11-2006

# A model for Anticipatory Event Detection

Qi HE

Kuiyu CHANG
*Nanyang Technological University*

Ee Peng LIM
*Singapore Management University*, eplim@smu.edu.sg

**DOI:** https://doi.org/10.1007/11901181_14

# A Model for Anticipatory Event Detection

Qi He, Kuiyu Chang, and Ee-Peng Lim

School of Computer Engineering,
Nanyang Technological University, Singapore 639798, Singapore
qihe@pmail.ntu.edu.sg, kuiyu.chang@pmail.ntu.edu.sg, aseplim@ntu.edu.sg

**Abstract.** Event detection is a very important area of research that discovers new events reported in a stream of text documents. Previous research in event detection has largely focused on finding the first story and tracking the events of a specific topic. A topic is simply a set of related events defined by user supplied keywords with no associated semantics and little domain knowledge. We therefore introduce the Anticipatory Event Detection (AED) problem: given some user preferred event transition in a topic, detect the occurence of the transition for the stream of news covering the topic. We confine the events to come from the same application domain, in particular, mergers and acquisitions. Our experiments showed that classical cosine similarity method fails for the AED task, whereas our conceptual model-based approach, through the use of domain knowledge and named entity type assignments, seems promising. We show experimentally that an AED voting classifier operating on a vector representation with name entities replaced by types performed AED successfully.

## 1 INTRODUCTION

Anticipatory Event Detection (AED)[1] refers to the problem of detecting the occurrence of a user-specified anticipatory event (AE). AED is a very hard problem since it requires a basic understanding of the AE semantics, which can vary by event type. Current news alert systems such as Google News Alerts[2] typically produce abysmal results for AED. For example, the search terms "China attacks Taiwan" (describing an AE that has not happened as of this writing) will generate numerous false alarm articles from Google News Alerts.

One way to look at AED is to think of it as finding the transition between two adjacent events in an *event transition graph* whose events are represented by news articles covering the event transition graph before and after a particular transition has consummated. Figure 1 shows an *event transition graph* with $n$ events and $n-1$ transitions for $topic_i$ (e.g. eBay buys Skype). A user may only be interested in receiving a notification when a particular transition has fired, and not be bothered about the remaining transitions. If sufficient number of news articles can be collected for each of the events, it would be theoretically possible to detect any of the $n-1$ transitions. In order to learn a particular transition, a model will have to be trained to classify articles as occurring "before" or "after"

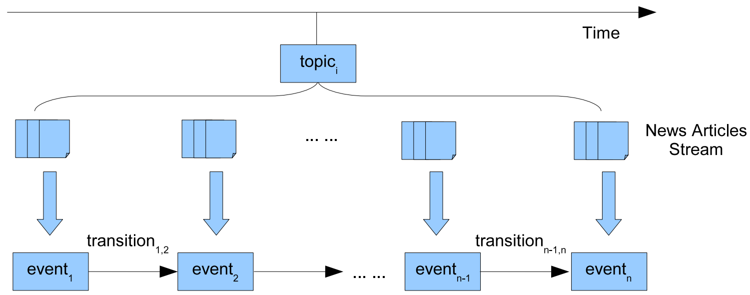


**Fig. 1.** Anticipatory event transition graph.

the transition. For example, given $transition_{1,2}$ in the *event transition graph* of Figure 1, we would like to detect the first story of $event_2$.

In this paper, we report new results on modeling the AED problem. To simplify the problem, we assume that 1) the topic is constrained to a particular domain, i.e. mergers and acquisitions, 2) the *event transition graph* is created manually, 3) we only detect a single event transition within the *event transition graph*.

## 2 Related Work

AED was previously proposed and tackled using a sentence classification approach[1] for detecting final scores of basketball matches. Moreover, AED falls under the broader family of problems collectively known as Topic Detection and Tracking (TDT), which includes traditionally, New Event Detection (NED), Topic Tracking (TT), and Retrospective Event Detection (RED), etc. TDT defines a evaluation paradigm that addresses event-based organization of broadcast news[3], with a significant focus on NED and TT for news [4][5][6][7][3][8][9][10]. AED differs from typical TDT tasks like NED/TT/RED primarily in two ways: 1) AED is concerned only with one particular user-predefined anticipatory event; 2) AED will return a hit if and only if the user-anticipated transition has consummated for that specified event genre. For example, suppose NED or RED is set up to return alerts for mergers and acquisitions events, then any news describing a new rumor or latest developments related to acquisition could result in one or more NED/RED hits. On the other hand, AED could be configured to return a hit if and only if a particular acquisition such as some company buying Skype is formally announced.

