Available online at www.sciencedirect.com

ScienceDirect

Journal of Informetrics 1 (2007) 131–144

Journal of INFORMETRICS
An International Journal

www.elsevier.com/locate/joi

Generating overview timelines for major events in an RSS corpus

Rudy Prabowo[a,*], M. Thelwall[a], Mikhail Alexandrov[b]

[a] *School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, WV11SB Wolverhampton, UK*
[b] *Autonomous University of Barcelona, Barcelona, Spain*

Received 31 August 2006; received in revised form 19 October 2006; accepted 23 October 2006

Abstract

Really simple syndication (RSS) is becoming a ubiquitous technology for notifying users of new content in frequently updated web sites, such as blogs and news portals. This paper describes a feature-based, local clustering approach for generating overview timelines for major events, such as the tsunami tragedy, from a general-purpose corpus of RSS feeds. In order to identify significant events, we automatically (1) selected a set of significant terms for each day; (2) built a set of (term–co-term) pairs and (3) clustered the pairs in an attempt to group contextually related terms. The clusters were assessed by 10 people, finding that the average percentage apparently representing significant events was 68.6%. Using these clusters, we generated overview timelines for three major events: the tsunami tragedy, the US election and bird flu. The results indicate that our approach is effective in identifying predominantly genuine events, but can only produce partial timelines.

Keywords: Feature selection; Clustering; Overview timeline

1. Introduction

The task of identifying significant events from real time news feed data is a standard one in data mining and event detection and tracking (Allan, Papka, & Lavrenko, 1998b; Yang, Pierce, & Carbonell, 1998). The Internet now hosts a range of readily accessible information formats that are new candidates for event detection, and these may come to replace or supplement traditional types, or may give rise to new event detection applications. Really simple syndication (RSS) is one such technology and has already become a widely used standard: it allows blogs and news sources to post-timely information to subscribers, for example, hourly or daily summaries of the most recent updates. RSS feeds have great potential to be used for public-opinion gathering (Glance, Hurst, & Tomokiyo, 2004; Gruhl, Guha, Liben-Nowell, & Tomkins, 2004), mainly because of the large numbers of blog authors maintaining sites with RSS feeds, although bloggers are not typical citizens (Adar, Zhang, Adamic, & Lukose, 2004; Lin & Halavais, 2004) and have a wide variety of motives (Herring, Scheidt, Bonus, & Wright, 2004). In addition, the concise RSS formats allow relatively low-bandwidth data gathering, even for a large number of different sources. When a major (or world) event, such as the Asian tsunami (26/12/2004), occurs, RSS feeds could therefore be used to generate an overview timeline of the event.

Our contributions are to develop an automatic method to achieve the following using RSS data.

* Corresponding author. Tel.: +44 1902 518584; fax: +44 1902 321478.
E-mail addresses: Rudy.Prabowo@wlv.ac.uk (R. Prabowo), Mike.Thelwall@wlv.ac.uk (M. Thelwall), dyner1950@mail.ru (M. Alexandrov).

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.10.002

(1) Find daily sets of significant terms (either nouns or noun phrases) which maybe associated with important events, i.e. the most discussed happenings (Section 3).
(2) Use the significant terms to build a set of (term–co-term) pairs and cluster the pairs. The clusters are our candidates for the day's significant events (Section 4).
(3) Generate overview timelines for major events by sorting the clusters by date (Section 5).

In this paper, we are primarily interested in the precision of the clusters in (2). More specifically, we assess the extent to which human judges agree that the automatically generated clusters genuinely describe a single event. A human-based evaluation was important to discover whether the results could be understood by potential end users, i.e. human interpreters.

To illustrate the timeline generation, three major events, 'tsunami tragedy', 'US election' and 'bird flu spreading', were selected as our case studies. Tables 5–7 show the generated timelines for each major event. Each timeline refers to one particular major event along with many related, subsequent events.

2. Related work

This section reviews existing work in the area of (1) term selection, (2) topic and event detection and tracking (TDT) and (3) timeline generation.

2.1. Term selection

Given a set of terms (e.g. words, word stems, nouns and noun phrases) in a document collection, selecting the most significant terms is the first step. This stage is common to a range of specific tasks, including information retrieval (IR), automatic text classification and time series analysis. The selected terms represent document features in the form of a term vector: a list of the most significant terms from the document and their frequencies in the document.

Either *tf·idf* (Salton & McGill, 1986) or *lnu* weighting (Singhal, Buckley, & Mitra, 1996) can be applied to assign each term a value which estimates its significance. These formulae take into account both local and global term frequency. In IR, the assigned value is then used as a starting point to: (1) compute the similarity between the documents available in the corpus and a user query and (2) rank search results in order of relevance to a user query. In an ideal scenario, each term should be assigned a degree of significance, such that an IR system can achieve a high precision level at 100% recall (Baeza-Yates & Ribeiro-Netto, 1999; Belew, 2000). It is not suitable for our event detection task, however, because important events may be identified through single highly significant terms (Swan & Allan, 2000).

The three formulae that have previously been used to identify significant events in blog or RSS corpora are variants of *tf* and *tf·idf* (Gruhl et al., 2004; Glance et al., 2004; Thelwall, Prabowo, & Fairclough, 2006). The formulae do not, however, depend on the full document space, but on a fixed time period as a time window of observations, and are used to quantify the 'burstiness' of a term within the fixed, short time period, i.e. the degree of importance of terms within the time period. The result changes if another time window is used, for example, 1 week earlier or later. While this feature is useful to keep track of the burstiness of terms for different time windows, it is less suitable for the initial identification of significant events. For this, the degree of significance of a term over a long period of time is required, e.g. 1 year.

The commonly used formulae for identifying significant terms in the area of automatic text classification are: χ^2, Mutual Information (MI) and Information Gain (I) (Sebastiani, 2002). Swan and Allan (2000) use a χ^2-based method to determine the degree of significance of terms on given dates. Nevertheless, it is not yet clear whether this is the best method for all types of data, and in the context of RSS, it is also worth harnessing the power of the Information Gain method (Prabowo & Thelwall, 2006).

