Web Mining for Understanding Stories through Graph Visualisation

Ilija Subašić Bettina Berendt
K.U. Leuven, Department of Computer Science, Belgium
firstname.lastname@cs.kuleuven.be

Abstract

Rich information spaces (like the Web or scientific publications) are full of "stories": sets of statements that evolve over time, manifested as, for example, collections of newspaper articles reporting events relating to an evolving crime investigation, sets of news articles and blog posts accompanying the development of a political election campaign, or sequences of scientific papers on a topic. In this paper, we propose a method and a visualisation tool for mapping and interacting with such stories. In contrast to existing approaches, our method concentrates on relational *information and on* local *patterns rather than on the occurrence of individual concepts and global models. In addition, we present an evaluation framework. A real-life case study is used to illustrate and evaluate the method and tool.*

1 Introduction

The Web has led to a proliferation of news (and other broadcast media like blogs) that continuously report on current events and other topics. Several search-engine innovations of the past few years like the grouping of news articles by topic in Google News have made it easier to keep abreast when one reads the news every day. However, a Web user who misses several days or who wants to gain an overview of major events and developments in a "story" that lies in the past, is today faced with a situation that is reminiscent of the early days of the Web. Search in most archives is based on keyword search and therefore returns an unmanageable number of results. Summarisations like that provided by Google Trends[1] or BlogPulse's Trend Search[2] show surges in publication and query activity in certain time periods, but these tools require that one knows which sub-topic to look for (and how to describe it in keywords).

The same problem arises in other areas with high publication intensity and readers who aim to gain, refresh, and/or extend overviews of topical developments – scholarly publications are a prime example.

This situation calls for systems that (a) identify topical sub-structure in a set (generally, a time-indexed stream) of documents constrained by being about a common topic, (b) show how these substructures emerge, change, and disappear (and maybe re-appear) over time, and (c) give users intuitive interfaces for interactively exploring the topic landscape and at the same time the underlying documents. In an extension of [20], we call the resulting problem *evolutionary theme patterns discovery, summary and exploration* (ETP3).

The first contribution of the paper is a (re-)appraisal of the ETP3 problem as one that requires a semi-automatic solution, and a proposal for a system that offers such a semi-automatic solution. Specifically, we believe that such a system should not be overly prescriptive. In particular, the user's interpretation of subdivisions within a topic will depend on her current tasks and other situational variables. We therefore aim, in contrast to the existing approaches, not at a global model of the topic (such as a clustering into exhaustive sub-topics); instead, we are interested in high-resolution local patterns and interaction options that support users in finding and exploring their own interpretations. The second contribution is an evaluation framework for ETP3 and a demonstration using a case study.

The paper is structured as follows: In Section 2, we give an overview of related research. Sections 3 and 4 present our solution approach "STORIES": Section 3 describes the computational method and Section 4 the tool. A case study demonstrates method and tool in Section 5. Section 6 describes the evaluation method and results. Section 7 concludes with an outlook.

2 Related work

Our work builds on several areas of research, in particular the identification and tracking of topics in text streams, the identification of "bursty" events, the use of co-occurrence information for content extraction, and information visualisation.

[1] `http://www.google.com/trends`

[2] `http://www.blogpulse.com/trend`

1550-4786/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDM.2008.138

Temporal text mining. [20] described evolutionary theme pattern discovery as one key subproblem of temporal text mining. They presented a fully automatic method that extracts subtopics and creates a graph that shows their life cycles and dependencies on each other. A mixture model was used to model documents as expressing (potentially several) themes (corresponding to sub-topics). These word clusters are tracked over time. The use of clustering models for finding emergent sub-topics and tracking them over time is also the subject of [24, 15]. The publications show that the methods can be applied to scholarly publications as well as Web news. Evolutionary theme pattern discovery is related to topic detection and tracking, specifically first story detection [2]. However, it is more fine-grained than TDT since it delves into a topic's substructure, and its aim is not only to classify something as a new (or old) topic, but to describe it. Evolutionary theme pattern discovery is also related to the document update problem in text summarisation, which is discussed in more detail in Section 6.

These methods rely on the notion of sub-topics that cover the space of reported content, such that it is difficult to identify local details and their changes over time.

Burstiness. (Sub)topics may be particularly interesting when they are *bursty* [16], i.e. when publication activity on them is very strong in a certain time period, picking up volume fast at this period's beginning and (usually) disappearing again as fast. Burstiness has been explored with respect to various domains and phenomena including "buzz" in text and news streams [11, 12, 13]. [11] group "bursty features" into "bursty events" based on co-occurrence, thereby creating an analogue of sub-topics.

So far, burstiness has only been investigated as a feature of single text features (words or topics). We extend this to an analysis of burstiness of associations.

Co-occurrence analysis. The analysis of bursty events points to the merits of focussing on specific parts of contents and their relations with each other, rather than on finding a global model. In general, the analysis of co-occurrences allows for a more fine-grained analysis of texts and has been investigated for example in text summarisation. [3] show that *topic signatures* [19] provide a simple and effective way to summarise multiple documents. [25] used co-occurrences to find historical associations between places and times in a digital library. He analysed how various interestingness measures rank these associations and showed that they behave differently, for example in the ranking of rare events. This indicates that different interestingness measures may be more or less adequate for the analysis of different corpora, domains and/or different tasks, an interpretation also supported by the findings of [10]. These authors found co-occurrence lift to be an adequate interestingness measure to analyse perceptions of (car) brands and markets in user forums.

[7] propose application-domain interpretations of temporal changes in the frequencies of co-occurrences. They argue that agents (person names in the texts) can exist independently of each other, join, split again, etc. These developments create specific "story lines".

All these approaches are restricted to analysing co-occurrences between typed elements (names, places, ...). We take a more general approach and identify "story lines" between arbitrary words or concepts.

[1] applied text summarisation to news streams, their focus was however more on finding the best sentences to be (re-)used in the summaries than on distilling concepts from these sentences. In contrast to this work, we focus not only on content that is new (i.e., different from what was reported before), but on content that is characteristic for a time period (i.e., also different from what was reported later).

