com) concerning the 2010 Haiti earthquake returns more than 26, 000 news stories just within an one-month span (January 12 through February 11, 2010). The sheer volume of articles represent a common challenge to firms that strive to stay abreast of important events as they develop over time, considering the number of articles to sift through. Conceivably, firms need a summary of each event as it progresses through different episodes (stages). Continued with the 2010 Haiti earthquake example, this event may have several important episodes that may include "the happening of the earthquake", "rescue and search", and "aids from the world".

In the following, we first review prior research related to event episode discovery in general and then brief review the frequent itemset-based hierarchical clustering technique we based on in this study in specific.

2.1 Research Related to Event Episode Discovery

Event episode discovery have been studied by Nallapati et al. (2004) and Wei & Chang (2005); both use the hierarchical agglomerative clustering (HAC) algorithm (Voorhees 1986) to identify episodes from a sequence of news articles of an event. Specifically, Nallapati et al. (2004) attempts to discover the event structure as inter-connected threading subjects by defining an event taxonomy: Story→Event→Topic. An event may have causal dependency with other events. “Topic” and “Event” in this taxonomy correspond to “Event” and “Episode” in our study, respectively. To discover event episodes, Nallapati et al. (2004) takes advantage of temporal localization of news stories; that is, when estimating the similarity of two news stories, a time decaying function is applicable and penalize pair of stories if they are apart temporally. Preliminary empirical results show that the time decaying function can improve the effectiveness of an existing event episode discovery technique.

Wei & Chang (2005) proposes another taxonomy, Story→Episode→Event, and builds intra-sequence and inter-sequence episode relationship. They attempt to discover event episodes and to capture event evolution by discovering the temporal pattern of the different episodes of an event. Unlike Nallapati et al. (2004), temporal localization of news stories is not considered, mainly because they aim at generalizing event episodes cross different events as well as discovering frequent event episodes and their underlying temporal relationships.

Research examining retrospective event detection (cluster detection) also relates to event episode discovery (Kumaran & Allan 2004; Kumaran & Allan 2005; Zhang et al. 2007; Yang et al. 2002). Topic detection and tracking (TDT) targets event-based organization of broadcast news, using a concrete set of evaluation-driven tasks related to general problem of identifying coherent topics (an event of interest) in a constantly expanding chronologically-ordered news stories obtained from multiple media sources and in different languages (Allan et al. 2005; Allan et al 2002). Retrospective event detection partitions (or clusters) all the news stories in a source corpus into topics (such as events). Retrospective event detection shares similar characteristics with event episode discovery but differs in the unit and granularity of analysis. For example, event episode discovery forms clusters (i.e., episodes) related to a focal event and retrospective event detection identifies events from a stream of news stories by segmenting different events described by these news stories. By and large, event episode discovery perform analyses at finer-grained level than does retrospective event detection. For example, retrospective event detection may identify the event of Haiti earthquake and all news articles related to it; however, the purpose of event episode discovery shall focus on discovering of the potential development stages of an event and the news articles pertaining to each stage, e.g. the news articles related to the stage of rescue and search.

2.2 Frequent Itemset-based Hierarchical Clustering Technique

Fung et al. (2003) proposed the frequent itemset-based Hierarchical clustering (FIHC) technique, which follows association rule mining by considering documents as transactions and features (of a document) as items. This technique identifies items of which document frequency exceeds a prespecified minimum support (g_f) and uses the identified frequent items as cluster centroids to perform document grouping. Specifically, it takes the identified frequent items as cluster labels and initially assigns each document to a set of candidate clusters according to its own frequent items, then determines the most appropriate cluster for each document d_j by evaluating the goodness score of the remaining document d_j in the cluster c_x. The goodness score is calculated as follows:

$$Score(c_x \leftarrow d_j) = [\sum_{i}(n(i) \times Cluster\_Support(i))] - [\sum_{i'}(n(i') \times Global\_Support(i'))] \qquad (1)$$

where i represents a global frequent item in document d_j as well as a frequent item in cluster c_x, i' represents a global frequent item in document d_j but not a frequent item in c_x, $n(i)$ denotes the weight of item i in document d_j, and $n(i')$ is the weight of item i' in document d_j.

