By **Dragomir Radev, Jahna Otterbacher, Adam Winkel,** *and* **Sasha Blair-Goldensohn**

NewsInEssence: Summarizing ONLINE NEWS TOPICS

A news delivery and summarization system, acting as a user's agent, gathers and recaps news items based on specifications and interests.

Internet users are turning more frequently to the Web for news rather than traditional media sources such as newspapers or television. This trend is likely to continue, according to a recent Forrester report [4], which found that Web veterans are more likely to cut back on reading print newspapers than people with less Internet experience. Indeed, the *New York Times*' online news source (nytimes.com) logs over 18 million distinct users monthly. Circulation numbers for the daily print edition are just over one million.

Reading news online offers many benefits over traditional media. Thousands of news sources are available, and speed of access has improved so that even geographically remote sites have easy access. Furthermore, nearly all news Web sites are free of charge.

Along with these benefits come challenges. NewsIsFree (newsisfree.com), a collection of links to news sites, currently lists more than 20,600 online news sources. With many of these sources adding dozens of stories daily, users can be overwhelmed with the sheer volume of news. For a reader interested in a given topic, this overload threatens to negate the benefits of online news because finding and reading all related stories becomes impractical.

NewsInEssence (NIE; www.newsinessence.com) [11], a news delivery and summarization system under development at the University of Michigan, helps alleviate these problems by acting as the user's agent to gather and summarize related online news articles. Given a user's topic specification (indicated via an example article or keywords), NIE searches across dozens of news sites to collect a group, or cluster, of related stories. It then generates a summary of the entire cluster, highlighting its most important content.

To build a news summarization service, it is important to consider how journalists write news stories. Most reporters are trained to use the inverse pyramid structure [3]: an article usually begins with a broad overview of the situation or event, followed by the finer details of the story. To the extent that writers follow this structure, it can be exploited by the summarizer.

Many summarizers, including NIE, create a summary by extracting salient sentences from the input documents. A challenging aspect of extractive multi-

ple-document summarization is that content and writing style may vary significantly from source to source. These stylistic differences can make it difficult to detect how two documents relate. This is particularly true on the Internet, as NIE may find itself comparing related articles published by news organizations in different countries, or intended for different audiences.

Other popular Internet news services (for example, AltaVista News or Google News) present clusters of related articles, allowing readers to easily find all stories on a given topic. However, these services do not produce summaries—a reader seeking a quick topic overview must choose between selecting a representative article to read in full or else skimming through all articles.

Since neither choice is ideal, systems like NIE and NewsBlaster (developed at Columbia University; newsblaster.cs.columbia.edu) provide summaries that give a representative gist of a cluster [5]. In addition, NIE uniquely allows the user to create personalized clusters and summaries.

NewsInEssence

NIE began as a research project at the University of Michigan in the summer of 2000, and has been online since March 2001. It offers user-driven clustering of articles, topic tracking, and multidocument summarization. NIE retrieves news articles from online news sources around the globe. In addition to nearly 20 U.S. sources, NIE retrieves news from the online versions of British, Canadian, South African, and Australian newspapers, as well as English-language versions of online newspapers from India, Singapore, and China, among others.

