Identifying, measuring and visualising the evolution of a story: A Web mining approach

Bettina Berendt[*]   Ilija Subašić[†]

October 10, 2008

Abstract

Rich information spaces (like the space of scientific publications or the Web) are full of a higher-order type of "stories": sets of statements that evolve over time, manifested as, for example, sequences of scientific papers on a topic, collections of newspaper articles reporting events relating to an evolving crime investigation, or sets of news articles and blog posts accompanying the development of a political election campaign. 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. A real-life case study is used to illustrate 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[1] 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[2] or BlogPulse's Trend Search[3] 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 and highly interactive interfaces for exploring the topic landscape and at the same time the underlying documents. In an extension of (Mei and Zhai 2005), we call the resulting problem *evolutionary theme patterns discovery, summary and exploration* (ETP3).

In the past years, a number of powerful methods for solving subsets of these three requirements have been proposed. However, systems that address all three challenges are still lacking. The first contribution of the current paper is a description of a system that does this.

The second 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 par-

[*]K.U. Leuven, Dept. of Computer Science, Belgium, firstname dot lastname at cs dot kuleuven dot be

[†]K.U. Leuven, Dept. of Computer Science, Belgium, firstname dot lastname at cs dot kuleuven dot be

[1]http://news.google.com

[2]http://www.google.com/trends

[3]http://www.blogpulse.com/trend

ticular, 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 the user in finding and exploring their own interpretations.

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 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 extracting content from text, and information visualisation.

Temporal text mining. Mei and Zhai (2005) 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 is used to model documents as expressing (potentially several) themes (corresponding to sub-topics). These word clusters are tracked over time with Kullback-Leibler divergence measuring similarity, and the lifecycle of themes as well as cross-theme transformations are modelled as a Hidden Markov Model. The use of clustering models for finding emergent sub-topics and tracking them over time is also the topic of (Schult and Spiliopoulou 2006; Janssens et al. 2007). 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 (Allan 2002). 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.

These clustering 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* (Kleinberg 2003), 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 (Fung et al. 2005; Gruhl et al. 2005; He et al. 2007). Fung, Yu, Yu, and Lu (2005) 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. Biryukov, Angheluta, and Moens (2005) show that *topic signatures* (Lin and Hovy 2002) provide a simple and effective way to summarise multiple documents. They use topic signatures to associate person names with other words, restricting the word class of the latter to maximise informativeness, and they use χ^2 and likelihood ratios as interestingness measures. Smith (2002) used co-occurrences to find historical associations between places and times in a digital library. He analysed how various interestingness measures rank these associations (raw counts, χ^2, log-likelihood, and mutual information) 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 inter-

pretation also supported by the findings of Feldman, Fresko, Goldenberg, Netzer, and Ungar (2007). These authors found co-occurrence lift to be an adequate interestingness measure to analyze perceptions of (car) brands and markets in user forums.

Choudhary, Mehta, Bagchi, and Balakrishnan (2008) 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 analyzing co-occurrences between typed elements (names, places, ...). We take a more general approach and identify "story lines" between arbitrary words or concepts.

Allan, Gupta, and Khandelwal (2001) 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. They aimed at finding sentences with high coverage and high novelty. Coverage is similar to the frequency/interestingness of conceptual patterns behind the sentences. In contrast to Allan et al., 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. Out of the vast literature on information visualisation, two approaches seem most relevant for our task (for an overview of many more, especially in the domain of literature-network visualisation, see (Chen 2003)). Smith (2002) 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. Wong, Cowley, Foote, Jurrus, and Thomas (2000) 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., Chen (2003), Chen (2006), Janssens, Glänzel, and Moor (2007). In contrast to that, we use a layout strategy that is more amenable to highlighting emerging and disappearing topics, and dynamic layout between successive time periods (morphing), the latter similar to (Leydesdorff et al. 2008; Leydesdorff and Schank 2008).

