NewsInEssence: A System For Domain-Independent, Real-Time News Clustering and Multi-Document Summarization

Dragomir R. Radev[⋆†], Sasha Blair-Goldensohn[⋆], Zhu Zhang[⋆], Revathi Sundara Raghavan[†]
[⋆]School of Information
[†]Department of EECS
University of Michigan
Ann Arbor, MI 48109
{radev,sashabg,zhuzhang,rsundara}@umich.edu

1. INTRODUCTION

NEWSINESSENCE is a system for finding, visualizing and summarizing a topic-based cluster of news stories. In the generic scenario for NEWSINESSENCE, a user selects a single news story from a news Web site. Our system then searches other live sources of news for other stories related to the same event and produces summaries of a subset of the stories that it finds, according to parameters specified by the user.

2. THE NEWSINESSENCE SYSTEM

NewsInEssence's search agent, NewsTroll, runs in two phases. First, it looks for related articles by traversing links from the page containing the seed article. Using the seed article and any related articles it finds in this way, the agent then decides on a set of keywords for further search. In the second phase, it attempts to add to the cluster of related articles by going to the search engines of various news websites and using the keywords which it found in the first phase as search terms.

In both phases, NewsTroll selectively follows hyperlinks with the aim of reaching pages which contain related stories and/or further hyperlinks to related stories pages.

Both general and site-specific rules help NewsTroll determine which URLs are likely to be useful. Only if NewsTroll determines that a URL is "interesting", will it go to the Internet to fetch the new page. A more stringent set of rules are applied to determine whether the URL is likely to be a news story itself. If so, the similarity of its text to that of the original seed page is computed using an IDF-weighted vector measure. If the similarity is above a certain threshold, the page is considered to contain a related article and added to the cluster. The user may use our web interface (Figure 2) to adjust the similarity threshold used in a given search.

Using several levels of filtering, NewsTroll is able to screen out large numbers web pages quite efficiently. The expensive operation of testing lexical similarity is reserved for the small number of pages which NewsTroll finds interesting. Consequently, the agent can return useful results in real time.

3. ANNOTATED SAMPLE RUN

The example begins when we find a news article we would like to read more about. In this case we pick a story is about a breaking story regarding one of President-Elect Bush's cabinet nominees (see Figure 1).

We input the URL using the web interface of the NEWSINESSENCE system, then select our search options, click 'Proceed' and wait for our results (see Figure 2).

In response to the user query, NewsTroll begins looking for related articles linked from the chosen start page. In a selection from the agent's output log in Figure 3, we can see that it extracts and tests links from the page, and decides to test one which looks like a news article. We then see that it tests this article and determines it to be related. This article is added to the initial cluster, from which the list of top keywords is drawn.

In its secondary phase, NewsTroll inputs its keywords to the search engines of news sites and lets them do the work of finding stories. Since we have selected good keywords, most of the links seen by NewsTroll in this part of the search are indeed related articles (see Figure 4). Upon exiting, NewsTroll reports the number of links it has considered, followed, tested, and retrieved (see Figure 4).

The system's web interface reports its progress to the user in real time and provides a link to the visualization GUI once the cluster is complete (Figure 5). Using the GUI, the user can select which of the articles to summarize (see Figures 6 and 7). Figure 8 shows the output of the cluster summarizer.

4. FUTURE WORK

We are currently working on the integration of Cross-Document structure theory (CST) [1] with NEWSINESSENCE. CST is used to describe relations between textual units in multi-document clusters. It is used for example to identify which portions of a cluster contain background information, which sections are redundant, and which ones contain additional information about an event.

5. REFERENCES

[1] Dragomir Radev. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *Proceedings, 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, October 2000.

.

Figure 1: Seed article.

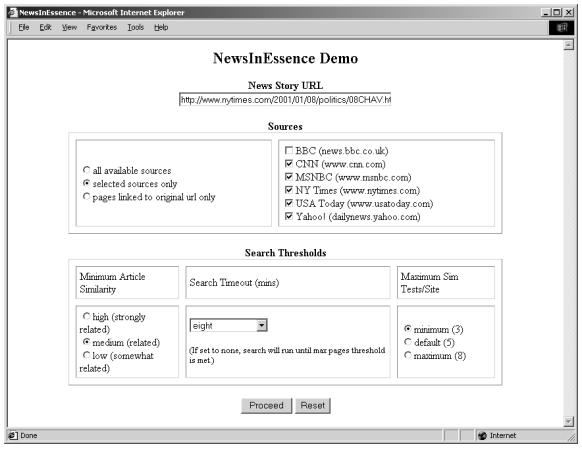

Figure 2: User interface.

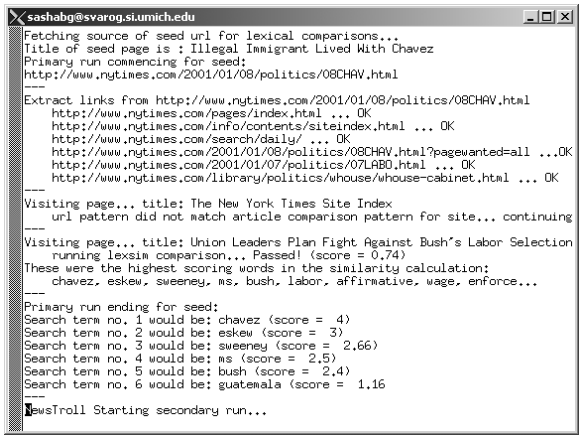

Figure 3: Run-time log (part I).

