

News Sync: Three Reasons to Visualize News Better

V.G.Vinod
Vydiswaran
University of Illinois
Urbana, IL

vgvinodv@illinois.edu

Jeroen
van den Eijkhof
University of Washington
Seattle, WA

jeroen@uw.edu

Raman
Chandrasekar

Microsoft Research, Redmond, WA

Ann
Paradiso

Jim
St. George

{ramanc, annpar, jamessg}@microsoft.com

ABSTRACT

News consumption patterns are changing, but the tools to view news are dominated by portal and search approaches. We suggest using a mix of search, visualization, natural language processing, and machine learning to provide a more captivating, sticky news consumption experience. We present a system that was built for three scenarios where a user wants to catch up on news from a particular time period, location, or topic. The results cover key events from that time period and are prioritized based on the user's interests. Further, users can interact with and explore stories of interest. An initial prototype is currently being piloted.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *information filtering, selection process.*

General Terms

Algorithms, Human Factors.

Keywords

News summarization, clustering, news adaptation, news exploration, news interfaces.

1. INTRODUCTION

The news landscape has undergone major changes with the advent of online media. While the readership of traditional newspapers has declined over the past few years, the consumption of news over the Internet has increased significantly. In a March 2010 survey of US Internet users [1] on the primary source used to find news, it was found that the Web/Internet is by far the most popular source (49%) as compared to Television (32%) and Newspapers (9%). As with other kinds of online information, the dominant mode of assessing news online is through search. According to a Pew Research survey conducted over Apr–Jun 2008 [5], 83% of those going online for news use search engines to find stories of interest. So, even though there are several dedicated news portals, consumption of news is triggered primarily through queries. Search engines today address this user behavior by integrating relevant news results with Web search results for news-related queries and provide news verticals and topic-specific news pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HCIR '10, August 22, 2010, New Brunswick, NJ, USA.

Copyright 2010 ACM 1-58113-000-0/00/0010...\$10.00.

However, the presentation of news is not optimal on these sites. Even as news is shifting online, the presentation of news is still driven by the print media. There is limited real estate on the search result pages to display news, and many news articles do not get surfaced on the site. Finding relevant news is more than just retrieving news results or restricting the search based on keyword queries over the news domain. Presentation of news needs to cater to specific user needs. We propose a use-case or scenario driven approach to selecting relevant news stories and presenting these appropriately to the user. Further, users should be able to explore the news landscape – getting to other related news articles, visualizing the connections between stories, getting background information on relevant people and concepts, commenting on and annotating stories, and sharing interesting items with friends.

In this paper, we present our system, called *News Sync*, which was developed to enable such enhanced news experience. The system is built over the New York Times Corpus released as part of HCIR Challenge. In Section 2, we present the motivation for this system with the help of three use-cases, and list key features that must be present in such a system in Section 3. In Section 4, we present our solution in detail and show how these use-cases can be addressed by the system. We conclude in Section 5 with related work and future steps.

2. NEWS EXPLORATION SCENARIOS

In this section, we present three specific scenarios for a user-driven news digest to illustrate our ideas. We propose techniques to select and present news according to user needs and preferences. While the techniques used may not be new, we suggest that the integration we propose will lead to a better news experience.

2.1 Scenario 1: Catching up on News

Consider the following scenario: Katie is an avid news reader who tracks news on a daily basis, often following up on specific news events several times a day. At times, Katie may be cut off from news, for example, when she goes on a long vacation. When she is back online, she may want to know what happened while she was away. She may want to skim through the major news stories that took place, including updates on the news she was following regularly before going on vacation.

This caters to a common, specific need of a news consumer wanting to catch up on news.

2.2 Scenario 2: Diaspora Digest

It has become fairly common for people to migrate to another country or city for work or studies. Though most of these expatriates try to keep abreast with the news from the country of origin, they lose touch with traditional sources of news. They visit news websites from the home country periodically to do so. If

Katie is from Berlin and residing in the US, she might not be interested in local news bulletin from Berlin, but might be interested in a summarized view of key events in Germany from the past week. She might be interested in the country's soccer team's performance round the year and also country-wide soccer competitions such as the German Cup.

This need caters not only to expatriates living in another country but also people migrating to other cities within a country.

2.3 Scenario 3: Following Celebrities

A longitudinal look at news is of great value for specific needs, such as following the activities of celebrities. Assume Katie is an admirer of Princess Diana and she wants to get a perspective of Princess Di's life history as described in the news. She would be interested in key events such as her marriage, her time as the princess, her divorce, and her death and subsequent investigations. The key idea here is that the user gets a historic perspective on celebrities using archival news content.