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 by time 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:[4]

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

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

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

This measure is based on the *domain relevance* metric (Navigli and Velardi 2004) 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": A high time relevance in a specific period signals that the term (or, in our case, the association) is bursty in that period.

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)*:

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

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

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 $TR > \theta_2 = 3$ 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

We applied the method to news articles downloaded from different sources on the Web, as indexed by Google News. In this section, we describe the data cleaning and further preprocessing 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

---

[4]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.

so far, such as (Debnath et al. 2005), 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. In order to not conflate data cleaning issues with content extraction issues, we therefore extracted the content into ASCII by manual copy-and-paste.

Text pre-processing. The documents were first tokenized; subsequently, several further preprocessing options were investigated. It turned out that stemming produced results that are too unintuitive for non-expert users, since stems are often not natural-language words. In contrast, lemmatization, performed by the TreeTagger[5] produced well-readable results.[6] The lemmas or, when the TreeTagger was not able to resolve a token, the original words were then used as terms. Term co-occurrences were measured in windows of $w = 5$ terms. Prior analyses had shown that stopword removal was not useful in this application, since it reduced the short texts too much, such that nearly all content-bearing words co-occurred. The window size for time periods l was set to 3.

All parameters can easily be changed, 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.

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 colour scheme accentuates time relevance differences, going from blue (high time relevance) via red (medium) to yellow (low). All programs can be executed on a local computer; after an initial download of documents from the Internet. 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

As a case study, we used a real-life story with a comparatively clear and well-known (and well-publicised) 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

---

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

[6] This is not perfect, because lemmatization does not conflate different word types with the same stem and in some cases may even *increase* the number of different tokens. For example, "missing" may be lemmatized to "miss" when it is a verb form, but remain unchanged when it is used as an attribute. In preliminary user tests, it was pointed out that a conflation of these two nodes would increase readability of the graphs.

[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 on the world might have would be reported, including a Web site that "soon became the [UK]'s number one community website" (http://www.telegraph.co.uk/news/main.jhtml ?xml=/news/2007/05/22/nmaddy122.xml). On the first anniversary of Madeleine's disappearance (which lay after the bulk of the work reported here was done), 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., (Thelwall 2006; Oka et al. 2006)) or the 2004 Tsunami (Mei and Zhai 2005).

[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.
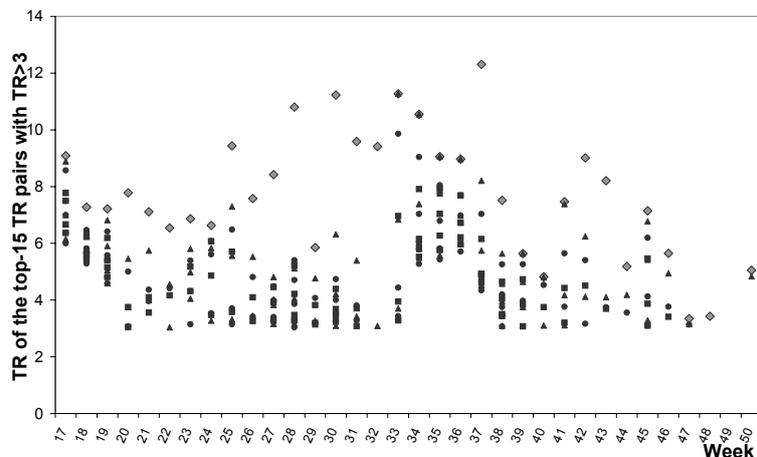
Figure 1: Burstiness profile of the three-week periods 17–50 (top-*TR* values highlighted)

misleading incidents like the arrest of two people unrelated to the case). All three suspects were cleared later; and the case was closed in July 2008, see (Wikipedia 2008).

The corpus. We used articles from the Google News archive[10] 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. This resulted in a corpus of 308 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), and there is no quality filter on the news sources in the Google News archive after some months.