2.2. TDT

In the context of the TDT task, an event is defined to be something that happens at a specific time and place, whereas a topic is defined more widely as a seminal event or activity, along with all directly related events and activities (Allan, Carbonell, Doddington, Yamron, Yang, 1998a). The term 'story'; is often used to describe the natural unit of text in which the information arrives, such as a single newswire report. The topic detection and tracking tasks focus on the identification of topics across stories. The event detection task is to cluster together stories that refer to the same event. The event-tracking task is concerned with assigning each incoming new story to the most appropriate event that it discusses, if any. Otherwise, the story is assumed to detect a new event (Allan et al., 1998b). In contrast, our task focuses on generating an overview timeline with respect to one particular major event and its related, subsequent events.

To detect and track events, the following method is used. Given a set of stories, assign each term found within each story a weight. The weighted terms of a story are stored as a vector, and regarded as the representation of the story. In the case of event detection, given a set of terms, stories that discuss the same event are clustered together. In the case of event tracking (1) the similarity between the term vector of a new story and all the term vectors of the existing stories is computed and (2) the most appropriate existing event is assigned to the new story, if any. Otherwise, a new event is recorded.

2.3. Timeline generation

As timestamped data, such as blog data and news articles, have become available, the timeline generation task has attracted a number of researchers. Glance et al. (2004) used blog data to carry out topic mining, detect key persons and produce a timeline for a topic or a key person. The method used is as follows: (1) select a set of significant phrases; (2) cluster all the selected phrases. Two phrases are clustered together, if the cosine similarity of their occurrence is greater than a threshold. Each cluster represents a topic; (3) generate timelines for the topics.

Swan and Allan (2000) focused on assigning each significant term a time period by using χ^2-tests. Swan and Allan (2000) ranked the significant terms according to their χ^2 values in a descending order, i.e. the term with the largest χ^2 value was at the top of the list. Then, they compared the time period of a term with all the lower ranked features. If their time periods overlapped, they carried out χ^2-tests to determine whether the two features were dependent, If yes, they marked the features as potential members of a cluster. Finally, they carried out a hierarchical agglomerative clustering on the marked features. The clusters were evaluated against a set of predefined topics (discussed in detail in Section 4.5).

Smith (2002) extracted a number of place names from historical documents and assigned each place name a date. Then, Smith (2002) ranked the collocations of place name and date pairs according to their log-likelihood values. The significant collocations can be displayed in timelines.

In contrast, we do not focus on a topic, but on the generation of a timeline for the wider concept of a major event. The generated timeline contains not only the major event, but also its related events. Our work is more related to Smith (2002), than to Glance et al. (2004) and Swan and Allan (2000). Our approach, however, is more general. We did not only exploit place names, but also other phrases as a way to build a set of (term–co-term) pairs and to cluster the pairs. The TDT corpus was built for single topic/event detection exercise, but was not annotated for the purpose of evaluating overview timelines of an event and its related events. We therefore could not use the TDT corpus for our experiment. Instead, we use RSS data, which is the type of data that we believe can be usefully exploited for timeline generation.

3. Significant term selection

The selection of significant terms was conducted in three stages.

(1) RSS items were collected and the text found within the items was processed;
(2) χ^2 and Information Gain (I) values were computed;
(3) A set of significant terms was selected.

3.1. Pre-processing texts

The following procedure was used to pre-process the texts found within a set of RSS items.

(1) Mozhdeh (Thelwall et al., 2006) was used for collecting data. The system monitored 19,587 RSS feeds hourly (daily for infrequently updated feeds). Each feed returns a set of items, with each item containing a separate set of information. The system stored each new item found.

Table 1
A 2 × 2 contingency table

	term_i	$\overline{\text{term}_i}$
date_j	a	b
$\overline{\text{date}_j}$	c	d

(2) For each item, the title, description and publication date were automatically extracted and stored in a plain text file. Each publication date was converted into its associated GMT time by using a date converter. For each file, a part of speech (POS) tagger (Brill, 1992) and noun phrase (NP) chunker (Ramshaw & Marcus, 1995) were used to tag and chunk the item texts, extracting nouns and noun phrases, including proper nouns. The tagger achieved 95% precision (Brill, 1992) and the chunker 93% (Ramshaw & Marcus, 1995). The tagger and chunker, which run on an intel P4 3.2 GHz, can process about 650K items per day. They could therefore operate in real time.

(3) Three inverted files were built.
   (a) TermItem: Used to determine to which item each term belongs;
   (b) ItemRSSFeed: Used to determine to which RSS feed each item be longs;
   (c) ItemDate: Used to determine when each item was posted in GMT time (i.e., publication date).

3.2. Computing χ^2 and Information Gain (I)

Given a term, term_i and a publication date, date_j, the 2 × 2 contingency table used for calculating a χ^2 value is constructed as follows (Table 1).

The χ^2 value of term_i with regard to date_j was calculated as follows:

$$\chi^2 = \sum_k \frac{(\mathrm{O}_k - \mathrm{E}_k)^2}{\mathrm{E}_k} \tag{1}$$

- $k = \{a, b, c, d\}$;
- O_k is the observed frequencies in a 2 × 2 contingency table with respect to term_i and date_j;
- E_k is the expected frequencies of O_k with respect to term_i and date_j.

A Yates continuity correction was applied to each χ^2 calculation, as the degree of freedom is 1. Large χ^2 values suggest that term_i and date_j are dependent upon each other.