The FIHC technique keeps document in the cluster that has the highest goodness score. When completed, each document belongs to one and only one cluster, with empty clusters removed. To avoid documents that pertain to the same topic (e.g., event) but are distributed over several clusters, the FIHC uses an inter-cluster similarity measure, defined as follows, to evaluate the similarity between clusters and merges the clusters when the similarity exceed a prespecific threshold (e.g., 1), hereby generating a natural topic hierarchy for ease of browsing and increased cluster accuracy.

$$Inter\_Sim\ (c_x \leftrightarrow c_y) = \sqrt{Sim(c_x \leftarrow c_y) \times Sim(c_y \leftarrow c_x)} \qquad (2)$$

The FIHC measures inter-cluster similarity by assessing the goodness of merging cluster c_y in c_x by aggregating all documents in c_y into a document as well as the goodness of merging cluster c_x with c_y by aggregating all documents in c_x into a document. Specifically, the goodness score of merging cluster c_y in c_x is defined as:

$$Sim(c_x \leftarrow c_y) = \frac{Score(c_x \leftarrow doc(c_y))}{\Sigma\ n(i) + \Sigma n(i')} + 1 \qquad (3)$$

where c_x and c_y are two clusters, $doc(c_y)$ represents the aggregation of all the documents in cluster c_y, i represents a global frequent item in $doc(c_y)$ as well as a frequent item in cluster c_x, i' represents a global frequent item in $doc(c_y)$ but not a frequent item in c_x, $n(i)$ denotes the weight of item i in $doc(c_y)$, and $n(i')$ is the weight of item i' in $doc(c_y)$.

3 TIME-ADJOINING FREQUENT ITEMSET-BASED EVENT EPISODE DISCOVERY

We propose time-adjoining frequent itemset-based event-episode discovery (TAFIED) technique by extending the FIHC because of its capability of dealing with the burst characteristic of features for event episode discovery. Our proposed technique addresses the limitations of the document-based time-decaying function commonly used in traditional feature-based document clustering techniques by properly considering the characteristics of online news articles, i.e., burst, novelty, and temporal locality. The proposed TAFIED groups documents by using frequent items as cluster centriods. It further considers temporal localization of documents in a cluster by revising the goodness measure between a cluster and a document by weighing the adjacentness of the time stamps of news articles pertaining to the same cluster (i.e., event episode). As a result, the TAFIED can generate clusters of which documents are likely to share frequent items appearing with temporal adjacency in a stream of news articles. As shown in Figure 1, the TAFIED technique takes as input a set of news articles pertaining to an event and discovers a set of associated episodes. The overall processing of the TAFIED consists of four phases: document preprocessing, cluster initialization, cluster distinction, and cluster adjustment, detailed as follow.

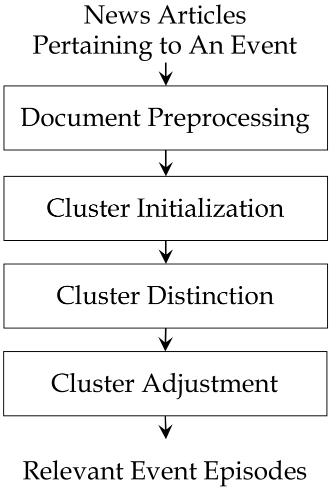

Figure 1. Overall Processing of the Proposed TAFIED Technique

Document preprocessing. In the document preprocessing phase, the TAFIED extracts meaningful terms (such as nouns, noun phrases, and verbs) from each news article. It applies a rule-based part-of-speech tagger to tag each word in a news article (Brill 2002; Brill 2004) and then employs a parser to select nouns, noun phrases, and verbs from the article. Stop words (i.e., non-semantic-bearing words) are removal and the remaining words are stemmed into their prototype.

Cluster initialization. In the cluster initialization phase, the TAFIED constructs a set of initial clusters and assigns each document to candidate clusters, on the basis of its frequent items. Specifically, frequent items (i.e., terms) are first identified from the entire news articles under analysis. The TAFIED determines term t_i as a frequent item if its document frequency (i.e., the number of news articles having term t_i) over the total number of news articles exceeds a prespecified minimum global support g_t. By taking each frequent item as a class label, a set of initial clusters are created and each news article is assigned to the candidate clusters, on the basis of its frequent items (i.e., class labels). That is, a news article can be labeled as members of multiple clusters; we may have as many clusters as the number of frequent items we identify. Let's assume that we have ten news articles, the identified frequent items, with g_t = 0.4, and their respective term frequency in each document are shown in Table 1.

	t_1	t_2	t_3	t_4	t_5
d_1	3	—	2	—	—
d_2	5	—	2	2	—
d_3	7	—	—	3	4
d_4	1	—	1	4	—
d_5	2	—	—	5	—
d_6	—	2	—	—	2
d_7	—	3	—	—	—
d_8	—	1	16	—	2
d_9	—	1	—	—	1
d_{10}	2	—	12	—	—

Table 1. Term Frequency of Frequent Items in Each Document

In the cluster initialization phase, five initial clusters are created, with five identified frequent items (i.e., t_1 to t_5) as cluster labels, respectively. Then, a document is assigned repeatedly to the corresponding clusters, according to its frequent items. For example, document d_1 is assigned to clusters t_1 and t_3, and document d_2 to clusters t_1, t_3, *and* t_4. Finally, we have a set of initial clusters with their respective member documents, as shown in Table 2.