Visualisation. The main focus of most of the above studies were challenges (a) and (b) mentioned in the Introduction. Visualisations are probably best suited to displaying the complex relationships found. [25] provided users with an interactive map browser for exploring the location-time co-occurrences. This is a good example of how to meet challenge (c) in a way that is adapted to the application domain. [30] show a domain-independent way of visualising pairwise associations of words that also takes into account when these associations were strong. They plot words against time and show co-occurrences by connecting lines in a format that is related to parallel coordinates. Their graphs provide an excellent overview of the occurrence or recurrence of pairwise associations over a whole timeline. However, because time takes up one visual dimension, higher-order patterns of associations cannot easily be detected. In contrast to this, we will show associations per time point/period. This "snapshot" idea is the same as that used in the graph sequences used for visualising scientific publications and topics by, e.g., [5, 6, 15]. In contrast to that, we use a layout strategy that is more amenable to highlighting emerging and disappearing topics, and offer the alternative of a dynamic layout between successive time periods (morphing), similar to [17].

3 The STORIES method

The basic assumptions of our method are that (a) there is a set of time-stamped documents that, when read by a human reader, reveal the story and its evolution and (b) the words in these documents also reveal the story and its evolution when processed by simple text mining methods. We conceptualise

- *story basics* as the high-ranking terms (words, compounds, named entities, concepts, ...) from all documents of a corpus of relevant documents, where the ranking reflects the importance of these terms in the corpus,

- *story elements* as the high-ranking relationships between story basics, where the ranking reflects the importance of these relationships in the corpus,
- *story stages* as networks of salient story elements in a certain time period, where salience is measured based on co-occurrence frequency and its relevance in a current time window and in the whole corpus,
- *story evolution* as the temporal sequence of story stages.

This basic scheme can be operationalised in several ways. To create a baseline, we have started with very simple versions of each of these constructs' operationalisations. Specifically, the method involves the following stages. First, a corpus of text-only documents is transformed into a sequence-of-terms representation. Subsequently, basic term statistics are calculated to identify candidates for story basics. We chose *term frequency TF* for the whole corpus, which is defined as *(# occurrences of the term in the whole corpus) / (# all terms in the whole corpus)*. We define the *content-bearing terms* as the 150 top-TF terms.

Next, the whole corpus T is partitioned into sets of documents that were published in time periods following one another, e.g. within one calendar week. Thus, T is the union of all document sets t_i, with $i = 1, \ldots, I$ the time periods.

For each t_i, the *frequency* of the co-occurrence of all pairs of content-bearing terms within a window of w terms in documents is calculated as follows:[3]

$$freq_i(b_1, b_2) = \frac{\textit{\# occ.s of both } b_1, b_2 \textit{within } w \textit{ terms in doc.s from } t_i}{\textit{\# all doc.s in } t_i}.$$

This measure of frequency and therefore relevance is normalised by its counterpart in the whole corpus to yield the measure *time relevance*:

$$\mathit{TR}_i(b_1, b_2) = \frac{freq_i(b_1, b_2)}{freq_T(b_1, b_2)}. \qquad (1)$$

This measure is based on the *domain relevance* metric [21] which measures the relevance of a term in a (subject-domain) subcorpus relative to the whole corpus. When used, as here, for time-specific subcorpora, it also measures "burstiness". Thresholds are applied to avoid singular associations in small sub-corpora and to concentrate on those associations that are most characteristic of the period and most distinctive relative to others . We define two sets *N(on-singular)* and *C(haracteristic)*:

$$N_i = \{(b_1, b_2) | (\textit{\# co-occurrences of } b_1, b_2 \; \textit{within } w \textit{ terms in articles from } t_i) \geq \theta_1\} \qquad (2)$$

$$C_i = \{(b_1, b_2) | \mathit{TR}_i(b_1, b_2) \geq \theta_2\} \qquad (3)$$

for some thresholds θ_1, θ_2. This gives rise to

- the *story stage* i: $N_i \cap C_i$. This can also be expressed as a graph with terms as nodes and associations as edges.
- the *story elements*: all edges of the story stage.
- the *story basics*: all nodes of the story stage.
- the *story evolution*: the sequence of story stages.

To obtain a smoother story evolution, we use the moving average of co-occurrence frequency values. This was done by replacing for each period t_i, the document base set in both numerator and denominator of the right-hand side of the*freq* definition by the union over periods $i, \ldots, (i+l-1)$.

Investigations of different parameter settings showed that in most cases, only associations with $\mathit{TR} > \theta_2 = 3$ are interesting and allow for a tractable graph. However, the advantage of an interactive approach is that we can let the user explore different values of θ_2 and thereby create their individual story stages. Visualisation options (see Section 4) help to accentuate the differences in time relevance. Users are also able to control θ_1.

4 The STORIES tool

The method can be applied to textual documents such as news obtained from the Web. In this section, we describe the data cleaning and further pre-processing applied to this kind of data.

Data cleaning represented a challenging first step in data preparation. Virtually all news sources present their content in Web pages with a multitude of other content: navigation menus, advertising, ... The best approaches developed so far, such as [9], essentially suggest to learn a wrapper by comparing different articles from the same source; the idea is that this will identify the "noise" by equality over different "content" pages (the "content" should be the only subtree in the DOM tree that changes). Unfortunately, this turned out to not work for many of the sources we investigated, because several elements of the DOM tree change across different articles, even if published on the same day in the same content area. We therefore included an automated wrapper-induction component in the tool; however in order to not conflate data cleaning issues with content

[3] This measure takes into account multiple co-occurrences within one document, in contrast to the *support* measure which uses the number of documents containing the co-occurrence as numerator. Prior tests showed that support did not find out salient co-occurrences well enough.

extraction issues, in the case study below we extracted the content into ASCII by manual copy-and-paste.