Let D be a binary date variable, $\{d, \bar{d}\}$ representing the presence or absence of a date and T be a binary term variable, $\{t, \bar{t}\}$ representing the presence or absence of a term. Information Gain, $I(D;T)$ was computed based on $H(D)$, an entropy value for D and $H(D|T)$, the conditional entropy of D given T and is defined as follows:

$$\begin{aligned} I(D;T) &= H(D) - H(D|T) \\ I(D;T) &= \left\{ -\sum_{j=d,\bar{d}} p(j) \cdot \log_2 p(j) \right\} - \left\{ p(t) \cdot \sum_{j=d,\bar{d}} H(j|t) + p(\bar{t}) \cdot \sum_{j=d,\bar{d}} H(j|\bar{t}) \right\} \\ I(D;T) &= -\left\{ p(d) \cdot \log_2 p(d) + p(\bar{d}) \cdot \log_2 p(\bar{d}) \right\} - \left\{ p(t) \cdot \sum_{j=d,\bar{d}} H(j|t) + p(\bar{t}) \cdot \sum_{j=d,\bar{d}} H(j|\bar{t}) \right\} \end{aligned} \tag{2}$$

The conditional entropy of D given T was computed as follows:

$$\sum_{j=d,\bar{d}} H(j|t) = -\left\{ p(d|t) \cdot \log_2 p(d|t) + p(\bar{d}|t) \cdot \log_2 p(\bar{d}|t) \right\} \tag{3}$$

$$\sum_{j=d,\bar{d}} H(j|\bar{t}) = -\left\{ p(d|\bar{t}) \cdot \log_2 p(d|\bar{t}) + p(\bar{d}|\bar{t}) \cdot \log_2 p(\bar{d}|\bar{t}) \right\} \tag{4}$$

In information theory (Shannon & Weaver, 1963), $I(D; T)$ is used to measure the average reduction in uncertainty about D that results from learning the value of T (MacKay, 2003). In our context, we (1) apply this notion of average reduction in uncertainty to determine the degree of closeness between D and T by learning the presence and absence of T in each RSS item, with regard to the $H(D)$ of one particular publication date and (2) use the value of $H(D|T)$ to quantify the degree of uncertainty that a term T is significant on a date D. The smaller a $H(D|T)$ value is, the higher the degree of certainty that T is a good indicator for D. For $H(D|T) = 0$, the degree of uncertainty in learning the value of D is 0, which means that T is highly indicative for a date D.

The time complexity of χ^2, as well as Information Gain (I) is $\mathcal{O}(\mathcal{T} \cdot \mathcal{D})$, where $\mathcal{T}$is the number of terms found within a set of RSS items and $\mathcal{D}$ is the number of publication dates used.

3.3. Selecting significant terms

From the 19,857 monitored RSS feeds, 880,536 different items were extracted between 01/01/2004 and 28/02/2005. A total of 1,736,715 unique terms were extracted from a total of 413 publication dates (some items were clearly old when collected). The 127,859 were single word terms and 1,608,856 were multi word terms. From these data, 2,912,581 term–date pairs were generated. Each term–date pair was assigned χ^2 and I values as described above. Our previous evaluation results (not using human evaluators) suggest that χ^2 would be the best of the three methods (Prabowo & Thelwall, 2006). Nevertheless, it is far from perfect as extremely high values can still occasionally be assigned to relatively insignificant terms. The results also showed that χ^2 and I had a strong degree of agreement when judging the term significance and I can aggressively remove some insignificant terms. In attempt to extract a high proportion of genuinely significant terms, only those which were judged to be significant by both χ^2 and I were selected, a total of 684,431.

4. Term clustering

This section discusses the way we automatically clustered together related terms and manually evaluated the clusters.

4.1. Clustering procedure

The ‘features’ of a document, i.e. its most significant terms, ideally should represent the essence of the document. Given a collection of documents, document clustering is the operation of grouping together similar (or related) documents, with respect to their features (Baeza-Yates & Ribeiro-Netto, 1999).

In our case, the clustering of items can be computationally intractable because of the size of the matrix: 880,536 items·684,431 features. Even with daily clustering, there are still thousands of items to process per day. To overcome the computational issue, a matrix of associations between significant terms only can be used as an alternative. This approach, however, does not take the context into account. For these reasons, we exploited term co-occurrence to provide context to a set of significant terms. The clustering procedure is described below.

For each daily set of significant terms, we extracted the terms which co-occurred with each significant term in the same sentence and generated (term, co-term) pairs. The co-terms do not necessarily need to be significant terms. Thus, a significant term was the trigger to initiate a context, which was represented by (term, co-term) pairs. To ensure that the clustering algorithm would not modify the context, unless other significant terms were linked to the existing co-terms, we used each significant term as an attribute and treated the co-terms as objects, i.e. we clustered the co-terms, rather than the significant terms. No clustering algorithm would have been necessary, if only one event occurred on a single day. This, however, was typically not the case.

Given a set of (term, co-term) pairs from a single day, only the pairs that occurred at least twice were selected (pair count threshold = 2) to remove infrequent pairs. Next, a covariance matrix, in form of a higher triangular (HT) form, was created. The cosine measure was used for measuring the closeness between two objects (co-terms). Finally, the MajorClust algorithm (Stein & Niggemann, 1999) was used to cluster the pairs.

Let G be a graph; $C = \{C_1, C_2, \ldots, C_n\}$ be the decomposition of G; $|C_i|$ be the number of nodes in C_i; λ be the minimum number of edges that must be removed to make G an unconnected graph. To get the best decomposition of

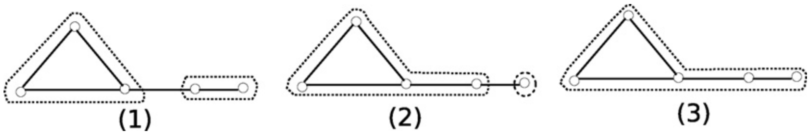

Fig. 1. An example to illustrate $\Lambda(\mathcal{C})$.

the graph, G, we need to compute and maximise the value of the weighted partial connectivity, $\Lambda(\mathcal{C})$:

$$\Lambda(\mathcal{C}) = \sum_{i=1}^{n} |C_i| \cdot \lambda_i \tag{5}$$

Let us use Fig. 1 as an example to illustrate $\Lambda(\mathcal{C})$.

The graph depicted above can be decomposed into three different possibilities. For (1), $\Lambda = 3{\cdot}2 + 2{\cdot}1 = 8$. For (2), $\Lambda = 4{\cdot}1 + 1{\cdot}0 = 4$. For (3), $\Lambda = 5{\cdot}1 = 5$. (1) yields the largest Λ value. Therefore, (1) is regarded as the best structural division of subgraphs. A $\Lambda(\mathcal{C})$-based clustering algorithm avoids long chaining effects, as illustrated in (3). The algorithm can be adjusted to handle a weighted graph.