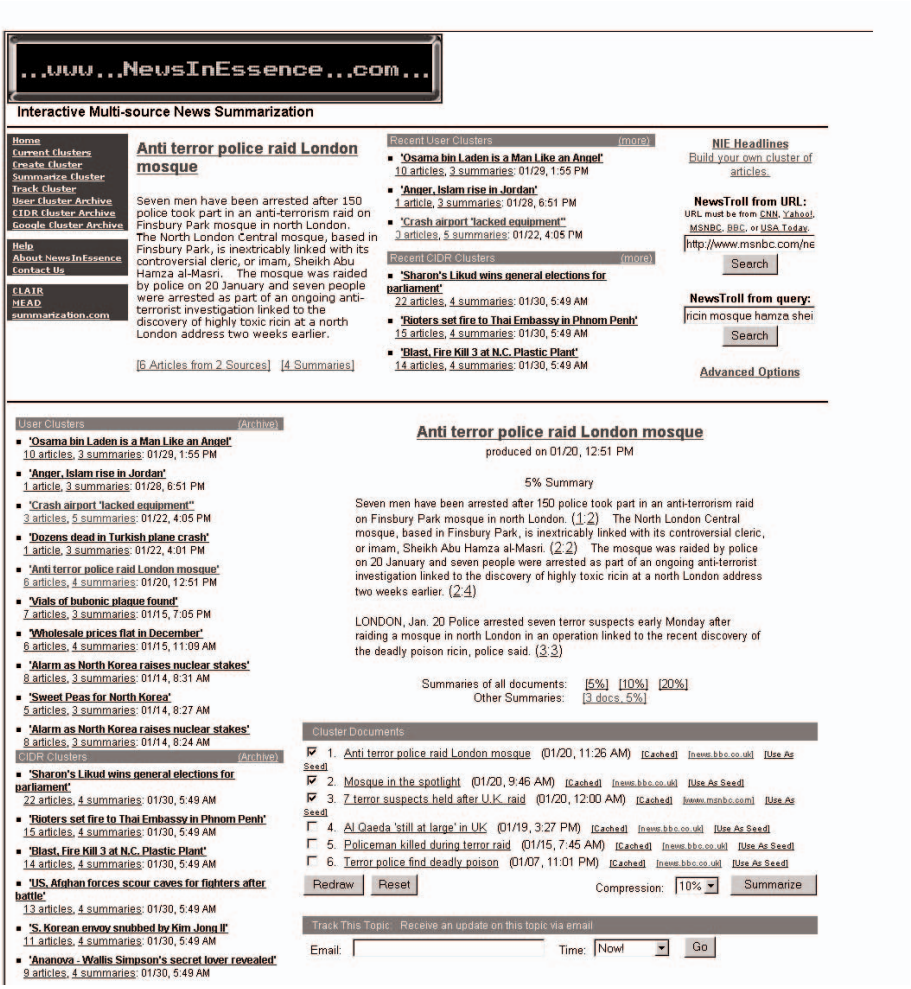

Figure 1. NewsInEssence front page.

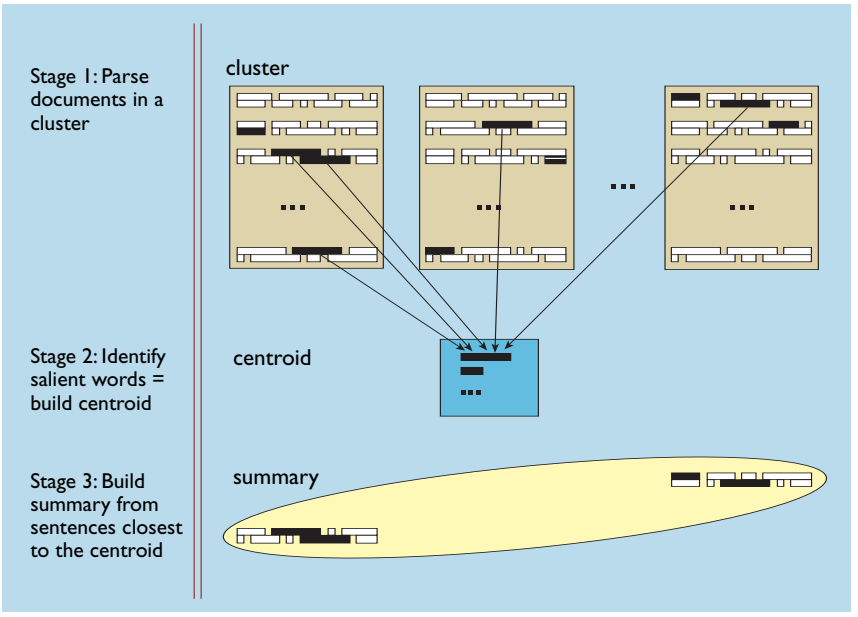

Figure 2. Centroid-based summarization.

The central object in NIE is the cluster, which consists of a number (typically 2–30) of topically related news articles. For each document in a cluster, NIE displays the article's title, source, publication date, and original URL (see Figure 1).

A teaser (a short cluster summary) dominates the top of the page. The teaser shown in Figure 1 displays two sentences from a cluster on a London police raid. Just below the teaser appear links to other existing summaries of the cluster. To the left is the navigation bar, which allows the user to quickly visit other NIE sections, such as an archive of past clusters.