Text pre-processing. The documents were first tokenized; subsequently, several further pre-processing options were investigated. Named entity recognition (NER) was done as a two-phase process. In the first phase, the Open Calais[4] semantic toolkit was used to extract NEs. Pilot tests showed that pronoun resolution did not work well on our materials; therefore pronoun resolution was filtered out using a stopword list. Since Open Calais operates on a per-document basis, it cannot map a term to named entities if the named entity does not appear in the document that is currently inspected, despite the fact that in the entire corpus the same term is mapped to the same named entity. To overcome this problem, in the second phase, each term x that was mapped to some named entity in at least one document in the first phase, was treated as follows: Let $x_1, ..., x_n$ be the NEs to which x was mapped in the first phase. Let x_{max} be the NE from $x_1, ..., x_n$ to which x was mapped most often. Then, in each document containing x but not x_{max}, we map x to x_{max}. (A similar NER solution was proposed by [8].) This was followed by lemmatization using the TreeTagger[5]. Stopwords were removed using the stopword list from the Terrier project[6], manually enhanced by HTML-code and application-specific words.

All parameters for text pre-processing can easily be configured, and the architecture provides the needed modularity for, e.g., using different interestingness measures and thereby re-using and/or evaluating other proposals for temporal text mining.

The graphical usage interface We implemented the method in a series of php scripts interacting with a MySQL database, and generated visualisations using GUESS[7]. The visualisations comprise static visualisations of the story stages of individual periods, and a morphing sequence that traces story evolution through the sequence of all periods. In addition to this "scanning", users can "(un)zoom" by adapting the period-window size l.

The visualisations are enhanced by salience slide rulers that allow the user to filter out story elements below individually set θ_1 (absolute number of occurrence of an association) or θ_2 (time relevance) thresholds. A configurable colour scheme accentuates time relevance differences. The figures included in this paper use a sequential scheme optimised for printing, going from black (high *TR*) to light grey (low). For on-screen viewing, different users expressed preferences for sequential schemes using other colours or for divergent schemes, in particular the harmonious colours from blue (high *TR*) via red (medium) to yellow (low). By clicking on an edge, the user gets a list of documents containing that co-occurrence in a browser window. All programs can be executed on a local computer; after an initial download of documents. A screenshot is shown in Fig. 3. In the remaining figures, the graphs have been extracted from the tool environment for better legibility.

5 A case study

For demonstration, we used a real-life story with a comparatively clear and well-known course of events: the disappearance of Madeleine McCann on May 3rd, 2007, and the development of the criminal investigation.[8]

Two main events in this investigation were the early suspicion of a man with the initials R.M.[9] as kidnapper, the discovery of Madeleine's blood in a car rented by the parents (established as hers by a DNA test) and the associated police questioning and suspicion of Madeleine's parents. These were interspersed by long periods of less media attention with little to report (or misleading incidents like the arrest of two people unrelated to the case).[10]

The corpus. We used articles from the Google News archive[11] between May and December 2007 (week 17 in which the girl disappeared until week 52) and restricted the results as follows: only English-language articles; for each month, the first 100 hits, and of those, only those that were still freely available in April 2008. After a first round of analysis, these were restricted to documents from weeks 17–37, the "eventful" weeks of that story. This resulted in a corpus of 215 documents. This was regarded as a good approximation of the real-life situation confronted by a deployed STORIES algorithm: Articles are found to be candidates based solely on keyword matching (in this case: using the first and last name of the missing girl as the query in the Google News archive), they come from sources of varying quality, and there is no ranking on the news sources in the Google News archive after some months.

[4] `http://www.opencalais.com`

[5] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

[6] `http://ir.dcs.gla.ac.uk/terrier`

[7] `http://graphexploration.cond.org`

[8] We wish to emphasise that in no way do we want to capitalise on the sad story of a missing child. However, in the present case, media attention was specifically asked for, at least in the beginning: Madeleine's parents established an unprecedented media campaign to ensure that any hints that anyone might have would be reported. On the first anniversary of Madeleine's disappearance, the family used the Web site to ask for an end of media attention. It is unfortunate that personal and public catastrophes seem to lend themselves most easily to automated story analysis, witness for example the 2005 London bombings (e.g., [26, 23]) or the 2004 Tsunami [20].

[9] In the text and figures of this article, we have anonymized all person names except that of the missing girl, which we need to report to identify our data, and consequently also her family name.

[10] All three suspects were cleared later; and the case was closed in July 2008, see [29].

[11] `http://news.google.com/archivesearch`

This set was extended by the set of all retrievable, English-language news articles referenced in the Wikipedia article [28], from the investigated time period. This provided another 91 articles. This selection constitutes a kind of opposite extreme of the first document selection, because the occurrence of an article in the reference list indicates that its content passed a manual quality control and was integrated into the Wikipedia article. Due to the collaborative authoring of the Wikipedia article, this selection can also be said to represent a wide variety of viewpoints and (potentially) consensus on the quality of the individual articles.

The combined corpus contained 306 articles with 174,886 words, resulting in an average of 572 words per article. The corpus contained 8,075 (6,089) unique words (lemmas).

Results Figures 2–5 show selected individual story stages.[12] In particular, Fig. 2 shows the *description of an event* (missing British child MM). Figure 3 illustrates how the key first suspect becomes an (also visually) *"central" element* of the story. Figures 5 (a) and (b) show how the interface is used by changing the threshold in order to *"uncover" a story stage*. Specifically, Fig. 5 (b) explains some of the reasons for the connections in Fig. 5 (a). An *eventless period* in a story is characterised by a small number of disconnected subgraphs like the ones in Fig. 4.

6 Evaluation

Temporal text mining is still a young area, so unlike for example in TDT, no standards exist yet for evaluating approaches, and the existing literature often restricts itself to plausibility checks. Therefore, the quest for an evaluation of the STORIES approach involves finding answers to the following questions:
(1) Can an existing evaluation framework and/or dataset be used as benchmark?
(2) How should the ground truth be defined?
(3) Can evaluation be (partially) automated to cut human evaluators' workload?
(4) What instructions should human evaluators get?
(5) How can the results be interpreted?
We address questions (1)–(5) in turn.