3. REQUIREMENTS FOR *News Sync*

To address the above scenarios, we propose a system we call *News Sync*. This allows Katie and similar news consumers to get adaptive, personalized news digests covering a period of time, a region, a topic, or a combination of these.

We list the following requirements for *News Sync*, a modified version of the requirements we presented in [13]:

1. **Control over news categories, topics, and sources:** The user should be able to specify the time period of interest. In addition, the user may specify if she is interested in news from particular sources, specific news categories, locations/regions, and/or specific topics.
2. **Personalized news feed:** The system should identify stories that are currently the most relevant to the user, based on past user behavior and user preferences.
3. **Variety in news content:** The system should show a variety of content across diverse categories, instead of, say, returning a list of ten "most popular" news links which may be restricted to one or two topics. Users can thus get an overall picture of key events first, before they delve into specific stories.
4. **Adaptive and integrated news presentation:** The news interface needs to be adaptive to the category of news and presence of multiple modes of news content. For example, news about Harry Potter over Summer 2007 should include, among other stories, the trailers from the movie "Harry Potter and the Order of Phoenix" (video), book reviews of "Harry Potter and the Deathly Hallows" (text, blogs) – which were both released in July 2007 – along with pictures and news about the Harry Potter theme park announced in May 2007 (images).
5. **Interactive and exploratory user interface:** The user should be able to interactively and directly modify time, location, and other parameters and have the system respond immediately with updated views of relevant news.
6. **Parameterized interface design:** Users should be able to set system parameters to get results at different specificities.
7. **Support source-tracing and finding related news:** The system should allow users to go from any news summaries to the original news articles. Further, the system should suggest other related news articles based on the news items viewed.
8. **Ability to share news:** Users should be able to comment on and share interesting news articles over their social network.
9. **Support news analyses by sentiment and points of view:** Users should be able to view stories pivoted/summarized on sentiment or different points of view.
10. **Keep the familiar list-view as back-off:** Even as the news interface gets a facelift, it may be prudent to maintain the list-based view as a back-off option to take advantage of familiarity with the concept.

Such a system would help frequent travellers, business customers who need to know the impact of ongoing news on their business, and avid news followers who spend a lot of time with news.

4. THE *News Sync* SYSTEM

In this section, we present a brief description of *News Sync*, the system we developed based on requirements listed in Section 3.

4.1 System Description

Figure 1 gives a schematic diagram of the *News Sync* system.

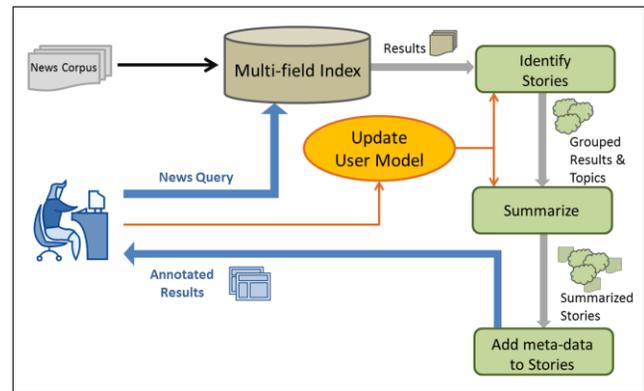


Figure 1. Schematic Diagram of *News Sync*

The key steps in the system are:

1. **Collecting a news corpus:** Our first step is to get access to the news articles for the time period of interest to the user population. Articles are processed with a named entity recognizer, to identify key concepts. In this prototype, we use the New York Times corpus, released as part of the HCIR 2010 Challenge. In addition to all articles published (or posted online) by New York Times from 1987 to 2007, the corpus also contains rich meta-data such as normalized list of people, locations, and organizations found in the articles.
2. **Indexing the corpus:** The New York Times corpus was indexed using Lucene.Net [3], such that each field can be queried individually. This involved removing frequently occurring words (stop words) and spurious characters, and additional pre-processing to normalize some fields, such as publication date, to make them searchable.

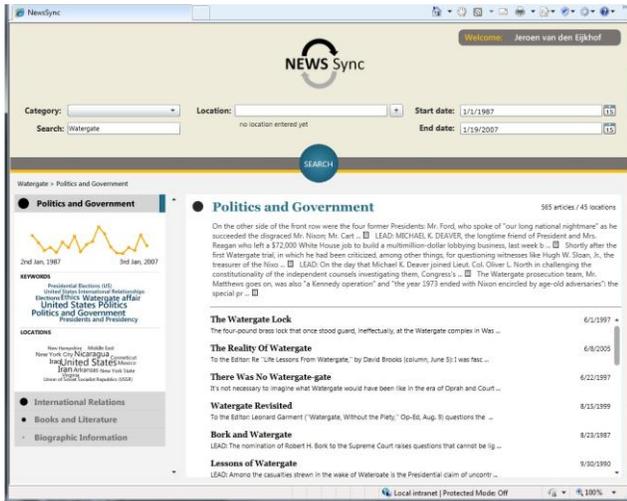


Figure 2. Screenshot of News Sync showing results for the query “Watergate” in the catching-up scenario.