Morever, NED was shown empirically to be a very hard problem if only simple vector space representation was used [6]. Yang et al.[8] reported a substantial performance gain by first classifying news articles into different topics, followed by applying 1NN to detect new events (NED). Kumaran et al.[10] applied text classification techniques and extracted named entities, but for detecting *all* new events of a particular category (using a model trained threshold) instead of finding the transition of a user-specified AE. Unlike AED, non of the above uses classification to detect *new* events.

Closely related to AED is RED, another NED derivative. Li et al.[11] attempts to identify events within a corpus of historical news articles with the help of time, user feedback, and content information. It assumes that the news event histogram of a particular event genre is Gaussian-distributed with each burst denoting a new event. RED cannot be used to solve the AED problem since it detects generic events, and requires multiple documents in order to form a statistically significant peak.

While similarity-based approaches had made limited inroads in TDT, others have tried incorporating domain knowledge to tackle the TDT problem[12][13]. Moreover, only a few existing work attempt to construct an *event transition graph*[14], which is the prerequisite for representing transitions between events in AED. Specific to news alerts, there has been previous work on presenting news to users in a meaningful and efficient manner[15][16].

Nallapati et al.[16] used interdependencies between news events to build an unsupervised relational structure similar to AED's *event transition graph*, but which was not used for AED. Another related work is Kleinberg's model for online change detection in data streams [17], which assumes that the points (news articles) in the stream are independently emitted by some underlying probability distribution; its goal is to detect any changes in distribution. Like NED, it requires more than one document to identify a significant change.

## 3 AED Model

Our proposed AED system first retrieves a set of *generic* acquisition news articles from Google News Alerts based on the user supplied list of domain specific keywords. The articles are then manually labelled as positive or negative with respect to a single transition, and fed into a classifier for training. To test this AED model, we manually created and labelled a separate and independent dataset comprising seven acquisition topics. For each topic, we use the trained generic AED classifer to detect the earliest news article published after the AE (in this case the announcement of an acquistion) has consummated.

### 3.1 Anticipatory Event Representation

An AED user preference is defined as a single transition in the event transition graph. In practice, topics of the same type (e.g. US Presidential elections) often involve a typical set of event transitions (e.g. nomination of party's Presidential candidates, nomination of party's Vice-Presidential candidates, election of party's Presidential team, election of Presidential team). Thus, it is reasonable to train an AED model using news about past-occurences of a similar nature.

Creating an event transition graph automatically based on arbitrary user specifications is extremely difficult. In our model, we assume that an event transition graph is already available, along with generic articles representing the "pre" and "post" states of a user preferred transition (preference). Our problem is thus reduced to applying *online AED* to a live stream of news articles with the goal of identifying the first story after a user-specified transition.

Figure 2 shows an example event transition graph describing typical states shared by most company acquisition topics. Suppose a user is interested in the event transition, $transition_{2,3}$ from $event_2$ ("In talk to acquire") to $event_3$ ("Announce acquisition"). As there are usually multiple news articles associated with the "Announce acquisition" event, AED will try to detect the first story among these. The complete AED framework is shown in Figure 3.