Initial Set of Clusters	Member Documents
c_{t1}	d_1, d_2, d_3, d_4, d_5, d_{10}
c_{t2}	d_6, d_7, d_8, d_9
c_{t3}	d_1, d_2, d_4, d_8, d_{10}
c_{t4}	d_2, d_3, d_4, d_5
c_{t5}	d_3, d_6, d_8, d_9

Table 2. Initial Clusters and Their Respective Member Documents

Cluster distinction. After the cluster initialization phase, each document is assigned to at least one candidate cluster. Because we assume that each news article pertains to one and only one event episode, the TAFIED, in the cluster distinction phase, evaluates the fitness of a document with respect to each candidate cluster and selects the most appropriate cluster, hereby producing the final set of clusters. We propose a fitness function to measure the likelihood that a document d_j belongs to a cluster c_x. By considering the temporal characteristic of a stream of news articles, we expect features (i.e., terms) of news articles that describe the same event episode should in principle exhibit characteristics such as burst, novelty, and temporal proximity, as well as sharing more important features. That is, news articles pertaining to a particular event episode should share many features appearing frequently and consecutively in news articles about that episode rather than in articles pertaining to another event episode. Formally, we define the fitness function between a document d_j and a cluster c_x as follows:

$$Fitness(c_x \leftarrow d_j) = \sum_{i=1}^{|F|} (\alpha \times CS(f_i, c_x) \times TFIDF(f_i, d_j)) \times TL(c_x) \qquad (4)$$

where F is a set of frequent items, f_i denotes a frequent item, $TFIDF(f_i, d_j)$ represents the within-document term frequency × inverted document frequency for frequent item f_i appearing in document d_j, $CS(f_i, c_x)$ is the cluster support calculated as the percentage of documents in c_x that contain f_i, α is a parameter to control the impact direction of cluster support, and $TL(c_x)$ is a temporal locality function for measuring temporal cohesion (or consecution) of documents when document d_j is assigned to cluster c_x.

Specifically, we use the TF×IDF measure to assess the novelty of feature f_i in the entire news articles, because a lower document frequency can get a higher TF×IDF value. Furthermore, the feature f_i is considered important to cluster c_x if it appears frequently in many documents of c_x. The Rationale is that a document that shares more important features with other documents in the same cluster should have a higher possibility of belonging to that cluster. Inversely, the possibility can be reduced if the document shares many unimportant features with other documents in c_x. Thus, we set α to 1 if $CS(f_i, c_x)$ is larger than or equal to a prespecified significance threshold c_t; and set α to -1 otherwise.

In the fitness function, we further consider temporal proximity of the documents in cluster c_x when assigning document d_j to a cluster. The fitness score should be reduced while assigning document d_j to a cluster c_x likely will widen the temporal difference between documents in that cluster. We therefore propose a temporal locality function $TL(c_x)$, which is defined as $\frac{e^{\lambda-\theta}}{1+e^{\lambda-\theta}}$ if $|c_x| > 1$ and 0.5 otherwise, where $\lambda = (|c_x|-1) \times w^2$ is the theoretically maximal temporal difference allowed between two time-ordered documents in cluster c_x, in which w is a parameter denoting the tolerant temporal difference between two time-ordered documents in c_x, and $\theta = \sum_{k=1}^{|c_x|-1} |t(d_k)-t(d_{k+1})|^2$ is the sum of the actual temporal difference between two time-ordered documents in c_x, in which $t(d_k)$ is the time stamp of document d_k in cluster c_x. The value of $TL(c_x)$ ranges between 0 and 1; a smaller temporal difference between the documents in c_x implies that these documents appear in close temporal adjacency or consecutively, and thus will result in a higher value of $TL(c_x)$. In addition, the value of $TL(c_x)$ decreases gradually as θ increases from 0 but decrease sharply as θ exceeds a threshold value. For example, with a cluster of five documents and w set to 2 to have λ = 16, we can observe the variance of $TL(c_x)$ value with the increase of θ in Figure 2. As shown, the value of $TL(c_x)$ decreases gradually as θ increases from 0 to 12, and decreases sharply as θ increases further. The value of $TL(c_x)$ reaches 0.5 when θ = 16; that is $\theta = \lambda$.