In our experiment, we used the MajorClust algorithm, which follows the notion of $\Lambda(\mathcal{C})$ optimisation, to cluster objects with respect to their attributes. At the initial step, the algorithm assigns each object to its own cluster. In the next iterations, each object is assigned to a cluster to which the closeness of the object is stronger than to other clusters. If more than one cluster exists, one of them is randomly selected (Stein & Niggemann, 1999; Alexandrov, Gelbukh1, & Rosso, 2005). This iterative optimisation strategy leads to the local optimum of the best possible division of objects. The algorithm does not need to know the number of clusters (k) in advance and can efficiently process a large graph, as it only looks for a local optimum and does not carry out a hierarchical clustering. These characteristics are suitable for our task for the following reasons. A collection of terms must be efficiently processed. A hierarchical, agglomerative clustering may offer a global optimum, but would be computationally expensive and we would also need to estimate the level where the dendogram is to be cut to get the desired clusters. A localised variant of the k-means algorithm may run faster than MajorClust, but the MajorClust runs reasonably fast, does not suffer from chaining effect and suits our data particularly well due to its graph-based approach. The time complexity of the algorithm is $\mathcal{O}(E{\cdot}C_{\max})$, where E is the number of edges representing the connections between clusters and objects and $C_{\max}$ is the largest cluster in the graph.

The outputs were a set of clusters, each of which contained at least one co-term. We also mapped each cluster into its associated significant terms. By doing this, we can keep the context in which a co-term was used. For example, cluster 7 (Table 6), contains the co-term, ‘Japan’. The two significant terms ‘bird flu fear’ and ‘bird flu virus’ determine the context in which the co-term ‘Japan’ occurs.

4.2. Evaluation procedure

The following procedure was used to evaluate the automatically generated clusters described above. Recall that we carried out a human-based evaluation, because we wanted to produce information that was meaningful to human interpreters and hence wanted to know whether the clustering procedure was producing information that could be understood by potential end users.

(1) A topic is defined as a seminal event, along with all directly related events (NIST Speech Group, 2005). We adopt the above definition and interpret it in a holistic manner, i.e. a topic is an abstraction of a series of events which are represented by groups of significant terms, which means that a whole has more value than its parts. In this respect, any significant term/an event should not be regarded as the full representation of a topic, but only as a single part of a whole.

From the 24,353 clusters, we manually selected clusters which were associated with 20 large topics (e.g. bird flu, tsunami, US presidential election 2204, global warming). As explained above, each topic is the abstraction of a series of events. A topic was chosen if there were at least 50 clusters which described an important/major event and its subsequent, related events. We used 20 large topics as a starting point to select a set of clusters which signified day’s events. For each group, we then randomly selected 5 clusters; the 100 ‘real clusters’.

(2) From the 100 real clusters, we generated a list of terms which contained all the terms in the clusters. Each term was assigned an index. From this term list, we randomly generated 100 pseudo clusters, each of which had the same length as a counterpart in the real set. These 100 ‘fake clusters’ formed the control group for the experiment.
(3) We shuffled the 100 real and 100 fake clusters and put them into a list. We then asked 10 staff members in our school to judge whether the clusters were good or bad and the results were returned anonymously to the first author. In cases where they were not sure we asked them to abstain. Here, a good cluster is defined to be a cluster which contains contextually related terms, i.e. the terms seem to relate to the same event.

4.3. Evaluation results

We used χ^2-tests to determine whether each individual assessor could tell the difference between the 100 real and the 100 fake clusters. The 10 χ^2 values range from 30.16 to 107.89, all greater than the critical value $\chi^2_{0.01} = 9.21$. Hence, there is a strong evidence to reject H_0. This means that for all the 10 assessors the proportion of good, bad and abstention for the 100 real clusters was significantly different from those for the 100 fake clusters. For the 100 real clusters, the average proportions of good, bad and abstention are 68.60, 14.40 and 17 and for the 100 fake clusters, 16.70, 64.10 and 19.20 and so the effectiveness of the clustering method is clear, although it is not perfect.

In addition, Friedman tests were carried out to determine whether the proportion of good clusters was significantly higher for the real than for the fake clusters across the 10 assessors as a group. Both the good and bad proportions of the real and fake clusters yield the same result: $\chi^2 = 10$ ($p = 0.002$). This confirms that the assessors could differentiate the real from the fake clusters a significant part of the time.

We encountered problems in the evaluation: to find a group of people who had broad knowledge about news stories. Two assessors recorded a very high level of abstentions (33 and 47). This may be because the assessors were not familiar with the particular news stories. Thus, the evaluation results may well underestimate the real performance of our clustering approach, from the perspective of an expert interpreter. Nevertheless, the results clearly indicate that the 100 real clusters are significantly better than the 100 fake clusters.

4.4. Further analysis of results

To better measure the performance of our clustering approach, we counted the number of good clusters for each topic, with regard to the 10 assessors and averaged the 10 numbers, as formally defined below.

$$G_{\text{avg}}(\tau) = \frac{\sum_{a=1}^{n} G(\tau, a)}{n} \qquad (6)$$

Here, a is an assessor, $n = 10$ is the total number of assessors and τ is a topic. The value $G(\tau, a)$ is the number of good clusters with regard to a topic, τ and an assessor, a. The $G(\tau, a)$ values range from 1 to 5 and $G_{\text{avg}}(\tau)$ is the average of all the $G(\tau, a)$ values. Each topic is represented by five clusters, as stated in Section 4.2. The judgement of an assessor was not taken into account if they abstained from all five clusters, which we took to mean that they were not familiar with the topic or were reluctant to do the evaluation. Hence, this method avoids the hypothesised bias from some of the assessors. Given a topic, the best mark would be 5, meaning that an assessor judged all the five clusters related to the topic to be genuine. The lowest mark would be 1, meaning that an assessor only judged one cluster to be genuine. Table 2 shows the $G_{\text{avg}}(\tau)$ of all 20 topics.

