

# Measuring graph topology for interactive temporal event detection

Bettina Berendt, Ilija Subašić

**The Web and other text collections are full of “stories”: sets of statements that evolve over time, made in fast-growing streams of documents. Even if one reads a specific source every day and/or subscribes to a selection of feeds, one may easily lose track; in addition, it is difficult to reconstruct a story already in the past. In this paper, we present the STORIES methods and tool for (a) learning an abstracted story representation from a collection of time-indexed documents; (b) visualizing it in a way that encourages users to interact and explore in order to discover temporal “story stages” depending on their interests; (c) supporting the search for documents and facts that pertain to the user-constructed story stages; and (d) navigating in document space along multiple meaningful dimensions of document similarity and relatedness. This combination provides users with more control of their customized story understanding, semantically in story space as well as between the underlying documents. In an evaluation, we investigated whether it is possible to use topological properties of the temporal story graphs for pointing users to more promising parts of the story space. The results indicated that global properties of the graphs are useful for this purpose, while properties local to individual nodes are less meaningful.**

## 1 Introduction

The Web, the Social Web, and other corpora like scientific publications are challenging for information-seekers because their content gets updated continuously, and it is hard to keep abreast. Even if one reads a specific source every day and/or subscribes to a selection of feeds, one may not “see the forest for the trees”. There is clearly a need for summarizing services, and these summaries should ideally provide a concise abstracted representation of information from all the pertinent parts of a source or source set. In addition, metadata should be utilised for such services. Last but not least, users will be able to profit most from summarizing services that provide convenient interfaces to both the abstracted summary and the underlying documents, and that allow for and encourage a flexible, (inter)active exploration of the space of the abstracted “topics” or “stories” on the one hand, and the space of the documents on the other hand.

In this paper, we present the STORIES approach that instantiates these ideas for a particular setting of media with a particular need for such interfaces: We focus on textual, time-indexed documents such as news and blogs. Both blogs and news constitute fast-growing corpora that report on recurring topics or unfolding stories. 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. Summarisation like that provided by Google Trends<sup>1</sup> or BlogPulse’s Trend Search<sup>2</sup> show surges in publication and query activity in certain time periods, but

these tools require that one knows which sub-topic to look for. Blogs and news are particularly interesting when viewed in relation to one another, raising questions such as who reports on stories first [19], who follows whom [7], and in general how the two media types mirror and relate to one another [10].

Our solution consists of four parts: (a) From a collection of time-indexed documents, a “story” is *learned*. This component employs text mining and summarisation. (b) This story is represented in a novel graphical way that encourages users to interact and explore in order to discover temporal “story stages” at different levels of granularity, depending on their interest. These *interactions in story space* results in the users discovering and constructing the story parts they wish to focus on. (c) The constructed story parts directly support *search*, serving as new indexes into the document set. (d) In the document set, a focussed local search for semantically related documents by navigation between documents is enabled. This *interaction in document space* relies on various document features extracted directly from the metadata and indirectly via text analysis. Thus, navigation directly supports the user in relating blogs as one source type with news as another source type – with respect to different questions concerning their (dis)similarity.

The contributions of this paper are threefold. First, it presents an integrated version of the recursive combination of learning, search, and interaction in both story space and document space. This integrated system builds on our previous work concerning learning and interaction in story space [24] and on the local analysis of document spaces that is part of the system described in [3]. Second, these systems are extended by enhanced search functionality. Third, we present an evaluation study that investigates whether global and/or local topological properties of the story graphs can be used to point the user to more “eventful” parts of the reported content.

The paper begins with an overview of related work. The subsequent Method and Tool sections are structured

<sup>1</sup> <http://www.google.com/trends>

<sup>2</sup> <http://www.blogpulse.com/trend>