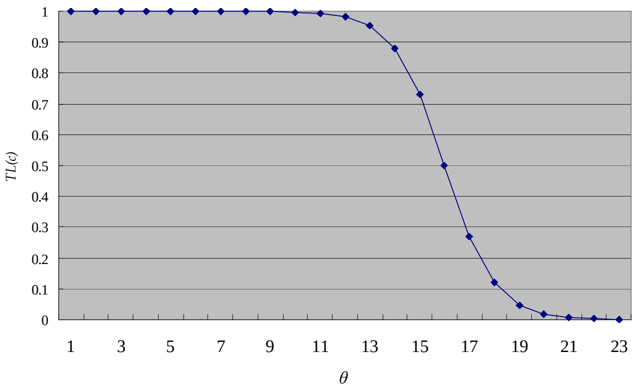

Figure 2. Variance of Temporal Locality Function

To demonstrate how to calculate the fitness score, we use as an example the initial set of clusters in

Table 2. We set the significance threshold c_t to 0.3 and the tolerant time gap w to 2, and calculate the fitness score of document d_3 to each of its candidate clusters as follows.

$$Fitness(c_{t1} \leftarrow d_3) = \left(\left(1\times\frac{6}{6}\times 7\times\log_2\frac{10}{6}\right)+\left(1\times\frac{4}{6}\times 3\times\log_2\frac{10}{4}\right)+\left(-1\times\frac{1}{6}\times 4\times\log_2\frac{10}{4}\right)\right)\times\frac{e^{20-29}}{1+e^{20-29}} = 0.001$$

where $\lambda_{t1} = |6\text{-}1|\times 2^2 = 20$ and $\theta_{t1} = |2\text{-}1|^2+|3\text{-}2|^2+|4\text{-}3|^2+|5\text{-}4|^2+|10\text{-}5|^2 = 29$

$$Fitness(c_{t4} \leftarrow d_3) = \left(\left(1\times\frac{4}{4}\times 7\times\log_2\frac{10}{6}\right)+\left(1\times\frac{4}{4}\times 3\times\log_2\frac{10}{4}\right)+\left(-1\times\frac{1}{4}\times 4\times\log_2\frac{10}{4}\right)\right)\times\frac{e^{12-3}}{1+e^{12-3}} = 7.802$$

where $\lambda_{t4} = |4\text{-}1|\times 2^2 = 12$ and $\theta_{t4} = |3\text{-}2|^2+|4\text{-}3|^2+|5\text{-}4|^2 = 3$

$$Fitness(c_{t5} \leftarrow d_3) = \left(\left(-1\times\frac{1}{4}\times 7\times\log_2\frac{10}{6}\right)+\left(-1\times\log_2\frac{1}{4}\times 3\times\frac{10}{4}\right)+\left(1\times\frac{4}{4}\times 4\times\log_2\frac{10}{4}\right)\right)\times\frac{e^{12-14}}{1+e^{12-14}} = 0.358$$

where $\lambda_{t5} = |4\text{-}1|\times 2^2 = 12$ and $\theta_{t5} = |6\text{-}3|^2+|8\text{-}6|^2+|9\text{-}8|^2 = 14$

The fitness score of d_3 with respect to candidate clusters c_{t1}, c_{t4}, and c_{t5} are 0.001, 7.802, and 0.358, respectively. We therefore remain document d_3 in cluster c_{t4}.

Distinct Set of Clusters	Member Documents
c_{t2}	d_6, d_7, d_9
c_{t3}	d_1, d_8, d_{10}
c_{t4}	d_2, d_3, d_4, d_5

Table 3. Distinct Clusters and Respective Member Documents

Cluster adjustment. Documents pertaining to a cluster may be assigned to different clusters if multiple dominant frequent items appear in these documents. Therefore, the TAFIED, in the cluster adjustment phase, merges clusters of which documents are highly relevant or similar. A combined cohesion measure is used to evaluate the appropriateness of merging two clusters. The combined cohesion of two clusters c_x and c_y is calculated as follow:

$$Combined\text{-}Cohesion(c_x \leftrightarrow c_y) = \sqrt{Cohesion(c_x \leftarrow c_y)\times Cohesion(c_y \leftarrow c_x)}\times TL(c_x \cup c_y) \qquad (5)$$