The average of all the $G_{\text{avg}}(\tau)$, listed in Table 2 was 3.57 (71.4%). Analogously, we adapted and applied Eq. (6), to compute $B_{\text{avg}}(\tau)$ for bad clusters and $A_{\text{avg}}(\tau)$ for neither good nor bad. The average of all the $B_{\text{avg}}(\tau)$ was 0.75 and the

Table 2
The $G_{\text{avg}}(\tau)$ of all the 20 topics

τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$	τ	$G_{\text{avg}}(\tau)$
τ_1	4.22	τ_5	4.10	τ_9	4.30	τ_{13}	4.30	τ_{17}	3.25
τ_2	3.40	τ_6	3.90	τ_{10}	2.56	τ_{14}	3.00	τ_{18}	3.13
τ_3	3.30	τ_7	3.11	τ_{11}	3.33	τ_{15}	3.80	τ_{19}	3.78
τ_4	4.40	τ_8	3.80	τ_{12}	3.70	τ_{16}	2.78	τ_{20}	3.20

Table 3
The number of real clusters judged bad by the assessors for different level of agreement

#Assessors (level of agreement)	#Real clusters judged bad	Cluster-id
5	6	11, 27, 37, 69, 87, 161
6	3	11, 27, 161
7	2	11, 27
8	1	11
9	0	–
10	0	–

average of all the $A_{\text{avg}}(\tau)$ is 0.68. Taking all the assessors's judgements into account, then the following results were obtained. The average of all the $G_{\text{avg}}(\tau)$ was 3.43 (68.6%). The average of all the $B_{\text{avg}}(\tau)$ was 0.72 and the average of all the $A_{\text{avg}}(\tau)$ was 0.85. Clearly, the average of all the $G_{\text{avg}}(\tau) = 3.43$ indicates a slight bias (2.8%) from some assessors.

In addition, we counted the number of real clusters which were judged bad by the assessors at the x level of majority, where $x = \{5, \ldots, 10\}$. Here, 'x' means that there are at least x assessors who think that a real cluster is a bad cluster. Table 3 lists the results. The six real, bad clusters were individually analysed to find the reason why they were judged bad. Four assessors who judged the six real clusters as bad clusters were willing to give their reasons for their judgement. Table 4 lists the six real clusters. The summary of all the RSS items which are associated with the six real clusters is listed below.

- 11: The experiment which shows that children are more vulnerable than adults, in regard to the adverse effects of air pollution.
- 27: The Vietnamese government took action to prevent bird flu from spreading in Yunnan province in China.
- 37: An experimental finding that shows that ethanol can promote cancer progression.
- 69: Vietnam deployed riot police at bird flu check points around Ho Chi Minh city.
- 87: The basic right to choose with whom someone wants to create their children.
- 161: The shifts in the apparel trade roil [i.e. agitate] the global economy which can threaten the living standards of poor nations.

The reasons for judging the clusters as bad clusters are listed below.

- Cluster 11 was judged bad, because the assessors thought that the term, 'adults', is irrelevant to the adverse effects of air pollution. In this case, only two term co-term pairs were used, as the associated RSS items are quite sparse. Thus, it leads the clustering algorithm to cluster the related co-terms together.
- Clusters: [27, 37, 161] were judged bad, because the assessors thought that the cluster members are contextually not related. This is due to the input data for the clustering algorithm, which was too sparse, even though the associated RSS items contain sufficient data. We only used the co-terms which can be found in the same sentence (a fixed window of observation). We might be able to avoid this problem by widening the window of observation, but this might not work well in another case, as some clusters may contain too many unrelated terms.
- Cluster 69 was judged bad, because the assessors thought that the term, 'riot police', is irrelevant to the other cluster members. There should be no connection between 'bird flu' and 'riot police'. This problem was anticipated from

Table 4
The six real clusters judged bad

Cluster-id	Cluster members
11	Adults, adverse-effects, air pollution
27	Bird flu intrusion, Yunnan-guards
37	Cancer progression, ethanol
69	Bird flu checkpoints, Hanoi Vietnam, Ho Chi Minh city officials, riot police
87	Basic human rights, children, strip
161	Apparel trade roil global economy, shifts, unravels

the beginning; it is difficult to judge two contextually related terms, unless the assessors know the event with which the cluster is associated.

- Cluster 87 was judged bad, because the assessors thought that the term, ‘strip’, is irrelevant to the other cluster members. This is due to a parsing error, as the POS tagger tagged the term as a noun.

In summary, given the six real clusters, there is one real cluster (cluster 69) which is actually a good cluster, in the sense of containing sufficient information to indicate a single event, but was judged bad by the assessors. The root cause of cluster 69 being judged bad is probably that the event described was relatively minor and quite specific with an unusual collection of terms. It is likely that the assessors did not know of the event or had forgotten it. The other five clusters are bad clusters, due to data sparseness, the length of window of observation and parsing errors.

We also found a chain effect of clustering, i.e. two unrelated events clustered together because the input data misleads the clustering algorithm. The following two examples illustrate the problem. Cluster 5 (in Table 5) grouped two unrelated events together. The term, ‘clean water’ in cluster 5 was significant on 28/12/2005 and became the trigger which initiated two different contexts. One was in the context of tsunami tragedy and another one was in the context of John Kerry’s speech about clean water. The term, ‘bloomberg’ (a media service) in cluster 1 (in Table 6) became the trigger which initiated one context, i.e. world crisis, which subsumed the two terms, ‘fuel costs’ and ‘bird flu’.