(1) Existing evaluation frameworks

The ETP3 problem can be decomposed into two subproblems: Evolutionary theme patterns discovery and summary on the one hand, and evolutionary theme patterns exploration on the other hand. Evolutionary theme patterns discovery is related to the *update task* first formulated in the Document Understanding Conference (DUC) 2007: "The update summary pilot task will be to create short (100-word) multi-document summaries under the assumption that the reader has already read a number of previous documents."[13] This contest supplied a test corpus of news stories (documents assigned to 10 topics, each divided into 3 time periods, were supplied), summaries of the updates manually generated by 4 independent human raters, and detailed evaluation reports (precision, recall and F1) of baselines and all the contenders. The evaluation reports were generated with the ROUGE software ("Recall-oriented understudy of gisting evaluation") that was kindly provided to us by its creator Chin-Yew Lin.

The DUC/ROUGE concept is not directly applicable to STORIES because it assumes that the summaries are natural-language texts, whereas we generate graphs. Yet, we created a way of applying the ROUGE evaluation framework and software to our representation (see (3) below).

However, the DUC/ROUGE dataset cannot be used to benchmark our approach. The reason is that the dataset is not a stream (a large set of documents following in quick succession and with usually relatively small differences to the previous one). Rather, it is (for each topic) a set of 3 small-cardinality (usually below 10) sets of documents that were published in 3 disjoint and subsequent time periods, but have very little connection to the other 2 periods' content. This resulted in *all* interesting co-occurrences being "bursty" in each of the 3 periods, resulting in an impossibility to select the really important ones.

Another candidate dataset is the Tsunami dataset used and provided by [20][14]. However, since it is not associated with a ground truth and since it is not straightforward to compare the output of STORIES with the output of the method of [20], this dataset could not be used either.

We therefore decided to use our own case-study dataset and to concentrate on defining a method for evaluation.

(2) Finding a ground truth

One of the biggest problems of finding a ground truth is that in many text tasks, the agreement between human raters is not very high [e.g., 27]. Thus, it is necessary to have a ground truth that reflects a wide range of human raters. In some evaluation frameworks, this goal is achieved by employing several ground truths by different people (e.g., 4 in the DUC/ROUGE evaluation, see (1) above).

Fortunately, the Web itself provides us not only with streams of news, but also with documents that come close to the goal of a multi-rater truth. Specifically, Wikipedia articles (especially those on contested themes such as our case study) are generally written and revised by hundreds of authors, cf. for example [4].

Wikipedia articles are often very long, full of detail, and only occasionally written with a story progression in mind.

[12] Preprocessing with $w = 5, l = 2$. Visualisations of the corpus with θ_2 adjusted for maximum visibility and $\theta_1 = 5$ throughout.

[13] http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html\#pilot

[14] http://sifaka.cs.uiuc.edu/~qmei2/data.html

Therefore, such a document must be transformed in order to serve as a ground truth to be used in a (machine or human) evaluation. We proceeded as follows: First, all sentences that contained a date were extracted from the article. To minimise bias and errors, we had two independent raters extract these sentences (and if necessary perform minor reformulations to make them understandable out of context). Only those assertions that both found in the text, plus a maximum of five others from each rater, were included in the final set of ground truth assertions. In the case study, this resulted in a total of 31 ground-truth *events*.

All ground-truth events were indexed by the calendar week in which they had occurred, such that they could later be assembled easily into the ground-truth of the time window (e.g., 3 weeks) that was covered by the method.

(3) Partially automating the evaluation

The goal was to present both the STORIES output and the ground truth to human evaluators in order to determine precision, recall and F1 values. However, in a pilot study with human raters, we had found that this was a very laborious task and could not easily be repeated for different settings because after seeing the first setting, a human rater knows the story.

Therefore, prior to presenting people with the ground truth and the algorithm output, the best parameter setting had to be found. Recall from Section 3 that the method has as parameters l (the number of weeks that make up a story stage, where story stages are overlapping when $l > 1$), w (the window size within the texts that is inspected for co-occurrences), θ_1 (the minimum total number of co-occurrences), and θ_2 (the minimum time relevance of a co-occurrence). The pilot study had also suggested that for humans to be able to read the graph, the *cardinality of the story stage* (the number of edges of each graph, $|StoryStage|$) should be limited.

We varied $l = [1, 3]$, kept $w = 5$ and $\theta_1 = 5$ based on common values found in the literature, and, starting from a value of $\theta_2 = 2$, varied $|StoryStage|$ from 10 to 30, in increments of 5. $\theta_2 = 2$ was an intuitive value based on the pilot study ("at least twice as frequent in this period than on average"), and 10 to 30 was considered to be a realistic range for human graph reading usability.

To evaluate this large number of combinations, we used the findings of [18], who showed that the automated word-pair matching rules of ROUGE correspond to human ratings in the following sense: The *ranking of quality assessments* by humans corresponds to the ranking of quality assessments by ROUGE. This does *not* necessarily mean that the *absolute values* of precision or recall correspond to each other. However, it means that the setting with the best ROUGE results should be chosen for presentation to humans.

ROUGE evaluates natural-language texts against other texts (the ground truth). STORIES outputs graphs instead of natural-language texts. These two forms of representation are not directly comparable (see for example [14]); however, pilot tests showed us that people interpret paths in the STORIES graphs in a similar way as sentences. We therefore used the following heuristic: We extracted all paths from each STORIES graph[15] and ordered them by descending average *TR* path weight. We then truncated these "pseudo-sentences" at 100 characters to generate ROUGE-style summaries.

ROUGE has different evaluating functions. In our case, the applicable ones were ROUGE-1 (overlap of unigrams), which however always favoured larger graphs – a result that is in conflict with the usability requirement of smaller graphs. The only other applicable function is ROUGE-SU4, which measures the overlap of skip-bigrams of at most length 4. A skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram cooccurrence statistics measure the overlap of skip-bigrams between an automatically generated text and a ground-truth text. We used the ROUGE parameter values that were employed in the DUC 2007 update task evaluation.

