Rich Interfaces for Browsing News in Blog Posts

Earl J. Wagner
ewagner
@u.northwestern.edu

Jiahui Liu
j-liu2
@northwestern.edu

Lawrence Birnbaum
birnbaum
@cs.northwestern.edu

Northwestern University, Evanston IL USA

ABSTRACT
Semantic models of news can enable richer interfaces for end-users to learn the context of news events referenced in blog posts. We present Brussell, a system that uses content-specific models of news event situations to perform anticipatory information retrieval, organize extraction results and present a novel, structured interface for navigating among the events of a news situation.

INTRODUCTION
A blogger commenting upon a news event in a blog post is likely to link to a relevant news article, often in the first sentence of the post. By providing a link, the author can assume that the reader has read the linked article and is familiar with the news event it covers. In fact, often the "value-add" for blog posts is not in repeating the information of the article, nor reporting new information, but in sharing the blogger's opinion or analysis. Bloggers use the news article as the starting-point for further discussion.

But what happens when the blog reader is unfamiliar with the news event described in the article? Often, a blogger will argue that a current event is caused by, or part of a trend involving, past events. Alternately, if the event that the blog post refers to is not recent, or if the post itself is old, the reader may be interested in finding out its outcome. We argue that software can help the user by explaining the context of news events through content-specific structured presentations.

BACKGROUND TO NEWS EVENTS IN BLOG POSTS
The issue of explaining the context of news highlights one important difference between blog posts and news articles. The blog format enables authors to write in a conversational style and presume familiarity with earlier commentary. Posts generally don't provide as much context as news articles and are often written with the expectation that readers will have been following the blog and are familiar with the events they discuss. Journalists writing news articles, on the other hand, provide "background" to current news events by explaining the events and factors that led to their occurrence.

Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW2009), IUI2009, Feb 8 2009, Sanibel Island, Florida, USA.

Often background events are related to the event covered by a news article by being part of the same overall situation - perhaps an earlier event in the situation, where by *situation* we mean a limited sequence of causally-related events, such as all of the newsworthy actions in a lawsuit. For example, an account of a lawsuit being settled by the defendant may refer back to the original filing of the lawsuit and both events are part of a particular lawsuit situation. Alternately, a news article may reference other similar or related situations. A similar lawsuit may be taking place simultaneously in another locale. Related lawsuits include a suit acting as a case precedent, or other suits involving some of the same participants, such as other suits against the defendant.

All of these relationships are part of the situational context that the user draws upon in making sense of the news event a news article describes. This context gives rise to specific questions, such as:

- What happened in this situation?
- How did this situation start? How did it end?
- What happened in the other situations referenced in this article?
- What other similar and related situations have these participants been involved in?

Neither conventional software and web sites for blogs and news, nor research systems for "semantic-blogging" provide content-specific support for answering these questions, however.

Without an in-page link, the user must find related articles manually in order to answer questions like these. She must identify relevant terms such as entity names and situation keywords. Then she cut-and-pastes them into a news search engine. Finally she sorts through lists of results to find relevant articles. These steps make for an inconvenient process familiar to anyone who reads news on the web. Even news timelines provided by advanced search engines are unable to provide content-specific overviews of a situation in accordance with the user's expectations of how it begins and continues.

We present ***Brussell***, a system that performs anticipatory information retrieval and model-based information extraction to support the user in exploring the situational

context of the news. Brussell retrieves news articles and extracts information to create models of news situations. When a user selects a situation, it presents a storyline with the major milestone events. Clicking on the event label loads an article that either immediately covers the event or is the earliest mention of the event. Evidence that an event took place, for its date and location, or for important attributes of participating entities can also be viewed in the form of collected textual snippets and links to source pages. All of these features work together to provide the user a "big picture" view of the news situation and answer high-level questions about it.

EXAMPLE

Consider the case of a user reading about the kidnapping of a BBC journalist on a blog post. Although the user is vaguely aware of this incident, he would like to find out more. With standard search technology, he would enter terms into a search engine and peruse the results in order to develop an overall sense of how the kidnapping situation transpired. Through Brussell, he can interact directly with the text referencing the situation. Moving the mouse over the text causes it to be highlighted (see Figure 1).

Right-clicking on this highlighted text opens a context menu presenting options for viewing the history of the situation and finding out more about its participants (see Figure 2). The user wants to see a summary of what happened, so he selects the first option, which updates the toolbar to show a storyline for the kidnapping with its major events and their dates (see Figure 3). With this high-level view of the kidnapping situation, he is able to see the how the situation unfolded through the occurrence of its milestone events. Alternately, selecting one of the menu items to learn more about a participant would load a similar timeline view presenting all of the situations the participant has been involved in.

Next, he wants to know exactly how the situation ended, so he selects the "release" event button that loads the most relevant page describing the event in detail (see Figure 4). By reading this page the user can learn more about the circumstances under which the journalist was released.

ARCHITECTURE

Brussell consists of a Firefox browser plugin and server software, which may both run on the same computer. When the user visits a web page the browser plugin sends the current page title and URL to the server, which responds with the (possibly cached) page situation references. A user can view situation references in a blog post, as in the example, or in other web pages such as news articles.