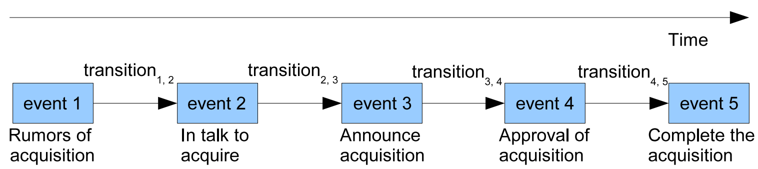


**Fig. 2.** Event transition graph for the "acquisition" topic.

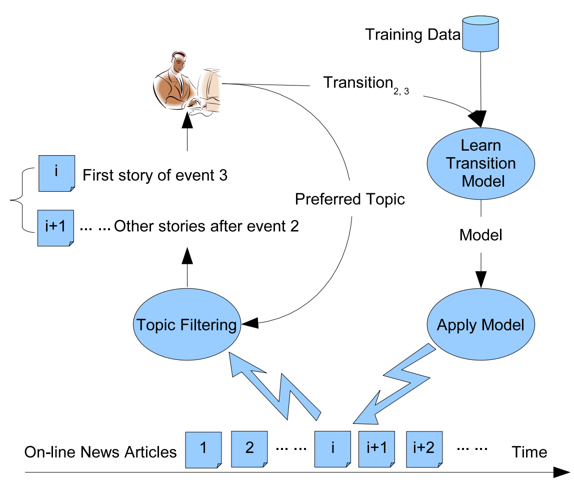


**Fig. 3.** Online AED system framework.

### 3.2 Named Entities and Text Classification in AED

**Named Entities Analysis in AED** In news stories, named entities of different types help provide essential context information. For example, company names involved in a merger and acquisition topic clearly helps to distinguish a particular topic from other topics. However, within a specified topic, named entities alone are not sufficient to determine an event transition boundary. As a matter of fact, we found experimentally that verbs and their senses actually carry more valuable information for determining a transition.

In our experiments, we use BBN's Identifinder[18] to identify 24 types of named entities, including *Animal*, *Contact info*, *Disease*, *Event*, *Facility*, *Game*, *Geo-political entities*, *Language*, *Law*, *Location*, *Nationality*, *Organization*, *Person*, *Plant*, *Product*, *Substance*, *Work of art*, *Date*, *Time*, *Cardinal*, *Money*, *Ordinal*, *Percentages*, and *Quantity*. Extracted named entities are then replaced in line by one of the 24 named entity types.

**Classification Methods** We tried three different feature representation methods and one classifier combining strategy to train the AED classifier, as follows:

*CONTENT* : Entire news content as features.
*TITLE* : Title as features.
*1SENT* : First sentence as features.
*VOTING* : Majority voting on above three classifier outputs.

The TITLE and 1SENT representations were inspired by the observation that human experts can usually decide if a news is a hit simply based on its first sentence and/or title. Moreover, the TITLE and 1SENT representation of a news article may not always carry useful features, and the AED decision will have to fall back to the CONTENT representation. For example, the first sentence "*Signature Control Systems is off to a busy start in early 2006*" does not contain features really relevant to the "acquisition" event transition. VOTING was thus used as a simple and effective way to improve the overall accuracy.

### 3.3 Evaluation Methodology

**Evaluating the AED Event Transition Classifier** We adopt the standard information retrieval measures, precision, recall, and f1-score to evaluate the performance of the various AED classifiers.

**Evaluation of Anticipatory Event Transition Detection** Suppose we are given a set of $N$ news articles $X = \{x_1, ..., x_N\}$ about a topic, and an event transition graph $E = \{e_1, ..., e_n\}$ comprising $n$ events. Each news $x_i$ is assigned a publication date/time represented by $t(x_i)$ and an event type in $E$ represented by $e(x_i)$, the latter of which is also known as the true event of $x_i$.

We assume that all news articles in $X$ are sorted in time ascending order, i.e. $t(x_i) \leq t(x_j)\ \forall i < j$, and all events in $E$ are sorted in time ascending order, i.e. $t(e_i) \leq t(e_j)\ \forall i < j$.

By applying our trained AED classifier on a news article $x_i$, we obtained its assigned event denoted by $s'(x_i)$. Given a $transition_{k-1,k}$ (i.e. user preference), the objective of AED is therefore to find the news article $x_m$ that satisfies:

$$x_m = \arg\min \{t(x_i) \mid \forall x_i \ where \ \ s'(x_i) = e_k\}$$

To make the time comparison easier between the detected first story $x_m$ and the event $e_k$, we also define the *true time* of $e_k$, $t(e_k)$, as follows:

$$t(e_k) = \min \{t(x_i) \mid \forall x_i \textit{ where } e(x_i) = e_k\}$$

Once the first story $x_m$ of the anticipatory event $e_k$ is determined by the AED classifier, all subsequent news articles, $x_j$, $j = (m+1), \ldots, N$ will be assigned to event(s) $e_k$ post $transition_{k-1,k}$. Occasionally, the first story identified by AED may be prematured, delayed, or undefined (never found). Accordingly, we define four evaluation criteria as follows:

*Accurate Alarm* : $t(x_m) = t(e_k)$. First story of $e_k$ found successfully.
*Delayed Alarm* : $t(x_m) > t(e_k)$. First story found was too late.
*False Alarm* : $t(x_m) < t(e_k)$. First story found was prematured.
*Miss* : $t(x_m) = undefined$. No $x_i$ in $X$ has $s'(x_i) = e_k$. AED fails to even identify the event!

