A SYSTEM FOR NEWS VISUALIZATION

S.R. SUBRAMANYA

School of Engineering and Computing National University San Diego, CA 92123, USA
E-mail: ssubramanya@nu.edu

Abstract- Digital news content has been witnessing tremendous growths both from traditional publishers as well as user generated content. Due to changing lifestyles, people have been increasingly watching news available on the Web sites on their desktops, laptops, and mobile devices for shot durations of time, in between activities. Thus, it is extremely important to provide facilities to get a glimpse of news on a Web site or on a mobile device in a manner that is appropriate for use in time-constrained situations. The corresponding user interface design is also challenging. In this project, we have designed and developed a system for gathering news from the Web and providing a visualization of the news using icons and brief texts at the appropriate locations on a map. This facilitates news consumption in a quick and effective manner.

Index Terms- Automatic news gathering, news visualization, news on the Web.

I. INTRODUCTION

The phenomenal advances in the computing, communications, consumer electronics, the Internet, and mobile technologies have led, among other things, to phenomenal increases in the generation, exchange, transmission, interaction, and consumption of vast amounts of digital content. In addition to traditional news sources, there has been an explosion of the user generated news content available on the Web and on social media (Twitter, Facebook, etc.).

We have also been witnessing tremendous influences of the above mentioned technologies on peoples' lifestyles in recent times – there have been marked increases in the levels of busyness and multitasking. As an example, a study found that a ten-second wait for a Web page to load can make 50% of consumers give up and leave the site. Researchers at Microsoft even found that a website begins losing traffic to competitors when it takes 250 milliseconds longer to load. With respect to reading/watching/listening to news, people (working / young adults) are less willing or unable to afford to allocate considerable time for leisurely reading of news. It has been established that pictures/graphics/icons convey lot more information compared to text using similar number of bits. It has also been shown that multiple media enhance the retention of information. Data visualization has been used in several domains – scientific data, retail industry, financial sector, etc. Graphical visualization also facilitates recognition of interrelationships among data much better than pouring over numbers and text.

The proposed news visualization system captures news from Web sites and extracts titles, keywords, and the initial paragraphs. It maintains a database of the news items. In the context of the project described in this paper, the major benefit of visualizations for reading online news is to get a quick glimpse of news in a region, a country, or the world, in contrast to clicking on headlines to read the news with text and pictures on a Web site. News items of particular interest could be pursued in detail using the URL in the textbox associated with the icons. This paper is organized as follows. The next section presents some background and newspaper readership trends driven by advances in technologies. Section III presents a few initiatives in the area of data and news visualization. The motivation, benefits, architecture, and implementation of the proposed system are given in Section IV. The future directions of the project are outlined in Section V, followed by conclusions.

II. BACKGROUND AND NEWSPAPER READERSHIP TRENDS

Although the amount of news generated has been increasing over the years, the recent technological advances have given a new impetus to the generation of digital news content – both conventional publishes as well as user generated content. This, together with the changing lifestyles of people, has had effects on the modalities of news consumption. The data from Pew Research Center about daily readership across age groups of (traditional) newspapers is shown in Fig. 1 below. It is evident that over the last 14 years there has been a steady decline over the years.

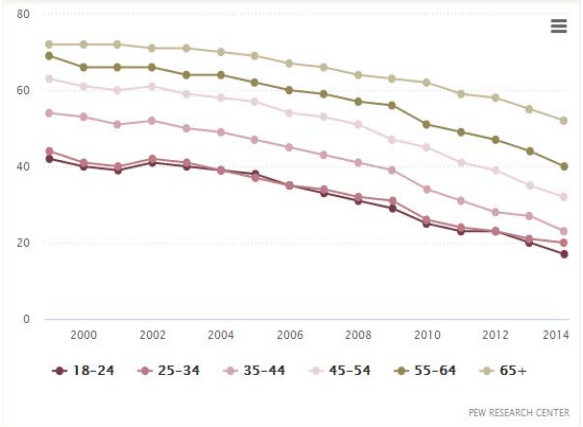


Figure 1. Trends of traditional newspaper readership

Almost all traditional newspaper publishers have a digital version accessible on the Web via fixed or mobile devices. The digital versions of the newspapers

could take several forms such as (i) Apps on mobile devices (smartphones and tablets), (ii) "epaper" version (PDF replica) which is a digital replica of the actual printed newspaper, (iii) metered/restricted Web sites, and (iv) e-reader editions. Since the readership of conventional newspaper has been decreasing, the newspaper publishers have been developing strategies for increasing the paid readership of their digital content. The circulation of top six digital newspapers (Oct. 2012 to Mar. 2013) is shown in Fig. 2.