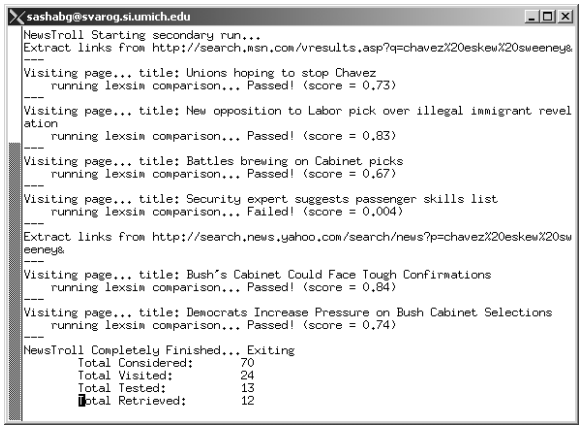

Figure 4: Run-time log (part II).

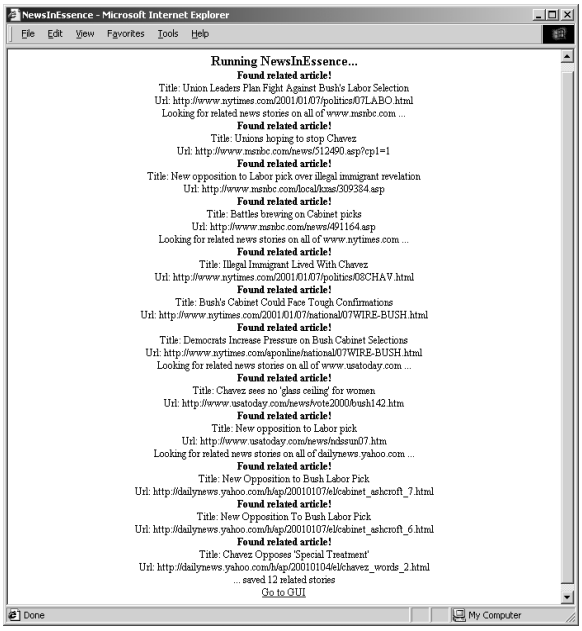

Figure 5: System progress.

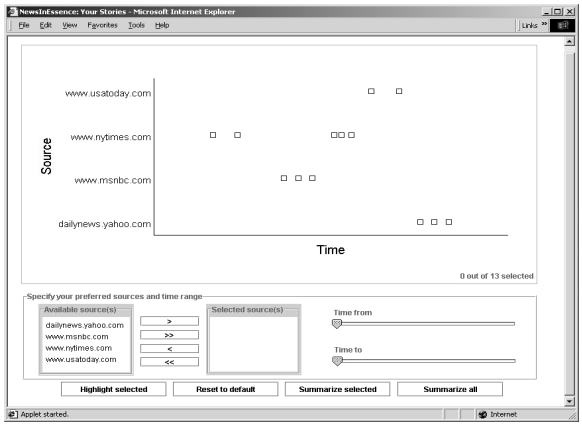

Figure 6: Cluster visualization.

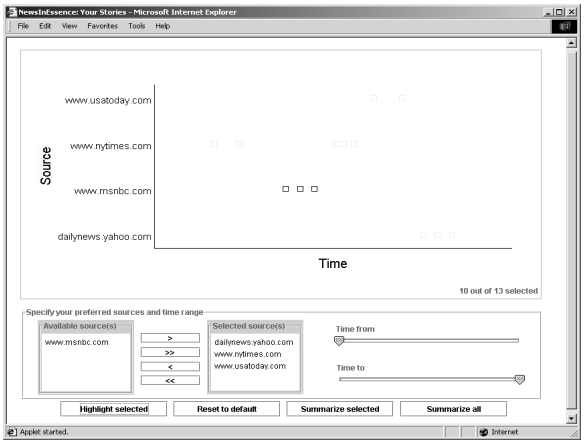

Figure 7: Selected articles.

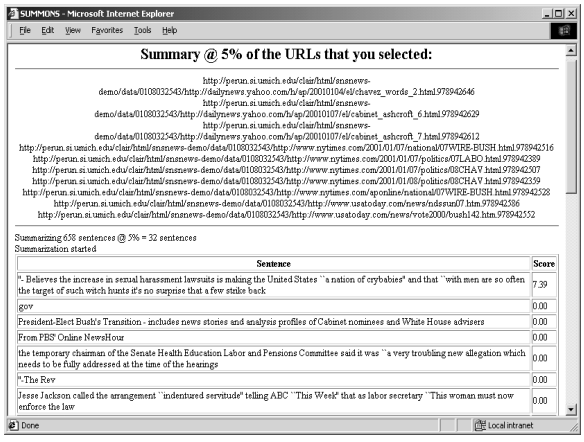

Figure 8: Summarization interface.