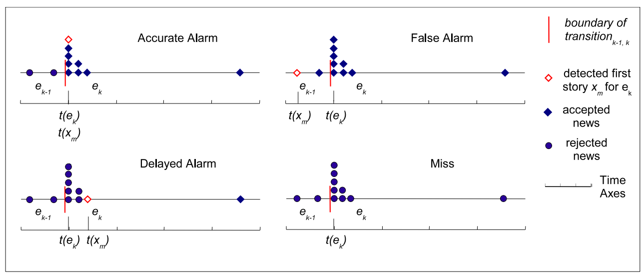


**Fig. 4.** An evaluation example for *Transition Detection* in AED.

Figure 4 graphically depicts each of the four evaluation criteria for AED *Transition Detection*. In Section 5 we will use the same type of graph to illustrate and analyze our experimental results. Specifically, we simply tally the total number of *false alarms*, *delayed alarms*, *accurate alarms*, and *misses* to evaluate the AED performance on a given set of events. For news alerts, an *accurate alarm* is the most desirable, followed by a *delayed alarm*. Otherwise, a *miss* is generally preferred over a *false alarm*.

## 4 Testbed

Two datasets were created specifically for evaluating the AED problem. For quality assurance purposes, each document in the two datasets was scrutinized and annotated by at least two people.

### 4.1 *Google Acquisition* Dataset

In order to learn an anticipatory event transition such as $transition_{2,3}$ in Figure 2, we manually created the generic *Google Acquisition* dataset. This dataset contains 346 as-it-happens news articles returned by Google News Alerts using the keywords “announce acquisition”, which corresponds to *event*3 in Figure 2, during the two-month period from Dec 19, 2005 to Feb 19, 2006.

Each article in *Google Acquisition* is manually labelled as one of two possible events, i.e. “pre” or “post” $transition_{2,3}$. Unfortunately, some articles can appear ambiguous even to a human expert. One general rule-of-thumb is to label the document based on overall context. For example, if the primary theme of an article revolves around the announcement/agreement/completion of an acquisition, we label it as post-$transition_{2,3}$; otherwise, it is labelled as pre-$transition_{2,3}$. We note that the latter case could also include irrelevant documents completely unrelated to acquisition.

To ensure consistency in labelling, a set of guidelines and rules was established, based on which 178 documents were labelled as positive and 168 as negative, which means that Google News Alerts returned 168 (48.6%) outright false alarms for the subscribed keywords “announce acquisition”. This is a typical result from a simplistic keyword-based news alert system.

### 4.2 *Acquisition7* dataset

We created another dataset, *acquisition7*, which covers seven recent acquisition topics as the test data for our proposed online AED solution. Each acquisition news topic in *acquisition7* is comprised of 20 news articles returned by Google News, approximately half of each (10) were reported before and after $transition_{2,3}$. The major difference between this dataset and the *Google Acquisition* dataset is that each document is not generic but instead tied to a specific acquisition. Further, there are no irrelevant documents in this dataset; a document occurs either before or after $transition_{2,3}$ for a specific acquisition.

The 7 acquisition news topics are listed in Table 1, where $t(e_3)$ refers to the true occurrence date for $event_3$ in Figure 2. The annotation of *Acquisition7*

**Table 1.** Make up of the *Acquisition7* dataset.

| *Acquisition Topics* | $t(e_3)$ |
|---|---|
| Adobe acquires Macromedia | Apr 18, 2005 |
| CNPC acquires PetroKazakhstan | Oct 26, 2005 |
| eBay acquires Skype | Sep 12, 2005 |
| Lenovo acquires IBM PC Division | Dec 08, 2004 |
| Oracle acquires PeopleSoft | Dec 13, 2004 |
| Oracle acquires Siebel | Sep 12, 2005 |
| SBC acquires AT&T | Jan 31, 2005 |

follows the same criteria as defined in Section 4.1.

## 5 Simulation Results

### 5.1 Experiment Setup

Lucene 1.4.3 was used to tokenize the news text content with stop word removal to create the corresponding document-word vector. In order to preserve time-sensitive past/present/future tenses of verbs, no stemming was done other than the removal of a few articles.