The right-hand side of NIE's home page displays the most recent cluster at the top, with links to previous clusters below. The "NIE headlines" link allows the user to create a new cluster from a current story. Clicking on "NewsTroll from URL" starts a Trolling search.

Finding Relevant News by NewsTrolling

NIE creates news clusters in two different ways. First, users can have NIE's NewsTroll component create clusters for them. Given an example document, or seed, NewsTroll searches online news sites to find related articles to the seed. NewsTroll can also search for news related to a query, for example, "London mosque raid." The second way to build clusters is through CIDR (pronounced "cider")—NIE's topic detection and tracking component. CIDR runs several times per day and groups all articles it downloads by topic.

When a user invokes NewsTroll with a seed article, it first follows hyperlinks from the seed looking for related articles. When it runs out of direct links, NewsTroll creates a list of keywords important to the seed article and any related articles found by link following (this first stage is skipped in a keyword-based search, with the words coming directly from the user). Next, NewsTroll queries search engines of several online news sources using the keywords. These search engines each return a list of articles that NewsTroll retrieves and compares to the seed. Stories judged to be sufficiently similar to the seed are then added to the cluster, while others are discarded.

NewsTroll also allows the user to specify parameters, such as which sources to use, that can customize the retrieved cluster. The user can choose to have a high, medium, or low article similarity threshold, which determines how closely related to the seed (or keywords) an article must be to be added to the cluster. In addition, the user may set a time limit on NewsTroll's search.

Multidocument Summarization

NIE's summaries are produced by MEAD, a public-domain summarizer [12] that uses a sentence-ranking procedure known as the centroid-based method (see Figure 2). NIE uses MEAD to automatically precompute summaries at set compression rates. However, a user can also have NIE create a cluster summary using customized input parameters. For example, the user can specify the desired summary length or exclude specified articles from the summary.

With NIE's tracking option, a user can request an update on a specific cluster to be sent directly to his or her inbox at a specified time. For example, a user interested in the London raid might want an update on the story's progress the next morning without having to visit several news sites looking for the desired information. Entering a request causes NewsTroll to run at the specified time, looking for articles written since the original cluster was built. NIE then sends an email message to the user with the summary resulting from the new articles.

Feature	My Yahoo	SUMMONS	NewsBlaster	Google News	NewsInEssence
User-specified Seeds	No	No	No	No	Yes
Number of Sources	Over 7,000	2–3	27	Over 4,500	35
Cluster by Topic	No	No	Yes	Yes	Yes
Cluster by Category (US, World)	Yes	No	Yes	Yes	No
Search	Yes	No	Yes	Yes	Yes
Precomputed Summaries	Yes	No	Yes	No	Yes
Customizable Summaries	No	No	No	No	Yes
Text Generation Used	No	Yes	Yes	No	No

System comparison chart.

User Demand

My Yahoo, Google News, NewsBlaster, SUMMONS (the first multidocument summarizer [6]), and NIE represent different directions in the trend toward summarization of news clusters. The table here compares and contrasts their characteristics.

According to a recent Forrester report [4], the news of the future must be "formatted but flexible." The ease of access to news via new technology such as wireless Web, mobile phones, and PDAs, has empowered readers and raised their expectations of news delivery services.

Just as these services must be flexible in delivery media, they must provide news in a manner consistent with the expectations of the Internet user accustomed to accessing the news they want anytime they want it.

In short, users want to get their news in a manner that is convenient, timely, and customized to their interests and needs. Given these factors, NIE and similar systems will likely become indispensable for the news consumers of the future.

Future Work

Although current systems, including NIE, do a good job of identifying information in source articles important to the user's query, the resulting summaries are often not like summaries written by humans. Linguistic theory tells us that humans are taught to organize text in a particular way, with the overarching structure of the text in mind [3]. Given the nature of extractive multidocument summarization, where sentences are taken from various source texts and put together to form a summary, such a structure does not exist. As a result, the summaries sometimes do not seem to flow as evenly as they should, and they may be difficult to understand.