The TAFIED considers merging two clusters if their combined cohesion score exceeds the specified merging threshold η (e.g., 1). To measure the cohesion of two clusters, the cluster to be merged (e.g., c_y) is considered as a document and its fitness with respect to the other cluster (e.g., c_x) is then calculated. A cohesion function is developed by extending the fitness function for assessing the fitness of a document and a cluster (used in the cluster distinction phase), and normalizes its output value to be between 0 and 2, in order to avoid any negative values. Formally, the cohesion function is defined as:

$$Cohesion(c_x \leftarrow c_y) = \frac{\sum_{i=1}^{|F|}(\alpha\times CS(f_i, c_x)\times \sum_{d_j\in c_y} TFIDF(f_i, d_j))}{\sum_{i=1}^{|F|}\sum_{d_j\in c_y} TFIDF(f_i, d_j)} + 1 \qquad (6)$$

where c_x and c_y are clusters to be considered for merging, $\sum_{d_j\in c_y} TFIDF(f_i, d_j))$ is the sum of TF×IDF values of f_i in each document d_j in cluster c_y, and $\sum_{i=1}^{|F|}\sum_{d_j\in c_y} TFIDF(f_i, d_j))$ is the sum of TF×IDF values of all frequent items in each document in cluster c_y.

As an illustration, let's calculate the combined cohesion score of merging clusters c_{t3} and c_{t4} in Table 3.

$Cohesion(c_{t3} \leftarrow c_{t4}) = 1.321 =$

$$\frac{(\frac{2}{3}\times(5+7+1+2)\times\log\frac{10}{6}\times 1)+(\frac{3}{3}\times(2+1)\times\log\frac{10}{5}\times 1)+(\frac{0}{3}\times(2+3+4+5)\times\log\frac{10}{4}\times(\text{-}1))+(\frac{1}{3}\times(4)\times\log\frac{10}{4}\times 1)}{(5+7+1+2)\times\log\frac{10}{6}+(2+1)\times\log\frac{10}{5}+(2+3+4+5)\times\log\frac{10}{4}+(4)\times\log\frac{10}{4}}+1$$

$Cohesion(c_{t4}\leftarrow c_{t3}) = 1.478 =$

$$\frac{(\frac{4}{4}\times(3+2)\times\log\frac{10}{6}\times 1)+(\frac{0}{4}\times(1)\times\log\frac{10}{4}\times(\text{-}1))+(\frac{2}{4}\times(2+16+12)\times\log\frac{10}{5}\times 1)+(\frac{1}{4}\times(2)\times\log\frac{10}{4}\times(\text{-}1))}{(3+2)\times\log\frac{10}{6}+(1)\times\log\frac{10}{4}+(2+16+12)\times\log\frac{10}{5}+(2)\times\log\frac{10}{4}}+1$$

$$Combined\text{-}Cohesion(c_{t3}\leftrightarrow c_{t4}) = \sqrt{1.321\times 1.478}\times\frac{e^{24\text{-}17}}{1+e^{24\text{-}17}} = 1.396$$

4 EMPIRICAL EVALUATION AND RESULTS

We empirically evaluate the proposed technique together with three prevalent techniques for benchmark purposes: FIHC, HAC (a traditional feature-based document clustering technique), and HAC with a time-decaying function (HAC+TD). In the following, we detail the document corpus we used, valuation design, parameter-tuning experiments, and comparative evaluation results.

Data collection. We evaluated the effectiveness of the proposed TAFIED with a set of events with known episodes and associated news documents. Specifically, we used the event corpus provided by Nallapati et al. (2004), which includes a total of 248 event episodes associated with 53 events and 1,468 relevant news stories selected from TDT2 and TDT3 corpora. In this event corpus, the number of relevant news stories pertaining to each event is not particularly large and is balanced across all the 53 events. The average length of news documents is 64.2 words and the average number of features identified for each event after the feature extraction phase is about 520. Table 4 provides a summary of the event corpus used in our evaluation.

	Average	**Minimum**	**Maximum**
Number of Stories per Event	27.7	16	30
Number of Stories per Episode	5.92	1	25
Number of Episodes per Event	4.68	2	8
Duration of Episode	8.32	1	103
Duration of Event	31.55	2	138

Table 4. Summary of Event Corpus

Evaluation design. We took a comparative approach to evaluate the effectiveness of the proposed TAFIED technique by comparing its performance with that of HAC, FIHC, and HAC+TD. The time decaying similarity function is defined as follows:

$$sim_{time\text{-}decaying}(d_i, d_j) = sim(d_i, d_j)\times\exp\left(-\frac{t(d_j)-t(d_i)}{T}\right) \qquad (7)$$

where $sim(d_i, d_j)$ denotes the cosine similarity between d_i and d_j, $t(d_i)$ is the timestamp of ith document, T is the time interval between the first document and the last document in the time-ordering document sequence pertaining to the focal event (i.e., $t(d_{|S|}) - t(d_1)$), in which $|S|$ is the total number of documents describing that event.