- Retrieving relevant news results:** When the user issues a query for news, the system converts the query to an appropriate Lucene [2] query. If the category or location is specified, they must appear in the document. If a date range is specified, only results from that date range are retrieved.
- Grouping news articles:** News needs to be presented in a manner that is easy to consume. This involves selecting the content to present and deciding how best to present it. In this work, we cluster articles to find related groups of articles. Each group may not be a single story thread, but this dimension reduction by clustering offers a more structured view into the articles. Recursive clustering can help us get to news stories, which are collections of tightly related articles. These news clusters may be adapted to the user model (user profile, explicit user preferences, and implicit interest tracking). We currently cluster on key concepts from articles, including named entities, descriptors, categories, and section headings obtained from article meta-data.
- Summarizing news clusters:** We adaptively summarize the clusters, to provide some insight into the articles in a cluster. Summarization is performed using a modified version of SumBasic [9].
- Add aggregated meta-data about the clusters:** Each news cluster is annotated with additional meta-data such as the news timeline, relevant categories, locations, and key concepts from the articles.
- Presenting and visualizing news:** Once the news clusters are annotated, they are presented to the user along with relevant meta-data. The meta-data, presented in the form of sparklines and tag clouds, can be used to further refine and explore news clusters.

The system is developed in C#. The interface is developed using Microsoft Silverlight [4], since it gives us access to animation and interactivity, and provides browser independence.

4.2 User Interaction

We now sketch the interaction flow for the system:

- Providing search parameters:** When Katie logs in, she is shown a tag cloud of key topics from the corpus. She can browse for news by providing one or more of four input

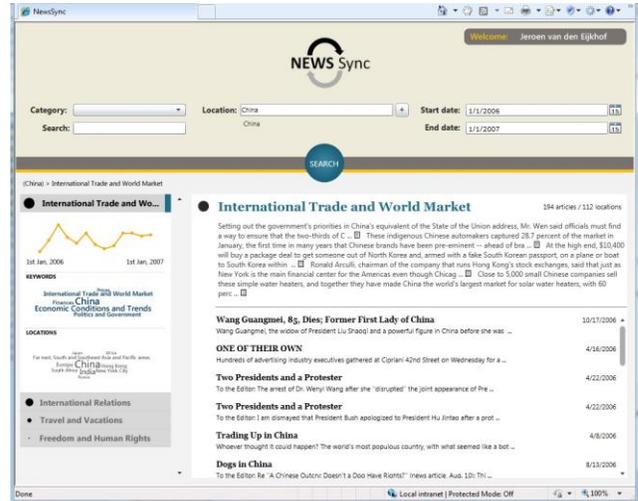


Figure 3. Screenshot of News Sync showing results for the diaspora query about China.

- parameters – the news category, topics of interest (keywords), location(s), and a date range of interest.
- Viewing news clusters:** When Katie enters a news query, consisting of one or more of the parameters, she is shown dynamically generated clusters of related articles. Figure 2 shows a screenshot of the results for the catching-up scenario query “Watergate”. The left panel of the result screen lists clusters, ordered by popularity and relevance. The top-most cluster is highlighted and the left panel displays additional properties about the selected cluster, such as key concepts and locations mentioned in the news articles. A sparkline shows the distribution of articles with time. The right panel gives additional information about the highlighted cluster. It shows a brief summary, followed by the list of relevant articles. The list shows the date of publication, headline, and lead paragraph for each article. Figure 3 shows a similar screen for the diaspora scenario, where the user is looking for information about China.
- Browsing news results:** Katie can either explore the articles in the current cluster or can look into other clusters from the left panel. If she clicks on the article headline, the article and all relevant meta-data is displayed (see Figure 4). If she clicks on another cluster from the left panel, the section with additional properties on the first story shrinks, and the newly selected cluster expands to show its properties. Katie can also select a portion of the timeline; as the date range is varied, the articles from that date range are highlighted in real-time. This allows Katie to zoom into news from a specific time period. If Katie is interested in exploring a particular topic further, she can select a topic and choose to dig deeper. A new query is then issued based on the chosen topic and the original query to get a refined search experience.
- Sharing results:** The interface also allows Katie to share the summary, articles, or stories with her friends on popular social networking sites. She can also save the query/results.
- Following user actions:** As Katie interacts with the system, her actions, queries, and parameter settings are stored. When Katie reads articles and shares it with her friends, the key concepts from the article are recorded in user models maintained per

user. The ranking and summarization of clusters are continuously adapted based on the user model.