4.5. Comparison with existing work

Swan and Allan (2000) focussed on generating the overview timelines of significant terms and carrying out a clustering on the significant terms. The Kappa statistic κ (Cohen, 1960; Siegel & Castellan, 1988) was used to measure pairwise agreement among assessors. For $0.67 \leq \kappa < 0.8$, a tentative conclusion can be drawn. For $\kappa \geq 0.8$, a definite

Table 5
An excerpt of an overview timeline for the ‘tsunami’ event

Id	Date	Cluster
1	2004-12-26	Asian-nations years-triggers-tsunami earthquake
2	2004-12-27	Asian-death-toll Asian-disaster death-toll earthquake-tsunami
3	2004-12-27	Australian-red-cross Asia-quake tsunamis-appeal
4	2004-12-28	Desperate-refugees tsunami death-toll-climbs
5	2004-12-28	Asian-tsunami disasters senator-Kerry spread aftermath clean-water
6	2004-12-29	Asian-earthquake tsunami pledge UK-government victims
7	2004-12-29	Basic-equipment Indonesian-tsunami monitoring-system
8	2004-12-30	Bush-administration pledges support tsunami-aid
9	2004-12-30	Aid-efforts Asian-tsunami pledge victims
10	2004-12-30	Tsunami-scientists earthquake-prone-nation public-safety-officials
11	2004-12-30	Banda-Aceh-Indonesia tsunami death-toll-jumps
12	2004-12-31	Asian-tsunami-tragedy aircraft-carrier-battle-groups navy-battle-groups tsunami-relief
13	2004-12-31	Banda-Aceh-Indonesia relief-efforts stricken-area victims aid tsunami-toll-climbs
14	2004-12-31	Basic-requirements basic-services devastating-tsunamis tragedy southeast-Asia
15	2005-01-01	Pope-John-Paul-II special-mass-early-Saturday tsunami-victims
16	2005-01-01	Diseases biggest-threat tsunami-survivors
17	2005-01-07	Aceh-destruction tsunami-destruction Annan
18	2005-01-10	Aid-money tsunami-relief public-tracking-system
19	2005-01-10	Aceh board-crashes relief-operation helicopter-crash
20	2005-01-22	Aceh-province early-warning-system tsunami lives

Table 6
An excerpt of an overview timeline for the ‘bird flu’ event

Id	Date	Cluster
1	2004-12-02	Bloomberg fuel-costs bird-flu
2	2004-12-08	Pandemic Asian-bird-flu virus
3	2004-12-09	Bird-flu-pandemic who governments
4	2004-12-10	Youngest-sars rapid-bird-flu-test Beijing Hong-Kong Xinhuanet-scientists
5	2004-12-18	Bird-flu virus-antibody blood-samples
6	2004-12-18	Japan outbreak bird-flu-virus culling
7	2004-12-19	Bird-flu-fear bird-flu-virus Japan
8	2004-12-20	Human-flu-virus bird-flu pandemic
9	2004-12-22	Avian-influenza first-human-infection bird-flu
10	2004-12-23	Dangers disease human large-scale outbreaks bird-flu
11	2004-12-31	Deadly-bird-flu-virus poultry fresh-outbreaks
12	2005-01-07	Bird-flu-case Vietnam girl
13	2005-01-08	Bird-flu-intrusion Yunnan-guards bird-flu-spreads
14	2005-01-21	Human-bird-flu-transmission birds virus Vietnam
15	2005-01-23	Bird-flu-deaths human-toll Vietnam possible-global-flu-pandemic
16	2005-01-24	Bird-flu-case bird-flu-infections positive-cases Vietnam-reports
17	2005-01-26	World-bird-flu-fear pandemic deaths
18	2005-01-28	Vietnamese-girls bird-flu southern Vietnam
19	2005-01-29	Bird-flu-checkpoints Hanoi Vietnam riot-police ho-chi-mink-city-officials
20	2005-01-30	Bird-flu-evolution Hanoi WHO-experts

conclusion can be drawn (Eugenio & Glass, 2004). Swan and Allan (2000) provided a list of TDT-2 topics, to four assessors and asked the assessors to assign each generated cluster to one/more of the TDT-2 topics. The assessors were allowed to define their own topic, if they were not happy with the topics provided. The percentage of agreement on how many topics each cluster should be assigned was 73.6%, with $\kappa = [0.045 - 0.315]$, and a $\kappa_{\text{average}} = 0.223$. Swan and Allan (2000) state that the low κ values were due to the notion of topic upon which the four assessors could not agree with each other. This is similar to our case, in the sense that the concept of topic is important, but yet inherently ill-defined. When the assessors were given a set of clusters which were already assigned a pre-defined TDT topic and were only asked to indicate whether they agreed, the percentage of agreement on the assigned topics was 86.7%, with $\kappa = [0.6 - 0.785]$, and $\kappa_{\text{average}} = 0.699$ (Swan & Allan, 2000).

In contrast, we focussed on generating overview timelines for major events from their related, subsequent events. We clustered the co-terms of significant terms, as we wanted to cluster all the terms which are contextually related and used the cluster to signify day’s events. We did not ask the assessors to assign a specific topic to each cluster. We carried out the κ test to measure the pairwise agreement among the assessors on judging the 200 clusters. The κ values = [0.21 − 0.60], and $\kappa_{\text{average}} = 0.36$. As all the κ values obtained were <0.67, there is no need to carry out κ test (Eugenio & Glass, 2004), for comparison. Clearly there is a low rate of agreement among assessors.

In our experimental setting, we did not ask the assessors to deal with the notion of a topic. Instead, we asked them to judge whether the cluster contains contextually related terms which may signify a day’s significant event. Clearly, our evaluation focuses on the context between cluster elements and not the mappings between a cluster and topics. We deliberately did not give the assessors the RSS items associated with a cluster or a set of pre-defined events, as it would introduce a bias in judging the cluster. This explains the reason for the low κ values obtained, because event knowledge is an additional complicating factor.

5. Clustering results

We use a qualitative approach to investigate how clusters relate to major events. The objective is to gain insights into the types of information indicated by the clusters and how this may vary by major event type. We automatically selected a portion of clusters which were related to three important events, 'tsunami', 'bird flu' and 'US presidential election', which happened in 2004. Each major event was termed 'an initial event', as it was the beginning of a series of 'subsequent events'. In this context, a subsequent event is an event which carries both of the following.

- The essence of an initial event, as it is directly or indirectly triggered by the initial event.
- Its own meaning, which extends the scope of an initial event.