This set was extended by the set of all retrievable, English-language news articles referenced in the Wikipedia article (Wikipedia 2008), from

---

[10]http://news.google.com/archivesearch

the investigated time period. This provided another 135 articles (only one article was found in both sets). 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 total number of words in the combined corpus was 183,834, resulting in an average of 415 words per article. The corpus contained 9,875 (7,602) unique words (lemmas).

Results Basic statistics are shown in Fig. 1. The figure shows the time relevance value of the top 15 co-occurrences in each three-week time window. Pairs with $TR \leq 3$ were filtered out. They show that the time periods differed strongly in the amount of time-specific reported information. They range from comparatively eventless weeks like 22 to weeks with highly specific information like 37. (The co-occurrence

with the highest overall *TR* value was actually a point event, the hiring of a press spokesman for the McCanns. Such co-occurrences are obviously much more frequent than in the preceding time periods (where these point events could not have been foreseen), and they are also more frequent than in the subsequent time periods (where such a point event usually stops being "news").

Figures 2–5 show selected individual story stages.[11] In particular, they show instances of (a) how story elements can change although story basics remain the same (cf. "M" in Figs. 2 vs. 4, or "suspect" in Figs. 2 vs. 4 vs. 5)[12], (b) how the interface thus helps the user to "track" developments between story stages, (c) how the interface helps the user to "uncover" a story stage to get more detail (cf. the top vs. bottom parts of Fig. 5), and last but not least (d) what a comparatively eventless story stage looks like (cf. Fig. 3).

6   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 and highly interactive interfaces for 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. (Mei

---

[11]Visualisations of the corpus with θ_2 adjusted for maximum visibility and $\theta_1 = 5$ throughout.

[12]Early on in the case, the police published a description of a suspect. This was independent of the questioning and later official declaration as suspect of R.M., which corresponds to the linkage of "suspect" in Fig. 2. R.M. quickly became the (only) official suspect for a long time, including renewed police interviews in week 27; the parents became official suspects in week 36.

Another couple had been implicated and even arrested in week 26, but they turned out to be con artists who had nothing to do with the case.

and Zhai 2005; Schult and Spiliopoulou 2006; Janssens et al. 2007) and the detection of bursty events, cf. (Fung et al. 2005).

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. Many other works in similar areas pay little attention to the usability of the proposed method and the way results could be presented to the users. Graphs that consist of elements of a co-occurrence network are an easy and understandable way of presenting the development of the story.

The most important area that we will investigate in future work is the evaluation of the approach. So far, evaluation with respect to a "ground truth" is mostly lacking from temporal text mining (with TDT, which however addresses different computational problems, a notable exception). An evaluation framework and a first evaluation study of the STORIES method are presented in (Subašić and Berendt 2008).

Another area of improvement is the detection of events represented by bursty features as suggested by Fung, Yu, Yu, and Lu (2005). 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. Named entities as more complex concepts are included in the current version of the tool (Subašić and Berendt 2008). 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 underlying text corpora.

The modular architecture of our tool will 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 mea-

sures for patterns (such as time relevance or other measures of burstiness).

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, for example the time relevance measure, 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. While further method and tool developments and evaluations will allow us to improve on some of these issues, we believe it is crucial to also find effective ways of communicating inherent limitations to users, in order to encourage a critical use of tools. Developing such meta-level communication strategies is another key goal of our future work.

In sum, looking at (any) information overload problem demands a user perspective, a fact that sometimes tends to be neglected in text mining research. In our tool-oriented approach, early and continued testing with users and iterative refinements based on these results become possible.

References

Allan, J., R. Gupta, and V. Khandelwal (2001). Temporal summaries of news topics. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9–13, 2001, New Orleans, Louisiana, USA*, pp. 10–18. ACM.

Allan, J. F. (2002). *Topic Detection and Tracking*. Berlin etc.: Springer.

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

Chen, C. (2003). *Mapping Scientific Frontiers*. London: Springer.

Chen, C. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology 57*(3), 359–377.