We used a normalized (unit length) binary document vector representation because we observed that co-occurrences of terms are far more important than the raw term frequency and inverse document frequency for AED. The normalized binary document-word vectors are than fed into SVM-light [19] for training and classification. SVM cost factors[20] were used to offset the slight imbalance in numbers between the positive and negative documents.

### 5.2 AED via Cosine Similarity

As a baseline, we evaluated AED performance using simple cosine similarity on the *Acquisition7* dataset. Standard TFIDF document representation was used with the following variations [10]: all terms (All Terms), all terms without named entities (No NE), and name entities only (NE only).

Each incoming news article is compared to *all* existing news articles (assumed to be negative or pre-transition) from all 7 topics. If the cosine similarity between this news and its nearest neighbor falls below a threshold, the incoming news is considered to have consummated the transition; otherwise it is classified as a negative news.

The similarity approach generated largely false alarms and misses, except for one accurate alarm, using various values of similarity threshold. Figure 5 shows the ratio of misses to false alarms for all 3 vector representations versus a gradually increasing similarity threshold. From Figure 5, we observe that starting from a low similarity threshold, the system was initially very strict (incoming news must be significantly different, i.e. has low cosine similarity compared to all existing news), resulting in high percentage of misses. As the similarity threshold is gradually increased, the system was able to detect some news, but almost all prematuredly as false-alarms, except for one accurate alarm detected by the "no NE" representation. Our results clearly show that similarity based approaches, which does not use a conceptual model, are too simplistic to detect a transition leading to a user-desired AE.

### 5.3 AED via Event-conditioned Novelty Detection

Event-conditioned 1NN novelty detection [8], essentially a topic-constrained cosine similarity approach, first classifies a news into a known topic before applying cosine similarity comparison between it and its nearest neighbor. To model this approach, we applied cosine similarity AED to all news within the same topic

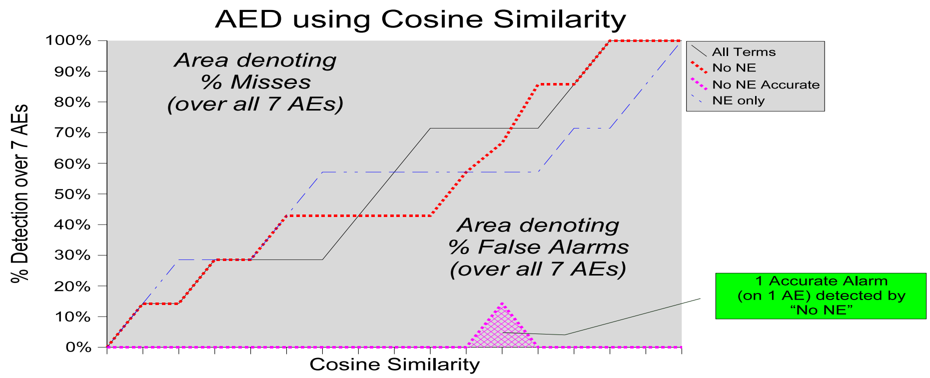


**Fig. 5.** AED using cosine similarity on Acquisition7 Dataset.

to obtain the results listed in Table 2. Clearly, event-conditioned novelty detection failed the AED task miserably as it generated all but one false alarms. This shows that even with topical constraints, similarity approaches still cannot perform AED reliably.

**Table 2.** AED results on *Acquisition7* using event-conditioned novelty detection.

| *Alarms:* | *Accurate* | *Delayed* | *False* | *Miss* |
|---|---|---|---|---|
| Adobe acquire Macromedia | | | √ | |
| CNPC acquires PetroKazakhstan | | | √ | |
| eBay acquires Skype | √ | | | |
| Lenovo acquires IBM PC Divison | | | √ | |
| Oracle acquires PeopleSoft | | | √ | |
| Oracle acquires Siebel | | | √ | |
| SBC acquires AT&T | | | √ | |

### 5.4 Validating the *Google Acquisition* Dataset

In order to validate the generic $transition_{2,3}$ trained model, we conducted two-fold cross-validated experiments using the four text classification approaches of Section 3.2 on the *Google Acquisition* dataset. The dataset is first split along the timeline into two equal parts: 1) news articles dating from Dec 19, 2005 to Jan 19, 2006, and 2) news articles dating from Jan 20, 2006 to Feb 19, 2006. One part was used for training with the other part used for testing and vice-versa.