For each initial event, all the clusters which contained the significant term which signified the event were automatically selected: 'tsunami', 'bird flu' and 'election'. For the 'US presidential election', we also selected the clusters which contained the terms, 'George W. Bush' and 'John Kerry', as they were the most influential presidential candidates in the 2004 US election. To illustrate the way in which we analysed each initial event along with its subsequent events, only 20 clusters for each initial event are listed in Tables 5–7.

5.1. Qualitative analysis

The 20 clusters listed in Table 5 shows an excerpt of an overview timeline for the tsunami of 26/12/2005. By manually analysing the RSS items – on each day – in which each term occurred, we found the following points.

Table 7
An excerpt of an overview timeline for the 'US presidential election' event

Id	Date	Cluster
1	2004-01-23	Election Internets
2	2004-01-24	John-Kerry Howard-Dean Iraq
3	2004-01-24	Election President-Bush White-House election national-security
4	2004-03-03	John-Kerry George-Bush Presidential-election
5	2004-05-07	George-Bush Iraq-prison-abuse-scandal re-election-campaign
6	2004-05-11	Impact president-falling-poll-numbers re-election-chances
7	2004-05-19	Bush-White-House credit twist strategically important-states machinery re-election
8	2004-07-06	Presidential-election chances winning
9	2004-07-11	Presidential-election postponement terror-attack elections
10	2004-07-14	Effect case elections plan later-date terror threat
11	2004-10-20	Impact finances presidential-election
12	2004-10-21	Citizens presidential-election country endorsements huge-deal
13	2004-10-28	Presidential-election voting redskins winner
14	2004-10-30	Minds quiet-issue liberals conservatives presidential-election
15	2004-11-01	Presidential-elections outcome year
16	2004-11-03	Outcome George-Bush re-election presidential-election
17	2004-11-06	Electronic-touch-screen voting machines
18	2004-11-15	Prospect election-result president
19	2004-12-14	Americans election real-determining-factor
20	2005-01-20	George-Bush inauguration US-president

- The estimation of the scale of tsunami. Cluster 1 highlighted the estimation of the scale of the initial event that affects five Asian nations.
- Assessing the need to install a tsunami monitoring system. Clusters: [7, 10, 20] were a specific issue about the need of having a tsunami monitoring system to prevent the tragedy in the future.
- The consequence and problems which occurred due to the tsunami. Clusters: [2, 4, 5, 11, 13, 14, 16] described major problems in the aftermath of the tsunami, such as the death toll, disease, refugees and the lack of basic requirements and services, such as clean water. Cluster-id 19 was an isolated incident about a helicopter crash near the Banda Aceh (Indonesia) airport during the relief operations.
- The actions taken to help tsunami victims. Clusters: [3, 6, 8, 9, 12, 18] referred to the Australian red cross appeals for donations and the UK and US government pledges to help the tsunami victims. For example, the US government sent navy aircraft carrier battle groups to deliver relief aid (cluster 12) and a public tracking system was set up to organise aid money (cluster 18). Clusters: [15, 17] highlighted two prominent public figures, Pope John Paul II and Mr. Kofi Annan (the secretary general of the United Nations), who expressed their condolences to the tsunami victims.

The 20 clusters listed in Table 6 shows an excerpt of an overview timeline for the spread of bird flu from 12/2004 to 01/2005.

- The global pandemic hypothesis. Clusters: [2, 5, 8, 9, 10, 11, 14, 15, 17, 20] indicated a predictive assessment from the World Health Organization (WHO) and scientists about the possibility that bird flu could be transmitted from human to human and become a global pandemic.
- The major problems which were raised due to the virus spreading. Clusters: [1, 5, 7, 9, 10, 12, 14, 15, 16, 18] showed the continuing fears and problems faced by the affected Asian countries, i.e. the danger of the bird flu virus for human life and national economies.
- The actions taken against the bird flu virus. Clusters: [3, 11] referred to WHO warnings of a possible bird flu pandemic. Cluster 4 was about a Chinese scientist who carried out a rapid bird flu test. Clusters: [6, 13, 19] referred to the actions which were taken to prevent the virus from spreading.

The two initial events, ‘tsunami’ and ‘bird flu’, are different in the way in which the initial event occurred. Within 01/12/2004–31/01/2005, the tsunami only occurred once, but had a devastating impact. In contrast, bird flu occurred more than once and periodically claimed human lives. Five (clusters: [5, 9, 10, 14, 15]) of the 20 clusters described both the danger of the bird flu virus and the global pandemic hypothesis, as the WHO and scientists on several occasions made predictive assessments, when bird flu claimed human lives. In contrast, in the tsunami example, the clusters were more about the problems raised due to the tsunami attack and the actions in bringing relief aid to stricken areas, organising financial support for reconstruction and addressing critical issues, such as the need for clean water.

Table 7 shows an excerpt of an overview timeline for the ‘US presidential election’ event. The majority of the clusters contained a number of events in which the presidential candidates made a political decision or a political judgement which led to a number of questions – concerning the election – as to whether:

- The influence of Internet on the election had been over hyped (cluster 1).
- John Kerry’s judgement over Iraq was right (cluster 2).
- John Kerry could win (cluster 4).
- Iraq prison abuse could affect the election result (cluster 5).
- President Bush’s falling poll ratings could have an impact on his re election (cluster 6).
- John Kerry’s choice of John Edwards could help or hurt his chances of winning (cluster 8).
- The presidential election should be postponed due to terrorist threats (clusters: [9, 10]).
- The presidential election would have an impact on individual finances (cluster 11).
- The presidential election would have an impact on the citizens of other nations (cluster 12).
- The outcome of the Washington Redskins football games could correctly predict the winner of this presidential election (cluster 13).
- Catholic consciences would recall the abortion issue in this election (cluster 14).
- American people were happy with the outcome (cluster 16).

- The voting machine worked well (cluster 17).
- President Bush would bring a new generation of conservative justices to the Supreme Court (cluster 18).
- The European people viewed the election differently (cluster 19).