We believe we can improve our summaries by using Cross-document Structure Theory (CST) relationships as well as revision techniques. Relationships such as Identity, Paraphrase, and Subsumption are the focus of CST [10]. The first step in using CST to improve summaries would be detection of relationships among the candidate sentences to be included. Once we have used CST to determine which sentences belong in the summary, as well as their ordering, a revision module could detect unclear passages in the summary and correct them.

This revision procedure is necessary to address the cohesion problems that cause some of the flow problems mentioned earlier. For instance, if a sentence begins with the pronoun "he," but the reader cannot tell who "he" refers to, the revision module should replace the pronoun with the correct name.

Another improvement might be to resolve temporal relationships in the summary [8, 9]. Since source articles may have been written at different times, it is important to make sure the reader can understand what happened and when, in order to fully comprehend the story or event. Usually, this cannot be accomplished by simply reordering the sentences. Adding temporal phrases such as "on Monday" or "two days later" that place the event described in a given sentence into the overall context of the summary may help the reader's understanding of the timeline of events, as well as making the summary seem more cohesive [2].

Conclusion

In a recent TechStrategy Report [1], the Forrester Group predicted all types of news, from national to local, will be available on demand over the next 15 years via a number of different media outlets. News sources will collaborate closely, which means that integrating diverse resources will be a necessity. We believe that NIE and its counterparts are the first steps toward this user-driven access.

Forrester also predicts that automated news technologies like NIE and NewsBlaster will be used to handle the summarization and rewriting of old or mundane news stories, allowing reporters to focus on more difficult reporting jobs. Soon, the Canadian Broadcasting Corporation will make pre-taped news reports available on demand, allowing users anytime access to news. In the more distant future, we can imagine this process being taken a step further: Users might specify their choice of virtual anchors—computer-animated video of human faces—to deliver, in real time, news collected and summarized according to their preferences by future generations of systems like NIE and NewsBlaster. C

References

1. Allen, L., Charron, C., and Roshan, S. Re-engineering the news business. Technical Report, The Forrester Group, June 2002.
2. Allan, J., Gupta, R., and Khandelwal, V. Temporal summaries of news topics. In *Proceedings of the 24th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval,* 2001.
3. Halliday, M.A.K. and Hasan, R. *Cohesion in English*. Longmans, London, 1996.
4. Kelley, C.M. and DeMoulin, G. The Web cannibalizes media. Technical Report, The Forrester Group. May 2002.
5. McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B., and Sigelman, S. Tracking and summarizing news on a daily basis with Columbia's NewsBlaster. In *Proceedings of Human Language Technology Conference (HLT 2002),* (San Diego, CA, Mar. 2002).
6. McKeown, K.R. and Radev, D.R. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* (Seattle, WA, July 1995).
7. Mitchell, C.C. and West, M.D. *The News Formula: A Concise Guide to News Writing and Reporting.* St. Martin's Press, New York, 1996.
8. Otterbacher, J.C., Radev, D.R., and Luo, A. Revisions that improve cohesion in multidocument summaries: A preliminary study. In *Proceedings of the Workshop on Automatic Summarization (including DUC 2002),* (Philadelphia, July 2002). Association for Computational Linguistics.
9. Pustejovsky, J. et al. The timebank project; www.time2002.org.
10. Radev, D.R. A common theory of information fusion from multiple text sources, step one: Crossdocument structure. In *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue,* (Hong Kong, Oct. 2000).
11. Radev, D.R., Blair-Goldensohn, S., Zhang, Z., and Raghavan, R.S. Interactive, domain-independent identification and summarization of topically related news articles. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries,* (Darmstadt, Germany, 2001).
12. Radev, D.R., Jing, H., and Budzikowska, M. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL Workshop on Summarization,* (Seattle, WA, Apr. 2000).

Dragomir Radev (radev@umich.edu) is an associate professor of information and electrical engineering and computer science at the University of Michigan, Ann Arbor. He also leads the CLAIR (Computational Linguistics and Information retrieval) research group.
Jahna Otterbacher (jahna@umich.edu) is a Ph.D. candidate in information at the University of Michigan, Ann Arbor.
Adam Winkel is a computer scientist and AI technical lead at ChoiceMaker Technologies in New York City.
Sasha Blair-Goldensohn is a Ph.D. candidate in computer science at Columbia University in New York City.