From the test results summarized in Figure 6 and Table 3, we see that the VOTING strategy is the overall best performer with the least number of false

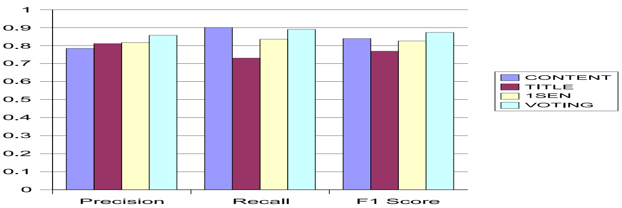


**Fig. 6.** Average test results of the four text classifiers on *Google Acquisition*.

alarms, while the CONTENT method gives a slightly higher recall at the expense of almost twice as many false alarms. The main problem with the CONTENT method is that it is easily affected by a few transition-alluding sentences in negative documents, such as "*Additionally, Magazine Acquisition announced that Morgan Stanley Real Estate and Onex Real Estate will be partnering with Sawyer Realty Holdings LLC ("Sawyer") in the TCT acquisition*", which understandably appears positive to a classifier. This is because the mere occurrence of the words "acquisition" and "announced" is sufficient to trigger the trained model, which uses a binary bag-of-words representation. The VOTING strategy thus combines the best results from CONTENT, TITLE, and 1SEN methods.

**Table 3.** Average test results on *Google Acquisition*. Best results are shown in bold.

| *Average* | *CONTENT* | *TITLE* | *1SEN* | *VOTING* |
|---|---|---|---|---|
| False Alarms | 22.5 | 15.5 | 17 | **13.5** |
| Misses | **9** | 24.5 | 15 | 10 |
| Precision | 0.7847 | 0.8110 | 0.8172 | **0.8571** |
| Recall | **0.9011** | 0.7308 | 0.8352 | 0.8901 |
| F1 | 0.8389 | 0.7688 | 0.8261 | **0.8733** |

Apart from deciding the best classification strategy, one other significance of this experiment is that it increased the precision of Google's returned news alerts from 51.4% to 85.7%, a more than 33% improvement! All in all, the high precision and recall figures confirmed that the *Google Acquisition* dataset is indeed suitable for modelling $transition_{2,3}$ for subsequent AED evaluations in Section 5.5.

### 5.5 AED via Classification

In this section, we test the generic AED classifier trained by *Google Acquisition* on the *Acquisition7* dataset. Figure 7 shows the true $transition_{2,3}$ boundaries for each of the 7 *Acquisition7* topics distributed along a timeline. Three AED outcomes are shown in Figures 8-10. Note that once the "first" story of $e_3$ has been identified by AED, all subsequent news articles are labelled post-$transition_{2,3}$.