The back-end system uses manually created situation model types (scripts) and currently supports *kidnappings*, *legal trials* and *corporate acquisitions* each of which has multiple possible outcomes and 8-12 possible events. The system runs daily to retrieve news articles from several news web sites via RSS feeds and store them in a Lucene index [5]. It

Figure 1. Viewing a reference to a news situation within a blog post.

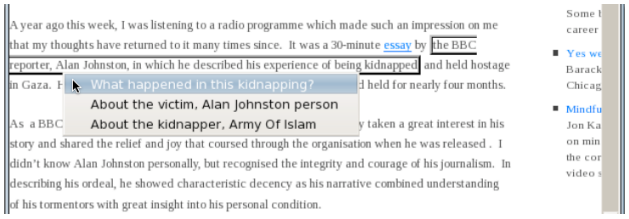

Figure 2. Asking about the situation.

Figure 3. Viewing a storyline for the selected situation

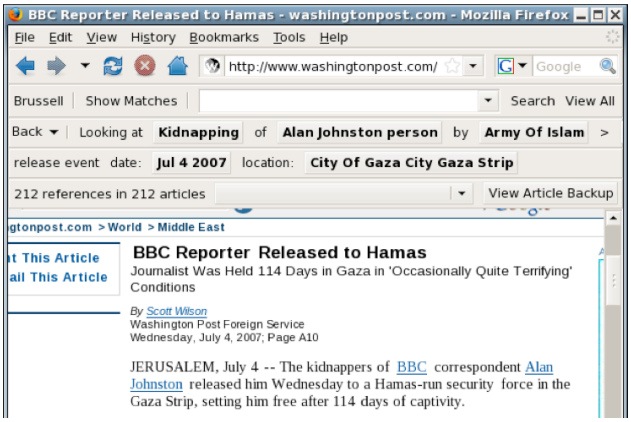

Figure 4. Viewing an article for the selected situation event.

then queries the database for new articles with keywords associated with the situation types it supports and reads through the returned articles to create and extend situation models instances of these types. These instances each cover a specific situation such as the kidnapping of the journalist above and include information from a few articles, up to several hundred if they are well-publicized.

Brussell uses GATE [3], a standard open-source information extraction system to extract situation information including event references, dates and locations, and entity information such as person names and occupations or organization names and nationalities. Extracting this information allows references such as "the British journalist abducted last year" to be resolved to a particular kidnapping situation instance. In fact, the same mechanism used for extracting information is used to identify situation references in page text, and in analyzing news articles the system caches the textual references for all of articles it processes.

RELATED WORK

Several semantic-blogging systems and been developed to support blog authoring for the Semantic Web. An extension to the RDF-based personal information management system Haystack supports right-clicking on any web page to create a blog post with a link to the page [6]. Like Brussell, it also improves the experience for reading blogs through structured views, including a graph of the comments made about a blog post. Unlike Brussell, it doesn't offer content-specific views of the news referenced in blog posts, however. The SWAD-E project represents the bibliographic entries referenced within blog posts and provides content-specific affordances for interacting with this information [2]. Another approach is taken by the Magpie Semantic Web browsing system, which allows rich interaction with references to entities within web pages [4]. Neither system models news events and situations and how these situations unfold, however.

Some systems extract formal knowledge from news to populate the Semantic Web such as SemNews, which processes news retrieved via RSS feeds [1]. Unlike Brussell, however, its emphasis is generating representations in the form of RDF triples rather than presenting structured views to the user.

Other work has focused on finding news articles relevant to blog posts by identifying and searching for important words in the post [8]. Though this approach doesn't provide a high-level storyline view of a news situation that organizes relevant news articles, it benefits from being domain-independent.

Further discussion of Brussell appears in [7].

REFERENCES

1. Akshay J., Finin, T., and Nirenburg, S. SemNews: A Semantic News Framework. 2006. *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06),* February 2006.
2. Cayzer, S. Semantic Blogging: Spreading the Semantic Web Meme. *Proceedings of XML Europe 2004*, Amsterdam, Netherlands, April 2004.
3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, PA, USA, July 2002.
4. Dzbor, M., Domingue, J., and Motta, E. Magpie: Towards a Semantic Web Browser. *ISWC 2003, Second International Semantic Web Conference*, Sanibel Island, FL, USA, October 2003.
5. Lucene. http://lucene.apache.org/java/docs/
6. Karger, D. R., Quan, D. What would it mean to blog on the semantic web? Journal of Web Semantics. 3(2-3): 147-157. 2005.
7. Wagner, E. J., Liu, J., Birnbaum, L., Forbus, K. D. Rich Interfaces for Reading News on the Web. *Proceedings of the 2009 International Conference on Intelligent User Interfaces*, Sanibel Island, FL, USA, February 2009.
8. Ikeda, D., Fujiki, T., Okumura, M. Automatically Linking News Articles to Blog Entries. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, Palo Alto, CA, USA, 2006.