Katie can also explicitly restrict her results to be from particular regions or categories. These customization preferences are recorded and subsequent results are tuned to these preferences.



Figure 4. News Sync article obtained after clicking an article in the summary view shown in Figure 2.

4.3 Evaluation

We are currently piloting an initial prototype of News Sync. We are deploying it to a small user-base to understand how users interact with the system, using implicit and explicit feedback. We are also conducting a survey to understand usage patterns and features that are popular.

5. RELATED WORK

Past literature has looked into generating a personalized webpage of news relevant to the user based on the topics of interest. Kamba et al. [8] conducted one of the early studies on presenting an interactive newspaper on the Web. They propose a system that builds web pages dynamically as the user browses the newspaper. Anderson and Horvitz [6] developed a personalized web page as a montage of links of frequently viewed pages that changes dynamically with the time at which the page is viewed. The system learns which pages are viewed regularly at certain time periods and presents content based on the user’s interests and browsing pattern. For example, a user might be shown weather forecasts and key news in the morning; the stock price ticker and work-related resources during the day; and traffic pattern and TV listings in the evening.

There has also been work in providing personalized newsfeeds. Gabrilovitch et al. [7] analyzed inter-/intra-document differences and similarities to recognize novel content in articles and how the information has evolved over time. This helps them develop measures to rank news by novelty, and pick the best (most novel) update to send to the user as a newsfeed. Other researchers, such as Tintarev and Masthoff [12] have studied different measures of similarity of news headlines to improve news recommendation.

There is a lot of relevant work in the realm of interface design. For example, Shneiderman [10] suggests use of dynamic queries to update the search results as users adjust sliders and other UI elements. Teitler et al. [11] suggest NewsStand, which proposes

using geographic information in news articles to overlay news on a map. This presents users with a geographic perspective of where the news comes from and helps them cluster and explore news based on location.

Some news ranking sites are able to show “popular” news for particular days or months, based on how many users clicked on or shared a news article.

6. CONTRIBUTIONS AND CONCLUSION

In this paper, we propose an approach to providing a captivating, sticky news consumption experience, using techniques from search, language processing, visualization and learning. We listed requirements for three news exploration scenarios. We presented our prototype, called News Sync, which is currently being piloted.

In the News Sync prototype, we provide controls to users to explore news by specifying topics, a time range, and/or locations of interest. We react immediately to user inputs to show not just the relevant articles, but additional information including clusters and summaries, tag clouds of locations and key concepts and a sparkline to show temporal trends. We also adapt results based on user preferences and a model of the user acquired over time, to ensure that the user gets maximally relevant content.

In this work, we have relied on news only from a single source, namely the New York Times. We hope to extend this to multiple sources, deal with different points of view and sentiments, and work with live news streams.

7. REFERENCES

- [1] Gather: The Changing Face of News Media, May 25th, 2010, www.emarketer.com
- [2] Lucene. <http://lucene.apache.org/java/docs/index.html>
- [3] Lucene.Net. <http://lucene.apache.org/lucene/net/>
- [4] Microsoft Silverlight. <http://www.silverlight.net/>
- [5] Pew Research Center for the People and the Press, 2008. Key News Audiences Now Blend Online and Traditional Sources – Audience Segments in a Changing News Environment. *Pew Research Center Biennial News Consumption Survey*. <http://people-press.org/reports/444.pdf>
- [6] Corin R. Anderson and Eric Horvitz. 2002. Web Montage: A Dynamic Personalized Start Page. In *WWW*.
- [7] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. NewsJunkie: Providing Personalized Newsfeed via Analysis of Information Novelty. In *WWW*.
- [8] Tomonari Kamba, Krishna Bharat, and Michael C. Albers. 1993. The Krakatoa Chronicle - An Interactive, Personalized, Newspaper on the Web. In *WWW*.
- [9] Ani Nenkova and Lucy Vanderwende. 2005. The Impact of Frequency on Summarization. MSR-TR-2005-101.
- [10] Ben Shneiderman. 1994. Dynamic Queries for Visual Information Seeking. In *IEEE Software*, 11(6), 70-77.
- [11] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. 2008. NewsStand: A New View on News. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*.
- [12] Nava Tintarev and Judith Masthoff. 2006. Similarity of news recommender systems. In *Workshop on Recommender Systems and Intelligent User Interfaces*, G. Uchyigit (Ed).
- [13] V.G.Vinod Vydiswaran and Raman Chandrasekar. 2010. Improving the Online News Experience. In *HCIR*.