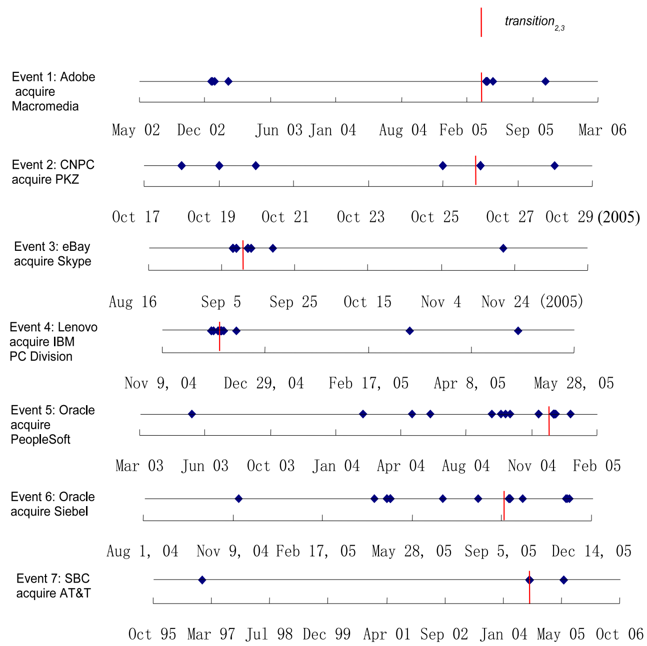


**Fig. 7.** $transition_{2,3}$ boundaries in *Acquisition7*.

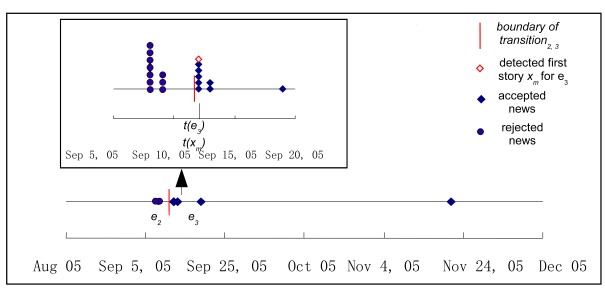


**Fig. 8.** Online AED of "eBay acquires Skype" found an accurate alarm, $t(x_m) = t(e_3)$.

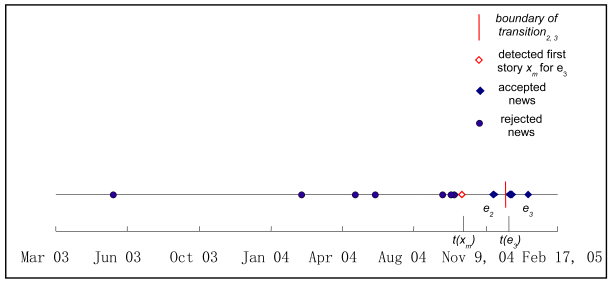


**Fig. 9.** Online AED of "Oracle acquires PeopleSoft" found a false alarm, $t(x_m) < t(e_3)$.

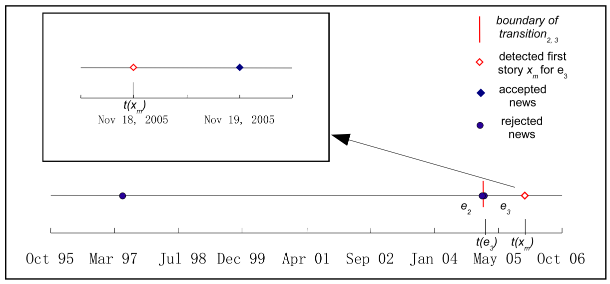


**Fig. 10.** Online AED of "SBC acquires AT&T" found a delayed alarm, $t(x_m) > t(e_3)$.

Table 4 gives a summary of the overall performances, which shows that AED based on the VOTING method generated 4 accurate alarms, 1 delayed alarm, 2 false alarms, and 0 misses. This means that the model trained by *Google Acquisition* was able to cover the main characteristics of all 7 acquisition topics.

Moreover, comparing this result with that of the CONTENT method as shown in Table 5, we found that the AED evaluation for Acquisition7 dataset is inconsistent with the two-fold cross-validation results for *Google Acquisition* dataset with respect to false alarms. In *Google Acquisition*, the VOTING method reduces false alarms, with the CONTENT method yielding the highest recall. The situation is completely reversed for *Acquisition7*. Based on the analysis in Section 5.4, we are inclined to trust the evaluation results for *Google Acquisition* better because the inconsistencies could simply be caused by the relatively small size of the *Acquisition7* dataset.

Nevertheless, in spite of the above inconsistencies, AED results achieved by both the VOTING and CONTENT methods were leaps and bounds ahead of the cosine similarity results (Figures 5) and event-conditioned novelty detection results (Table 2). This is actually a very encouraging outcome for a preliminary investigation into AED, and thus provides strong support and credibility to our AED model and solution.

**Table 4.** AED results on *Acquisition7* using the VOTING method.

| *Alarms:* | *Accurate* | *Delayed* | *False* | *Miss* |
|---|---|---|---|---|
| Adobe acquires Macromedia | $\surd$ | | | |
| CNPC acquires PetroKazakhstan | $\surd$ | | | |
| eBay acquires Skype | $\surd$ | | | |
| Lenova acquires IBM PC Division | | | $\surd$ | |
| Oracle acquires PeopleSoft | | | $\surd$ | |
| Oracle acquires Siebel | $\surd$ | | | |
| SBC acquires AT&T | | $\surd$ | | |

**Table 5.** AED results on *Acquisition7* using the CONTENT method.

| *Alarms:* | *Accurate* | *Delayed* | *False* | *Miss* |
|---|---|---|---|---|
| Adobe acquires Macromedia | $\surd$ | | | |
| CNPC acquires PetroKazakhstan | $\surd$ | | | |
| eBay acquires Skype | $\surd$ | | | |
| Lenovo acquires IBM PC Division | | | | $\surd$ |
| Oracle acquires PeopleSoft | | $\surd$ | | |
| Oracle acquires Siebel | $\surd$ | | | |
| SBC acquires AT&T | | $\surd$ | | |