We used cluster recall and cluster precision to measure the effectiveness of each technique for discovering event episodes. The cluster recall and cluster precision, which anchor at the association of documents from the same cluster (or episode), are fundamental to the performance measure of document clustering techniques (Roussinov & Chen 1999; Wei et al. 2006; Wei et al. 2009). Given each event in our evaluation corpus, assume the known event episodes be the true event episodes of the target event. The cluster recall (*CR*) and cluster precision (*CP*) pertaining to the target event are respectively defined as $CR = \frac{|CA|}{|TA|}$ and $CP = \frac{|CA|}{|GA|}$, where *TA* is the set of associations of documents in the true event episodes, *GA* is the set of associations of documents in the event episodes generated by a technique investigated, and *CA* is the set of associations of documents that exists in both true and generated event episodes.

For illustration, let's consider a sequence of documents $S = <d_1, d_2, d_3, d_4, d_5, d_6, d_7>$ pertaining to an event. Let *S* be classified into three true event episodes, EP_1, EP_2, and EP_3, where $EP_1 = \{d_1, d_2\}$, $EP_2 = \{d_3, d_4, d_5, d_6\}$, and $EP_3 = \{d_7\}$. Accordingly, there exists seven associations of documents, including $\{(d_1, d_2), (d_3, d_4), (d_3, d_5), (d_3, d_6), (d_4, d_5), (d_4, d_6), (d_5, d_6)\}$ in the true event episodes. On the other hand, let the event episodes identified by an investigated technique be G_1 and G_2, where $G_1 = \{d_1, d_2, d_3\}$ and $G_2 = \{d_4, d_5, d_6, d_7\}$. Correspondingly, nine associations of documents, including $\{(d_1, d_2), (d_1, d_3), (d_2, d_3), (d_4, d_5), (d_4, d_6), (d_4, d_7), (d_5, d_6), (d_5, d_7), (d_6, d_7)\}$ exist in the generated event episodes. Accordingly, four associations of documents, including $\{(d_1, d_2), (d_4, d_5), (d_4, d_6), (d_5, d_6)\}$ exist in both true and generated event episodes. Hence, we can have $CR = 4/7$, and $CP = 4/9$.

Once the cluster recall and cluster precision are attained for each event in our evaluation corpus, we then apply the weighted average method to measure the overall effectiveness across all events. To examine the trade-off between cluster precision and cluster recall, we employed the precision/recall trade-off (PRT) curve (Gordon & Kochen, 1989; Buckland & Gey, 1994; Manning & Schutze, 1999; Wei et al., 2006, 2009), which reveals the effectiveness of a technique with various merging threshold values, i.e., the intercluster similarity threshold for HAC and HAC+TD, and the cluster merging threshold for FIHC and TAFIED. For the HAC and HAC+TD techniques, we examined merging threshold between 0 and 1, in increments of 0.02. Besides, for the FIHC and TAFIED technique, the value of combined cohesion ranges from 0 to 2 and two clusters are suggested to be merged while it is larger than 1. Thus, we examined the merging threshold between 1 and 2, in increments of 0.02. In general, PRT curves close to the upper-right corner are more desirable than those near the point of origin.

Parameter tuning experiments. We performed a series of experiments to select appropriate values for the parameters essential to each investigated technique, using a random sample of news articles pertaining to 10 events. For the TAFIED technique, we needed to tune several parameter values, such as minimum global support (g_t), significance threshold for cluster support (c_t), and the tolerant time gap (*w*). We examined g_t over the range from 0.02, 0.05, 0.1 to 0.5 (increments of 0.1), c_t between 0 and 1.0 (increments of 0.1), and *w* ranging from 1.5 to 4 (increments of 0.5). Overall, the TAFIED technique seemed most effective with $g_t = 0.02$, $c_t = 0.2$, and $w = 2.0$. We adopted these parameter values for our subsequent experiments.

For the FIHC, we needed to tune its minimum global support (g_f) and significance threshold for cluster support (c_f). We assessed values of g_f over the range from 0.02, 0.05, 0.1 to 0.5 (increments of 0.1), and c_f between 0 and 1.0 (increments of 0.1). According to our experimental results, the FIHC seemed to perform best with $g_f = 0.02$ and $c_f = 0.8$. For the HAC, we chose the TF×IDF as the feature selection metric and tuned two important parameters: the number of features (k_h) and the document representation scheme (r_h). We examined the effects of k_h ranging from 50 to 250 (increments of 50) and r_h by the binary and TF×IDF schemes. According to the experimental results, we set k_h as 150 and selected the binary scheme for r_h, which appeared most appropriate. Furthermore, because the parameters that need to be tuned for HAC+TD are identical to that of HAC, we therefore followed the same procedure for HAC parameter tuning. We finally adopted 150 and TF×IDF scheme as the number of features and the document representation scheme for HAC+TD for performance reasons.