The resulting ranking for the corpus was (pairs denote $|StoryStage|$ and l): $20-2, 30-3, 25-2, 15-3, 20-3, 25-3, 10-3, 15-2, 30-2, 10-2, 10-1, 25-1, 30-1, 15-1, 20-1$. Thus, 20 edges and a window of two weeks produced the best story stage descriptions.

(4) Procedure of the manual evaluation

Two raters from different backgrounds, both with a good command of English and only a superficial knowledge of the story, volunteered to rate the STORIES summaries. They were given the 20-2 graphs for a temporal subset of the case-study corpus. The raters received the GUESS software and the graphs together with a driver script, an Excel sheet with one tab for each time window of 2 weeks and one ground-truth event per line, and a set of instructions. They then worked individually at their own pace.

The raters were asked to inspect the graphs in temporal order using the slide ruler for θ_2, starting from a high value so as to "uncover" the graph, and to rate the first 20 edges as follows: If it describes an aspect of an event, then annotate the event with the edge number (visualised by a change in the script). If multiple matches seem appropriate, multiple annotations should be made. The raters were asked to stop when they reached 20 edges or before if the graph "stopped making sense". The filled-in Excel sheets were the basis for the following quantitative evaluation.

(5) Results and interpretation

The measured outcomes were precision at $n = 5$, 10, 15 and 20 (since edges were numbered, the top-TR edges could easily be identified) and recall at the same n, the latter

[15] using an adapted version of the Gaston software [22]

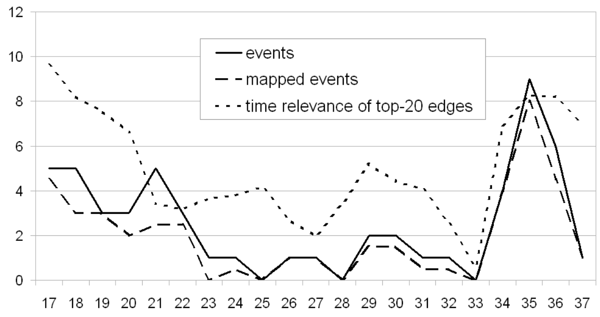


Figure 1. Events (ground-truth), edges and their burstiness profile (average TR), and represented events.

defined as the number of correctly retrieved events for the top-n edges. This notion of recall differed slightly from the standard one because the number of events differed between periods (such that 1 mapping edge would give rise to a recall of 0.2 in a period with 5 events, but 0.5 in a period with 2 events). Some graphs contained fewer than n edges, for these, precision and recall at n are not defined.

The results are shown in Table 1. They illustrate that (a) the judgements by both raters were highly similar in terms of overall quality; (b) recall was quite high – on average, nearly half the events were found in a graph as small as 5 edges, and over 80% in a graph of 20 edges; (c) precision was uniformly acceptable but strictly lower than recall (about one third of all edges were content-bearing). The relatively large values of the standard deviation indicate that the quality of representation varies by week. To investigate whether certain "ground-truth event patterns" cause these variations, we illustrate in Figure 1 more detail by plotting the number of events against the number of events that were represented in the graphs (for 20 edges, averaged over the two raters). It indicates that the number of events does *not* influence the quality of representation as measured by recall. The figure also plots the average *TR* values of the top 20 edges; they too follow the same pattern as the events. Thus, the burstiness measured by TR is a good measure of ground-truth "eventfulness".

Only in week 25 is there is a marked difference between a high average TR and a low number of events. The reason is that in week 26 (which affects 25 due to $l = 2$), a couple had been implicated and arrested. These soon turned out to be con artists who had nothing to do with the case. The incident is not reported in Wikipedia (which we used as "ground truth"), but made headlines at the time.

7 Conclusions and outlook

This paper has presented a new problem in the area of temporal text mining: the tracking of story evolution. More specifically, the *ETP3* (evolutionary theme patterns discovery, summary and exploration) problem consists of (a) identifying topical sub-structure in a set (generally, a time-indexed stream) of documents constrained by being about a common topic, (b) showing how these substructures emerge, change, and disappear (and maybe re-appear) over time, and (c) giving users intuitive interfaces for interactively exploring the topic landscape and at the same time the underlying documents. The problem is related to, but extends known problems, in particular evolutionary theme pattern discovery, cf. [20, 24, 15] and the document update problem [18], as well as the detection of bursty events, cf. [11].

By using simple co-occurrence measures on elements that make up a story through the STORIES method, we created a tool that allows users to look at and actively explore story evolution from their individual perspectives. A case study on a well-publicised story over a long period of time showed the usefulness of the proposed method. An easily-usable, interactive GUI for tracking story evolution is a specific focus of this work. Graphs that consist of elements of a co-occurrence network are an easy and understandable way of presenting the development of a story.

We also presented an evaluation framework for approaches to the ETP3 problem and demonstrated the representation quality of the approach, using a real-life case study. This represents an advancement over the state of the art because so far, evaluation with respect to a "ground truth" is mostly lacking from temporal text mining (with TDT and the DUC update tasks, which however address different computational problems, notable exceptions). In the future, we want to extend this framework to also allow for a comprehensive cross-evaluation of different methods (such as "global" clustering or (P)LSA-based methods vs. "local" co-occurrence analysis) and interestingness measures for patterns (such as time relevance or other measures of burstiness). In addition, we will complement our IR/data-mining oriented evaluation by usability assessments.

Another area of improvement is the detection of events represented by bursty features as suggested by [11]. In order to discover more precise story elements, next versions of the method will also look more into further natural language processing methods, including the investigation of more complex terms and concepts (e.g., n-grams) and syntactical analysis including POS tagging. These variations will be investigated with respect to their usefulness for different kinds of corpora (news, blogs, scientific publications, ...). Also, the end-user tool will be developed further to provide more interactions with the corpora.