Choudhary, R., S. Mehta, A. Bagchi, and R. Balakrishnan (2008). Towards characterization of actor evolution and interactions in news corpora. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30–April 3, 2008. Proceedings*, Volume 4956 of *Lecture Notes in Computer Science*, pp. 422–429. Springer.

Debnath, S., P. Mitra, N. Pal, and C. Giles (2005). Automatic identification of informative sections of web pages. *IEEE Transactions on Knowledge and Data Engineering 17*(9), 1233–1246.

Feldman, R., M. Fresko, J. Goldenberg, O. Netzer, and L. H. Ungar (2007). Extracting product comparisons from discussion boards. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28–31, 2007, Omaha, Nebraska, USA*, pp. 469–474. IEEE Computer Society.

Fung, G. P. C., J. X. Yu, P. S. Yu, and H. Lu (2005). Parameter free bursty events detection in text streams. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pp. 181–192. VLDB Endowment.

Grossman, R., R. J. Bayardo, and K. P. Bennett (Eds.) (2005). *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21–24, 2005*. ACM.

Gruhl, D., R. V. Guha, R. Kumar, J. Novak, and A. Tomkins (2005). The predictive power of online chatter. See Grossman, Bayardo, and Bennett (2005), pp. 78–87.

He, Q., K. Chang, E.-P. Lim, and J. Zhang (2007). Bursty feature representation for clustering text streams. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26–28, 2007, Minneapolis, Minnesota, USA*. SIAM.

Janssens, F. A. L., W. Glänzel, and B. D. Moor (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12–15, 2007*, pp. 360–369. ACM.

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

Leydesdorff, L. and T. Schank (2008). Dynamic animations of journal maps: Indicators of structural change and interdisciplinary developments. *Journal of the American Society for Information Science and Technology 59*(11), 1810–1818.

Leydesdorff, L., T. Schank, A. Scharnhorst, and W. D. Nooy (2008). Animating the development of social networks over time using a dynamic extension of multidimensional scaling. Keynote address at the *Fourth International Conference on Webometrics, Informetrics, and Scientometrics & Ninth COLLNET Meeting*. 28 July – 1 August 2008, Berlin.

Lin, C.-Y. and E. Hovy (2002). Automated multi-document summarization in neats. In *Proceedings of the second international conference on Human Language Technology Research*, San Francisco, CA, USA, pp. 59–62. Morgan Kaufmann Publishers Inc.

Mei, Q. and C. Zhai (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. See Grossman, Bayardo, and Bennett (2005), pp. 198–207.

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

Oka, M., H. Abe, and K. Kato (2006). Extracting topics from weblogs through frequency segments. In *Proc. of WWW2006 3rd Annual Workshop on the Weblogging Ecosystem*. `http://www.blogpulse.com/www2006-workshop/papers/wwe2006-oka.pdf`.

Schult, R. and M. Spiliopoulou (2006). Discovering emerging topics in unlabelled text collections. In *Advances in Databases and Information Systems, 10th East European Conference, ADBIS 2006, Thessaloniki, Greece, September 3–7, 2006, Proceedings*, Volume 4152 of *Lecture Notes in Computer Science*, pp. 353–366. Springer.

Smith, D. A. (2002). Detecting and browsing events in unstructured text. In *Proceedings of the 25th Annual ACM SIGIR Conference*, pp. 73–80. VLDB Endowment.

Subašić, I. and B. Berendt (2008). Web mining for understanding stories through graph visualisation. In *Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM 2008)*, New York. IEEE Press.