Results. We empirically examined the effectiveness of our proposed TAFIED in comparison with the benchmarks, i.e., FIHC, HAC and HAC+TD techniques. We perform the best-versus-best comparison using the most appropriate parameter values selected for each technique examined. As shown in Figure 3, our proposed TAFIED technique achieves obviously greater effectiveness in discovering event episodes than all other techniques across all specific merging thresholds. In addition, the traditional feature-based technique HAC which did not consider the characteristics of news articles has achieved the less effectiveness in event episode discovery. On the other hand, the HAC+TD, which considered the temporal localization by using a time decaying function to adjust the similar between news articles, arrived at the better performance than the HAC and FIHC techniques which do not consider the temporal characteristic of news articles. This experiment result has responded to prior research that the time decaying function showed its usefulness in improving the effectiveness of HAC technique in discovering event episodes. Finally, the frequent-itemset-based technique FIHC, which based on frequent items to cluster news articles, has to some degree considered the burst of features (words) in the same event episodes and made the documents share more important features to be grouped together, and therefore, resulted in the better performance than the feature-based HAC technique. Overall, our evaluation results suggest that the proposed TAFIED technique that considers the temporal characteristics of news articles noticeably improves the effectiveness of discovering event episodes.

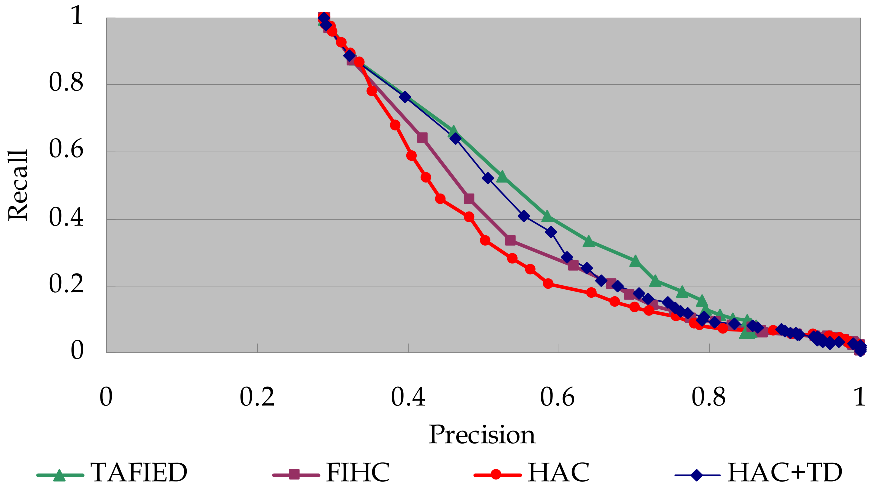

Figure 3. Comparative Evaluation Results

5 CONCLUSION AND FUTURE RESEARCH DIRECTIONS

When performing environment scanning, organizations typically deal with a numerous of events and topics about their core business, relevant technique standards, competitors, and market, among many others, where each event or topic to monitor or track generally is associated with many news documents. To reduce such information overload and information fatigues when monitoring or tracking events, it is essential to develop an effective event episode discovery mechanism to organize all news documents pertaining to an event of interest.

In this study, we propose time-adjoining frequent itemset-based event-episode discovery technique (TAFIED). Using the traditional feature-based clustering approach HAC, HAC with a time-decaying function, and the frequent itemset-based clustering approach FIHC as performance benchmarks, the empirical evaluation results suggest our proposed TAFIED technique outperforms its benchmark in cluster recall and cluster precision. In addition, our result reveals that time decaying function can benefit to the effectiveness of event episode discovery.

Some future research works related to this study are highlighted as follows. First, the news articles we adopted for the empirical evaluation consisted of only 53 events. To generalize the conclusions of our research study, it is required for collecting additional news events and their respective news articles and performs the evaluation on another set of event corpus. Second, a news article has been assumed pertaining to one and only one episode in this study. In effect, a news article can possibly cover the development of the event across multiple episodes. Therefore, the extension of our event episode discovery technique for dealing with multi-episode news documents is desirable. Finally, as mentioned, event episode discovery can be the initiation of developing multi-document summarization techniques. As a result, it is both interesting and desirable to extend the scope of this research to develop an episode-based multi-document summarization technique on the basis of the proposed event episode discovery technique.

References

Allan, J. Papka, R. Lavrenko, V. (1998b). Online new event detection and tracking. In Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37-45, ACM Press, Melbourne, Australia.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998a). Topic detection and tracking pilot study: Final report. In proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194-218.