Week	p20 U	p20 M	p15U	p15 M	p10 U	p10 M	p5 U	p5 M	rp20 U	rp20 M	rp15 U	rp15 M	rp10 U	rp10 M	rp5U	rp5 M
17	0.55	0.4	0.47	0.47	0.5	0.40	0.8	0.40	0.8	1	0.8	1	0.6	1	0.6	0.4
18	0.40	0.25	0.33	0.27	0.40	0.30	0.40	0.60	0.8	0.4	0.6	0.4	0.6	0.4	0.40	0.4
19	0.55	0.2	0.47	0.2	0.50	0.3	0.2	0.4	1	1	0.67	0.67	0.67	0.67	0.67	0.67
20	0.2	0.3	0.2	0.33	0.1	0.3	0.2	0.2	0.33	1	0.33	1	0.33	0.67	0.33	0.67
21	0.15	0.25	0.20	0.27	0.3	0.4	0.20	0.6	0.4	0.6	0.4	0.6	0.4	0.6	0.2	0.6
22					0.1	0.30	0.2	0.40					0.67	0.67	0.67	0.33
23					0	0	0	0					0	0	0	0
24			0.27	0	0.3	0	0.6	0			1	0	1	0	1	0
26					0.3	0.1	0.4	0.2					1	1	1	1
27					0.25	0.1	0.2	0.2					1	1	1	1
29	0.05	0.10	0.07	0.13	0	0.20	0	0	1	1	1	1	0	1	0	0
30	0.2	0.25	0.13	0.27	0.2	0.3	0	0.2	1	1	1	1	1	1	0	0.5
31	0.29	0	0.2	0	0	0	0	0	1	0	1	0	0	0	0	0
32					0.29	0	0.2	0					1	0	1	0
34	0.35	0.35	0.4	0.47	0.4	0.6	0.6	0.8	1	1	1	1	0.6	1	0.6	0.6
35	0.3	0.60	0.33	0.53	0.3	0.70	0.2	0.60	0.75	1	0.75	1	0.75	1	0.13	0.75
36	0.67	0.67	0.47	0.47	0.40	0.40	0.20	0.20	0.83	0.83	0.67	0.67	0.5	0.5	0.17	0.17
37	0.15	0.15	0.07	0.07	0.10	0.10	0.20	0.20	1	1	1	1	1	1	1	1
Avg.	0.32	0.29	0.28	0.27	0.25	0.25	0.26	0.28	0.83	0.82	0.79	0.72	0.62	0.64	0.49	0.45
Std.dev.	0.19	0.19	0.15	0.18	0.16	0.21	0.23	0.25	0.24	0.32	0.24	0.38	0.36	0.40	0.40	0.36

Table 1. Precision and recall for $n = 5, 10, 15, 20$ edges for raters U, M over the weeks (eventless weeks are excluded). Empty cells in a row "n" denote a week with $< n$ edges.

Automatic language processing of the type presented here has a number of limitations. These concern both natural-language understanding and media reception. For example, methods that focus on words/terms, whether local or global, cannot detect negation well. Our method cannot detect possible multiple meanings of one term (homonyms), and a dictionary would be needed to conflate different terms with the same meaning (synonyms). Frequency-based interestingness measures like our time relevance generally single out dominant themes (or ways of reporting) and, by design, neglect outliers that may still be important. Also, the method at present has no notion of or differentiation between news sources of different quality. Further method and tool developments and evaluations will address these issues.

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proc. SIGIR 2001*, pp. 10–18. ACM, 2001.

[2] J. F. Allan. *Topic Detection and Tracking*. Springer, Berlin etc., 2002.

[3] M. Biryukov, R. Angheluta, and M.-F. Moens. Multidocument question answering text summarization using topic signatures. *Journal on Digital Information Management*, 3(1):27–33, 2005.

[4] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. *Information Visualization*, 7(1):34–48, 2008.

[5] C. Chen. *Mapping Scientific Frontiers*. Springer, London, 2003.

[6] C. Chen. Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*, 57(3):359–377, 2006.

[7] R. Choudhary, S. Mehta, A. Bagchi, and R. Balakrishnan. Towards characterization of actor evolution and interactions in news corpora. In *Proc. ECIR 2008*, LNCS 4956, pp. 422–429. Springer, 2008.

[8] C. Clifton, R. Cooley, and J. Rennie. Topcat: Data mining for topic identification in a text corpus. *IEEE Trans. Knowl. Data Eng.*, 16(8):949–964, 2004.

[9] S. Debnath, P. Mitra, N. Pal, and C. Giles. Automatic identification of informative sections of web pages. *IEEE Trans. Knowl. Data Eng.*, 17(9):1233–1246, 2005.

[10] R. Feldman, M. Fresko, J. Goldenberg, O. Netzer, and L. H. Ungar. Extracting product comparisons from discussion boards. In *Proc. ICDM 2007*, pp. 469–474. IEEE Computer Society, 2007.

[11] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *Proc. VLDB '05*, pp. 181–192. VLDB Endowment, 2005.

[12] D. Gruhl, R. V. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proc. SIGKK'05*, pp. 78–87. ACM, 2005.

[13] Q. He, K. Chang, E.-P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *Proc. 7th SIAM Int. Conf. Data Mining*. SIAM, 2007.

[14] W. Huang and P. Eades. How people read graphs. In *Proc. APVis '05*, pp. 51–58. Australian Computer Society, Inc.

[15] F. A. L. Janssens, W. Glänzel, and B. D. Moor. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proc. SIGKDD'07*, pp. 360–369. ACM, 2007.

[16] J. M. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[17] L. Leydesdorff and T. Schank. Dynamic animations of journal maps: indicators of structural change and interdisciplinary developments. *JASIST*, 59(11): 1810–1818, 2008.

[18] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Proc. WS Text Summarization Branches Out (WAS 2004)*, 2004.

[19] C.-Y. Lin and E. Hovy. Automated multi-document summarization in neats. In *Proc. Second Int. Conf. Human Language Technology Research*, pp. 59–62. Morgan Kaufmann, San Francisco, CA, 2002.

[20] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proc. SIGKK'05*, pp. 198–207. ACM, 2005.