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

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

Wong, P. C., W. Cowley, H. Foote, E. Jurrus, and J. Thomas (2000). Visualizing sequential patterns for text mining. In *INFOVIS*, pp. 105–111.
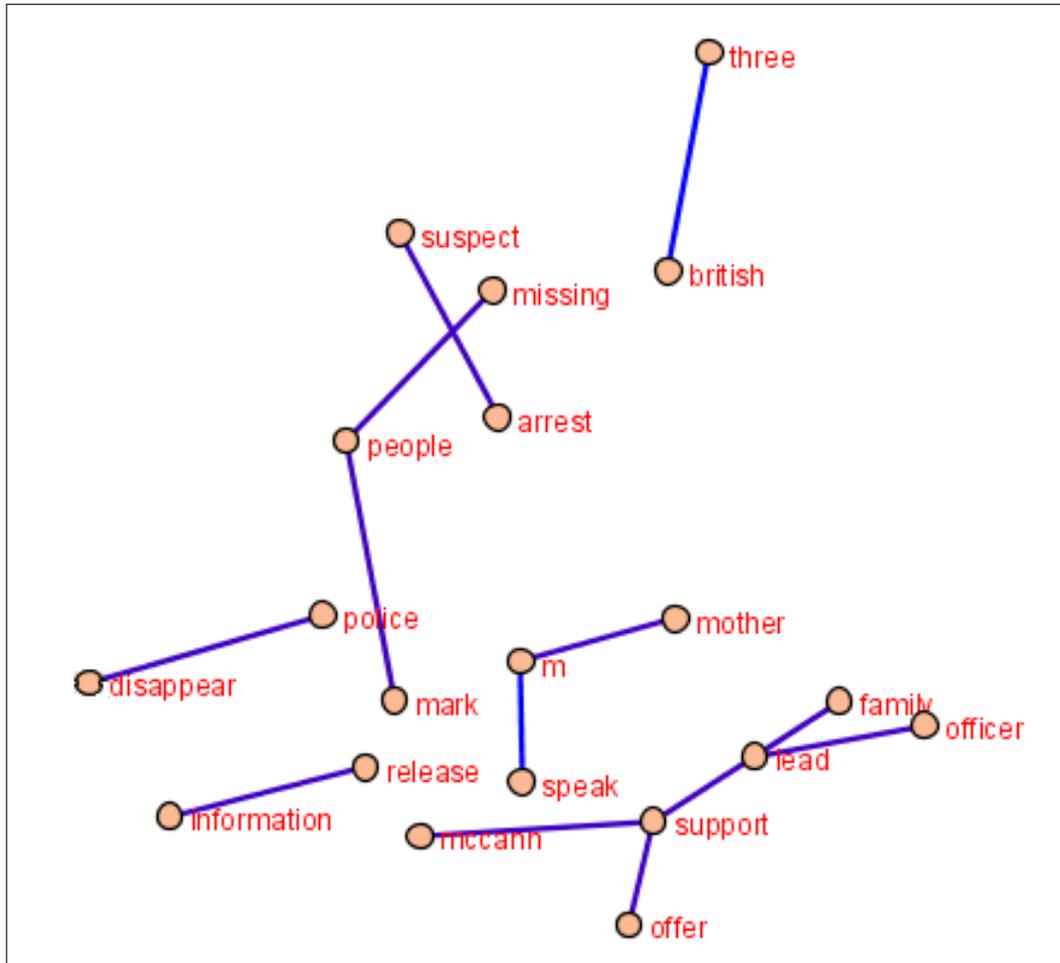
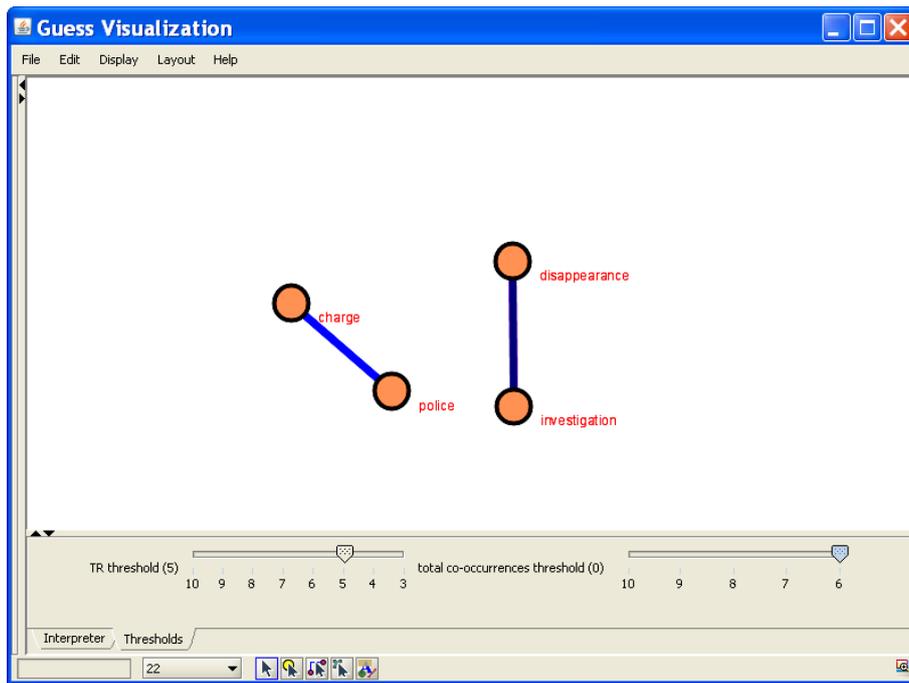Figure 2: Week 20 ($TR = 6$): The house of R.M.'s mother was searched, and, being questioned, he spoke to the police.