## 6 Conclusion

We have made five main contributions in this paper: 1) we formally defined and formulated a conceptual model for the AED problem and identified its associated research issues, 2) proposed a new way of applying named entities for AED, 3) proposed a principled way to assemble generic training data for learning one AE transition, using the user's AE preferences, 4) verified the feasibility of AED in practice for one restricted domain, 5) compared our method with two classical cosine similarity methods. The encouraging results in this paper showed AED to be applicable in practice, thus paving the way for future work.

We have made a number of simplifying assumptions in this study: 1) we assumed that an *event transition graph* matching the user's query is available, based on some domain knowledge, 2) we only detect a single transition, among many other possible transitions in the graph, and we claim the transition corresponds to the user specified list of keywords (i.e. user preferences), 3) we constrained our testbed to a particular genre of AE, that of mergers and acquisitions. One possible future AED research focus is simply the relaxation of these assumptions, which will involve significant challenges.

Naturally, the holy grail of AED is to detect any number of AE transitions of arbitrary genres. This is akin to having a live assistant constantly scanning newsfeed monitoring a set of AEs. Surely, current state-of-the-art technologies will not be able to attain this in the foreseeable future. However, we could still improve AED by incorporating additional information such as frequency/time of documents/words and user feedback. We hope to eventually come up with an effective and practical AED system, perhaps initially for some restricted domain, and which overcomes limitations of existing systems like Google News Alerts.

## References


1. He, Q., Chang, K., Lim, E.P.: Anticipatory event detection via sentence classification. In: IEEE SMC Conf. (2006) to appear
2. Google: (Google news alerts, http://www.google.com/alerts.)
3. Allan, J.: Topic Detection and Tracking. Event-based Information Organization. Kluwer Academic Publishers (2002)
4. Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., Caputo., D.: Topic-based novelty detection: Final report. In: DARPA Broadcast News Transcription and Understanding Workshop. (1999)
5. Jin, H., Schwartz, R., Sista, S., Walls, F.: Topic tracking for radio, tv broadcast, and newswire. In: DARPA Broadcast News Workshop. (1999) 199–204
6. Allan, J., Lavrenko, V., Jin, H.: First story detection in tdt is hard. In: 9th ACM CIKM Conf. (2000) 374–381
7. Stokes, N., Carthy, J.: Combining semantic and syntactic document classifiers to improve first story detection. In: 24th ACM SIGIR Conf. (2001) 424–425
8. Yang, Y., Zhang, J., Carbonell, J., Jin, C.: Topic-conditioned novelty detection. In: 8th ACM SIGKDD Conf. (2002) 688–693
9. Brants, T., Chen, F., Farahat, A.: A system for new event detection. In: 26th ACM SIGIR Conf. (2003) 330–337
10. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: 27th ACM SIGIR Conf. (2004) 297–304
11. Li, Z., Wang, B., Li, M., Ma, W.Y.: A probabilistic model for retrospective news event detection. In: 28th ACM SIGIR Conf. (2005) 106 – 113
12. Fukumoto, F., Suzuki, Y.: Event tracking based on domain dependency. In: 23rd ACM SIGIR Conf. (2000) 57–64
13. Maguitman, A., Leake, D., Reichherzer, T., Menczer, F.: Dynamic extraction topic descriptors and discriminators: towards automatic context-based topic search. In: 13th ACM CIKM Conf. (2004) 463 – 472
14. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: 11th ACM SIGKDD Conf. (2005) 198–207
15. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: 26th ACM SIGIR Conf. (2003) 314–321
16. Nallapati, R., Feng, A., Peng, F., Allan, J.: Event threading within news topics. In: 13th ACM CIKM Conf. (2004) 446–453
17. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: 30th VLDB Conf. (2004)
18. Bikel, D.M., Schwartz, R., Weischedel, R.M.: An algorithm that learns what's in a name. In: Machine Learning. (1999) 34(1–3):211C231
19. Joachims, T.: (Svm-light, http://svmlight.joachims.org/)
20. Morik, K., Brockhausen, P., Joachims, T.: Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In: 16th ICML Conf. (1999) 268–277