[21] R. Navigli and P. Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguistics*, 30(2):151–179, 2004.

[22] S. Nijssen and J. N. Kok. A quickstart in frequent structure mining can make a difference. In *Proc. SIGKDD'03*, pp. 647–652. ACM, 2004.

[23] M. Oka, H. Abe, and K. Kato. Extracting topics from weblogs through frequency segments. In *Proc. WWW2006 WS Weblogging Ecosystem*, 2006. `http://www.blogpulse.com/www2006-workshop/papers/wwe2006-oka.pdf`.

[24] R. Schult and M. Spiliopoulou. Discovering emerging topics in unlabelled text collections. In *Proc. ADBIS 2006*, LNCS 4152, pp. 353–366. Springer, 2006.

[25] D. A. Smith. Detecting and browsing events in unstructured text. In *Proc. SIGIR 2002*, pp. 73–80. VLDB Endowment, 2002.

[26] M. Thelwall. Blogs during the london attacks: Top information sources and topics. In *Proc. WWW2006 WS Weblogging Ecosystem*, 2006. `http://www.blogpulse.com/www2006-workshop/papers/blogs-during-london-attacks.pdf`.

[27] J. Véronis. Sense tagging: Does it make sense? In *Corpus Linguistics 2001 Conf.*, 2001. `http://citeseer.ist.psu.edu/veronis01sense.html`.

[28] Wikipedia. Disappearance of Madeleine McCann, 2008. `http://en.wikipedia.org/w/index.php?title=Disappearance_of_Madeleine_McCann&oldid=215814790`.

[29] Wikipedia. Disappearance of Madeleine McCann, 2008. `http://en.wikipedia.org/w/index.php?title=Disappearance_of_Madeleine_McCann&oldid=243566210`.

[30] P. C. Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas. Visualizing sequential patterns for text mining. In *Proc. INFOVIS*, pp. 105–111, 2000.

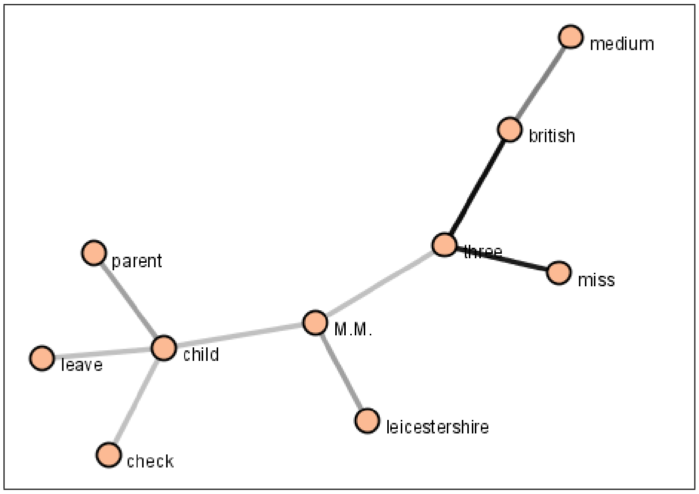
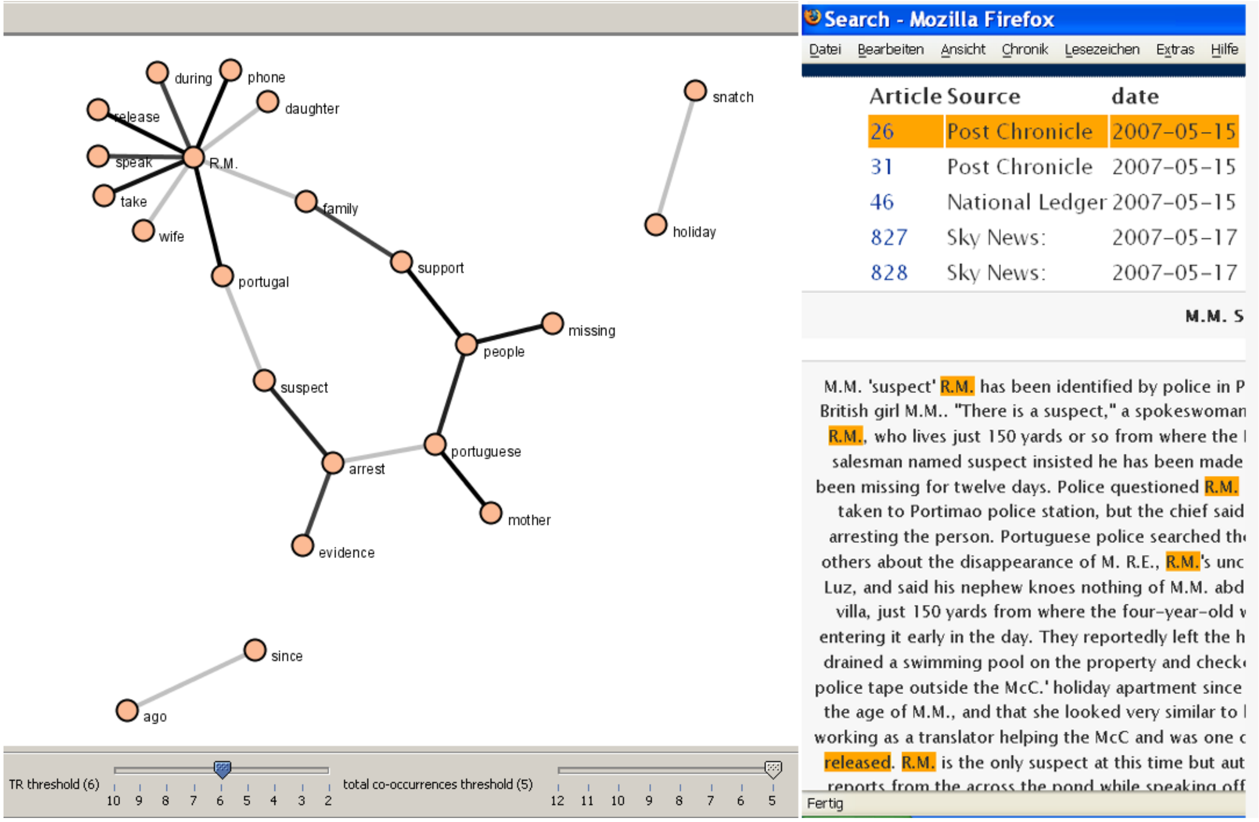


Figure 2. Week 17 ($TR \geq 3$): Description of an event: a missing child.