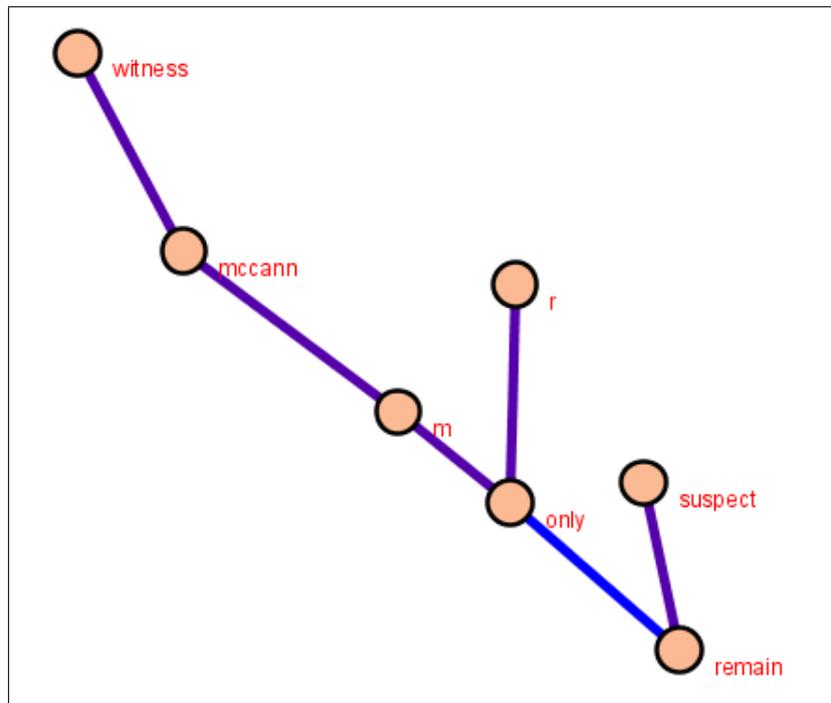Figure 3: Week 22 ($TR = 5$): An eventless time; story stage shown in the GUI.