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., and Amstutz, P. (2005). Taking topic detection from evaluation to practice. In Proceedings of the 32th Annual Hawaii International Conference on System Sciences (HICSS), Big Island, Hawaii.

Allan, J., Lavrenko, V., and Swan, R. (2002). Explorations within topic tracking and detection. In Topic Detection and Tracking: Event-based Information Organization, James Allan (Ed.), Kluwer Academic Publishers, pp. 197-224.

Beil, F., Ester, M. and Xu, X. (2002). Frequent term-based text clustering. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 436-442.

Brill, E. (1992). A simple rule-based part of speech tagger. In Proceedings of the Third Conference on Applied Natural Language Processing: Association for Computational Linguistics, pp. 152-155, Trento, Italy.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), pp. 722-727, AAAI Press, Menlo Park, CA.

Fung, B., Wang, K., and Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In Proceedings of The SIAM International Conference on Data Mining (SDM’ 03), pp. 59-70.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In Proceedings of NAACL-ANLP 2000 Workshop on Automatic Summarization, Association for Computational Linguistics, pp. 40-48, Seattle, WA.

Granat, J. (2005). Event mining based on observations of the system. Journal of Telecommunications and Information Technology, (3), 87-90.

Jennings, D. and Lumpkin, J. (1992). Insights between environmental scanning activities and porter’s generic strategies: An empirical analysis. Journal of Management, 18 (4), 791-803.

Kumaran, G. and Allan, J. (2004). Text classification and named entities for new event detection. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Sheffield, United Kingdom.

Kumaran, G. and Allan, J. (2005). Using names and topics for new event detection. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 121-128, Vancouver, British Columbia, Canada.

Liu, R.L. (2004). Collaborative multiagent adaptation for business environmental scanning through the Internet. Applied Intelligence, 20 (2), 119-133.

Makkonen, J. (2003). Investigations on event evolution in TDT. In Proceedings of HLT-NAACL 2003 Student Workshop, pp. 43-48.

Makkonen, J., Ahonen-Myka, H., and Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. Information Retrieval, 7 (3-4), 347-368.

Nallapati, R.M., Feng, A., Peng, F., and Allan, J. (2004). Event threading within news topics. In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management （CIKM 2004）, pp.425-432, Washington, D.C..

Papka, R. (1999). On-line new event detection, clustering, and tracking,” Unpublished Doctoral Dissertation, University of Massachusetts, Amherst, MA.

Roussinov, D.G. and Chen, H. (1999). Document clustering for electronic meetings: An experimental comparison of two techniques. Decision Support Systems, 27, 67-79.

Tan, S., Teo, H.H., Tan, B. and Wei, K. (1998). Environmental scanning on the Internet. In Proceedings of The International Conference on Information Systems, pp. 76-87, Helsinki.

Voorhees, E.M. (1986). Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management, 22 (6), 465-476.

Wei, C.P. and Chang, Y.S. (2007). Discovering event evolution patterns from document sequences. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 37 (2), 273-283.

Wei, C.P. and Lee, Y.H. (2004). Event detection from online news documents for supporting environmental scanning. Decision Support Systems, 36, 385-401.

Wei, C.P., Chiang, R.H.L., and Wu, C.C. (2006). Accommodating individual preferences in the categorization of documents: A personalized clustering approach. Journal of Management Information Systems, 23 (2), 173-201.

Wei, C.P., Hu, P., and Lee, Y.H. (2009). Preserving user preferences in automated document-category management: An evolution-based approach. Journal of Management Information Systems, 25 (4), 109-143.

Wei, C.P., Wu P.F., and Lee, Y.H. (2004). Use of text summarization for supporting event detection. In Proceeding of the Eighth Pacific Asia Conference on Information Systems (PACIS), pp. 1098-1111, Shanghai, China.

Yang, Y. and Chute, C.G. (1994). An example-based mapping method for text categorization and retrieval. ACM Transaction on Information Systems, 12 (3), 252-277.

Yang, Y., Carbonell, J.G., Brown, R.D., Pierce, T., Archibald, B.T. and Liu, X. (1999). Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14 (4), 32-43.

Yang, Y., Pierce, T. and Carbonell, J.G. (1998). A study on retrospective and on-line event detection. In Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.28-36, ACM Press, Melbourne, Australia.

Yang, Y., Zhang, J., Carbonell, J., and Jin, C. (2002). Topic-conditioned novelty detection. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (CIKM'02), pp. 688-693, Edmonton, Alberta, Canada.

Yi, J. (2005). Detecting buzz from time-sequenced document streams. In Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE-05), pp. 347-352, Hong Kong.

Zhang, K., Li, J.Z., and Wu, G. (2007). New event detection based on indexing-tree and named entity. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215-222, Amsterdam, Netherlands.