Article	Source	date
26	Post Chronicle	2007-05-15
31	Post Chronicle	2007-05-15
46	National Ledger	2007-05-15
827	Sky News:	2007-05-17
828	Sky News:	2007-05-17

M.M. S

M.M. 'suspect' R.M. has been identified by police in P
British girl M.M.. "There is a suspect," a spokeswoman
R.M., who lives just 150 yards or so from where the
salesman named suspect insisted he has been made
been missing for twelve days. Police questioned R.M.
taken to Portimao police station, but the chief said
arresting the person. Portuguese police searched th
others about the disappearance of M. R.E., R.M.'s unc
Luz, and said his nephew knoes nothing of M.M. abd
villa, just 150 yards from where the four-year-old v
entering it early in the day. They reportedly left the h
drained a swimming pool on the property and check
police tape outside the McC.' holiday apartment since
the age of M.M., and that she looked very similar to
working as a translator helping the McC and was one c
released. R.M. is the only suspect at this time but aut
reports from the across the pond while speaking off

Figure 3. Week 18 ($TR \geq 6$): R.M. emerges as a central figure; story stage shown in the GUI. The browser windows shows documents containing the co-occurrence of the edge "release"–"R.M.".

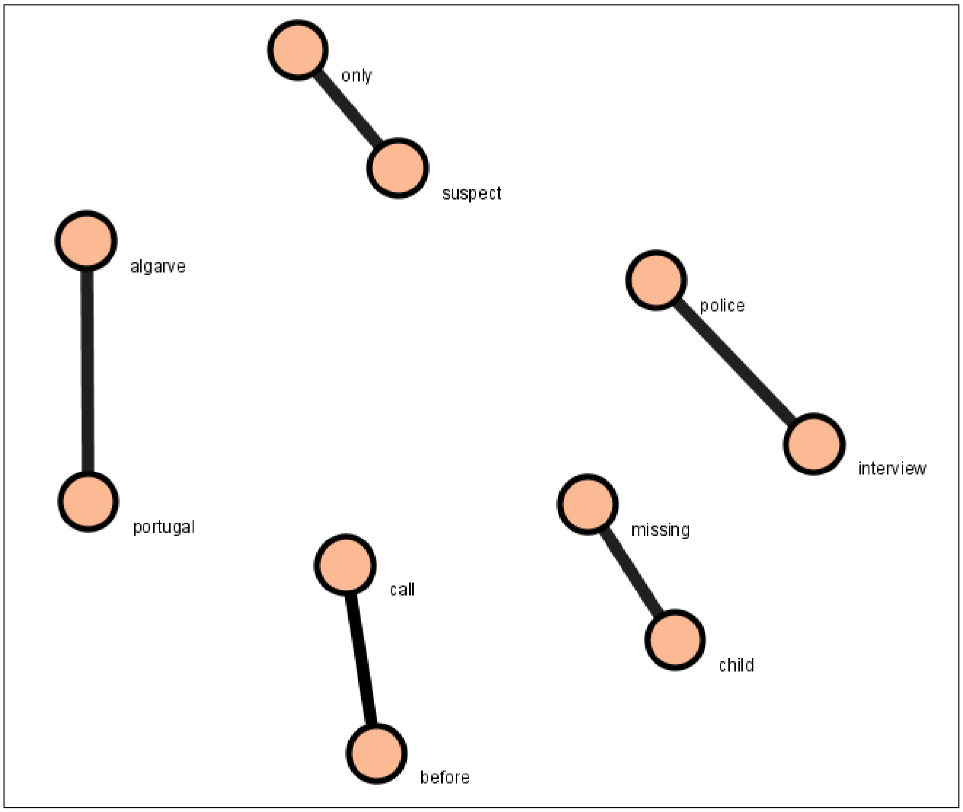


Figure 4. Week 26. ($TR \geq 3$): An eventless time.

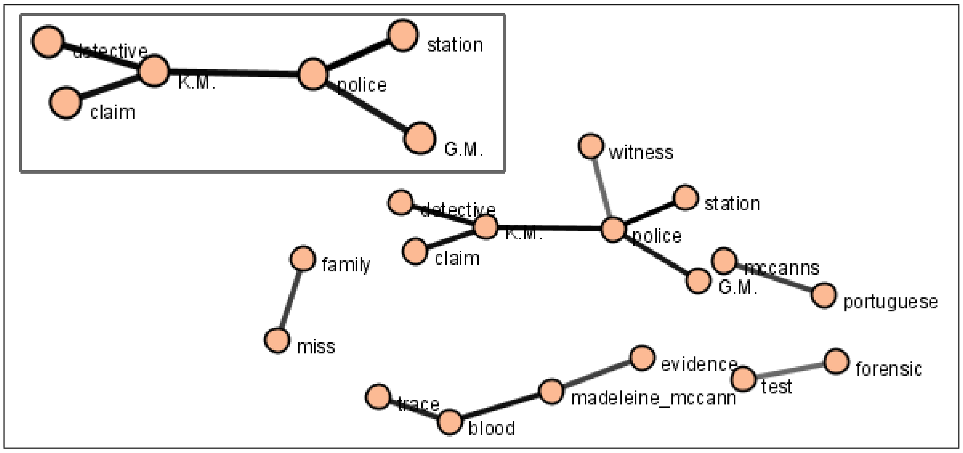


Figure 5. Week 34. Event uncovering. Top ($TR \geq 10$): The police are questioning K.M. ... Bottom ($TR \geq 5$): ... in relation with the blood found in the car.