Clusters: [2, 3, 7, 8, 15] referred to different types of political judgements and decisions. Clusters: [16, 20] referred to the outcome of the US presidential election 2004 and inauguration in 20 January 2005.

At the beginning, the US presidential election event triggered many questions and some responses and the outcome of the election occurred at the end. The election and bird flu events had the same characteristics, in the sense of triggering predictive assessments in a particular situation. This indicates the nature of the two events which have an inherent uncertainty about their final outcome, animal-human/human-human virus transmission for bird flu.

5.2. Discussion

Instead of operating at the document (i.e. RSS item) level, we operated at the level of features and selected a portion of significant terms. The terms formed a basis for further data processing. By using co-occurrence with significant terms, for each day we clustered all the terms which were contextually related. The generated clusters, however, were sometimes incorrect, as discussed in Section 4.3, and our approach did not achieve a very high level of success. Despite these weaknesses, our approach can be applied to generate overview timelines for major events, as described in Section 5.1. The case studies show that the clusters were genuinely related to the major events and could be used to form a narrative, albeit partial.

The human-based evaluation and analysis, as explained in Sections 4.2 and 5.1, are useful for determining the degree of the effectiveness of our approach from human interpreter's perspective. An automatic, metric-based evaluation, such as described in Ng and Han (2002) or Alexandrov et al. (2005), would have also been useful, if we would have had a set of predefined clusters as a gold standard, so that we could measure the cluster quality by measuring the number of overlaps between the predefined and automatically generated clusters. In our experimental setting, we did not have the predefined clusters with which our results could be compared. For these reasons, it was not possible for us to conduct an automatic evaluation.

There are some delays in identifying events, when compared with the other news media. An individual may use an RSS feed to express and post their comments and the propagation of news from a media source to the individual may take time. For example, the intention of Pope John Paul II to offer a special Mass on New Year's Eve for the tsunami victims was reported by Catholic World News (CWN) on 31/12/2005. Our cluster which is associated with the event was dated at 01/01/2005 (1 day later).

6. Conclusions and future work

Our method automatically produced clusters of terms from RSS feeds, which were assessed by human evaluators to see whether they appeared to signify a single news event. The low level of agreement among the 10 assessors ($\kappa_{\text{average}} = 0.36$) indicated the difficulty of the human task of reliably identifying an event from a small set of terms rather than problems with the clustering algorithm itself. The evaluation of 100 real clusters carried out by the assessors indicated that the average percentage of good clusters was 68.6%, which was much higher than the 16.7% for bad clusters. Thus, the method was clearly effective to some extent ($p < 0.01$). Our clustering approach, however, produces some incorrect clusters, as well as some clusters that it would be unreasonable to expect a non-expert to identify.

Despite these issues, our clustering approach can be applied to generate overview timelines for major events. The case studies show that the method can be applied to obtain coherent, human identifiable events and form a narrative, albeit partial. In future work, we hope to adopt our approach to fully automatically generate a short summaries of major events and their subsequent events from RSS feeds.

Acknowledgements

The work was supported by a European Union grant for activity code NEST-2003-Path-1. It is part of the Critical Events in Evolving Networks project (CREEN, contract 012684). We thank the reviewers for their helpful comments.

References

Adar, E., Zhang, L., Adamic, L. A., & Lukose, R. M. (2004). Implicit structure and the dynamic of blogspace. In *Proceedings of the 13th international WWW conference – workshop on weblogging ecosystem – aggregation, analysis and dynamics*.

Alexandrov, M., Gelbukh1, A., & Rosso, P. (2005). An approach to clustering abstracts. In *Proceedings of the 10th international conference on applications of natural language to information systems (NLDB 2005)* (pp. 275–285).

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 37–45).

Baeza-Yates, R., & Ribeiro-Netto, B. (1999). *Modern information retrieval* (1st ed.). ACM Press/Addison Wesley.

Belew, R. K. (2000). *Finding out about—a cognitive perspective on search engine technology and the WWW* (1st ed.). Cambridge University Press.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of the 3rd conference on applied natural language processing (ANLP 1992)* (pp. 152–155).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Eugenio, B. D., & Glass, M. (2004). The kappa statistics: A second look. *Computational Linguistics*, *30*(1), 95–101.

Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. In *Proceedings of the 13th international WWW conference – workshop on weblogging ecosystem – aggregation, analysis and dynamics*.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international WWW conference* (pp. 491–501).

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii international conference on system sciences (HICSS-37)*.

Lin, J., & Halavais, A. (2004). Mapping the blogosphere in America. In *Proceedings of the 13th international WWW conference – workshop on weblogging ecosystem – aggregation, analysis and dynamics*.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms* (2nd ed.). Cambridge University Press.

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016.

NIST Speech Group. (2005). The topic detection and tracking phase 2 (TDT2) evaluation plan. [http://www.nist.gov/speech/tests/tdt/tdt98/ (accessed 15 June 2005)].

Prabowo, R., & Thelwall, M. (2006). A comparison of feature selection methods for an evolving RSS feed corpus. *IPM*, *42*(6), 1491–1512.

Ramshaw, L. A., & Marcus, M. P. (1995). Text chunking using transformation-based learning. In D. Yarovsky & K. Church (Eds.), *Proceedings of the 3rd workshop on very large corpora (VLC 1995)* (pp. 82–94).

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval* (1st ed.). McGraw-Hill, Inc.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.

Siegel, S., & Castellan, J. N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–29).

Smith, D. A. (2002). Detecting and browsing events in unstructured text. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 73–80).

Stein, B., & Niggemann, O. (1999). On the nature of structure and its identification. In P. Widmayer, G. Neyer, & S. Eidenbenz (Eds.), *Proceedings of the 25th international workshop on graph-theoretic concepts in computer science* (pp. 122–134).

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, & P. Ingwersen (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 49–56).

Thelwall, M., Prabowo, R., & Fairclough, R. (2006). Are raw rss feeds suitable for broad issue scanning? A science concern case study. *JASIST*, *57*(12), 1644–1654.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study on retrospective and online event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 28–36).