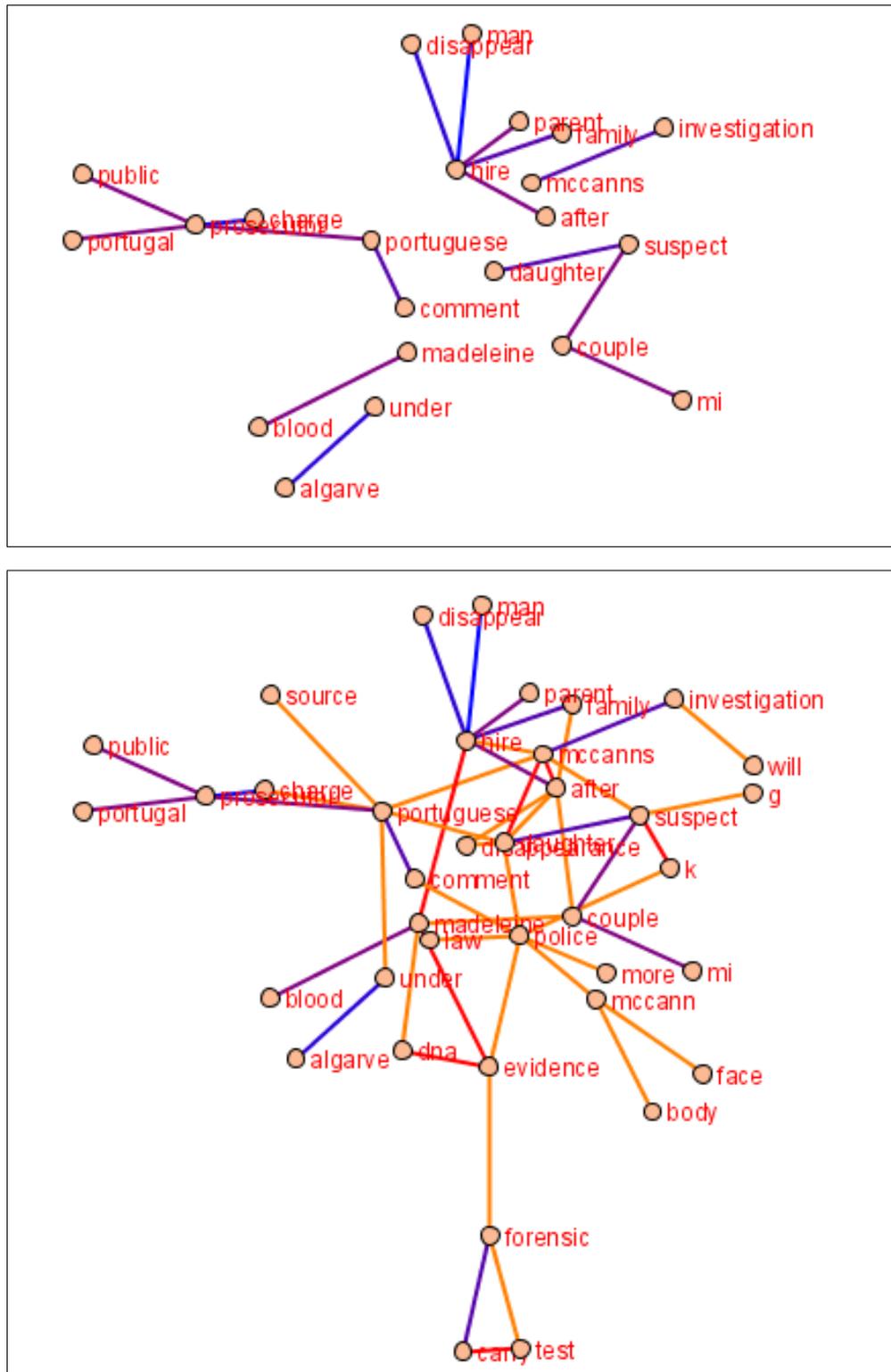Figure 4: Week 28 ($TR = 5$): R.M. (who has been questioned again) is still the only suspect.

Figure 5: Week 34. Top ($TR = 6$): Madeleine's blood found in the car, and a couple are the suspects. Bottom ($TR = 4$): The couple are Madeleine's parents, K. and G. (Mi. is the couple's spokesman.)