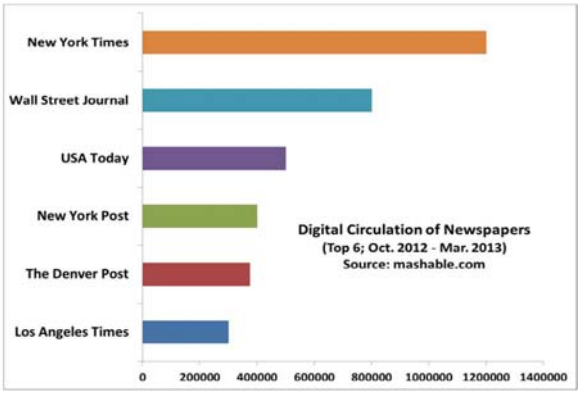


Figure 2. Readership of digital version of newspapers

According to Newspaper Association of America (source: naa.org), the number of subscribers to digital version of newspaper content reached a new peak in August 2015, totaling 179.3 million adult unique visitors. This is a 10% increase in adult unique visitors measured by comScore since August 2014, and is double the growth rate of the Internet of 5%. About 50% of the digital content subscribers use only mobile devices (smartphones or tablets) for their newspaper digital content. The SenseMaker Report by the Newspaper Association of America [2] derives from the data about consumption from the Scarborough Research survey of some 206,000 U.S. audience of different media sources across platforms. Some points are: (i) 164 million (69%) of U.S. adults read newspaper media content in print or online in a typical week, or access it on mobile devices in a typical month; (ii) the majority (59%) of young adults (ages 18-24) read newspaper media content in print or online in a typical week or access it on mobile devices in a typical month; (iii) the mobile newspaper audience was 34 million adults in 2012 (there was an increase of 58% in an average month in 2012 compared with 2011).

III. RELATED WORK

The data visualization types play different roles in understanding data. In order to address the span of visualization types, a new taxonomy for static (non-interactive) visualizations is given in [3]. The taxonomy classifies static visualizations according to the underlying data structures, the visual encoding of the data, and the perceptual tasks enabled by these encodings. It contains twelve main visualization categories and several popular sub-types for each category. In addition, a set of properties that aid in the characterization of the visualizations are given.

The study in [3] used 2,070 single-panel visualizations, categorized with visualization type (e.g., bar chart, line graph, etc.), collected from news media sites, government reports, scientific journals, and infographic sources. The findings therein suggest that quantifying 'memorability' which would serve as a metric for the utility of information would be helpful in designing effective visualizations.

In [9, 10] the authors describe "Mapping Texts", new methods for finding and analyzing meaningful patterns embedded in massive collections of digital newspapers. It was designed to fit into an ecosystem of tools, processes, and methods in text-mining, analysis, and visualization. It combined the two most promising methods for analyzing large-scale datasets: data- and text-mining (for discovering meaningful patterns embedded in large bodies of text) and data visualization/mapping (for grouping, discovering, analyzing, and making sense of those patterns).

Using a collection of 232,500 pages of historical newspapers digitized for the Chronicling America project, the project in [9, 10] developed two interactive visualizations as part of the system. The visualizations allow one to explore both the quality of those digitized newspapers and the major language patterns. As for the quality, the objectives were to see how much information was available for any particular time and place, and be able to measure the amount of information which was still recognizable and useable after the digitization process.

In interactive infographics, the user can be guided in the desired way by selecting the technology and data presentation method that serves the purpose. In [5] the authors explain this situation as: "Sometimes you may want the user to browse your information in order to derive meaning that is relevant or interesting to them. Or you may use interactive infographics to guide someone through a specific narrative in a linear fashion, so they grasp the specific story that you want to tell. Keep in mind that this isn't an either/or decision; you can use the narrative/editorial angle to tell people what you see as significant, interesting, or entertaining in order to entice them to view the content, and then encourage them to explore the data further to find information that is relevant to them".

Google News gathers news from thousands of sources worldwide and presents information with pictures and text. It updates entries continuously throughout the day. It groups together related headlines and photos. By default, it automatically displays the latest headlines. Google Alerts facilitates monitoring a developing news story (e.g. keep current on a competitor or industry; learn where a company is cited or quoted; get the latest on a celebrity or event; keep tabs on your favorite sports teams; find when people

link to your site; discover new websites on a certain topic, etc.). A news visualization system from [1] is shown in Fig. 3. A treemap visualization algorithm helps display the enormous amount of information and the relationships between data and the unseen patterns in news media. Mousing over a block pops up a window with a picture and text related to the news item. It provides the date/time, main news source, links to other sources, etc.

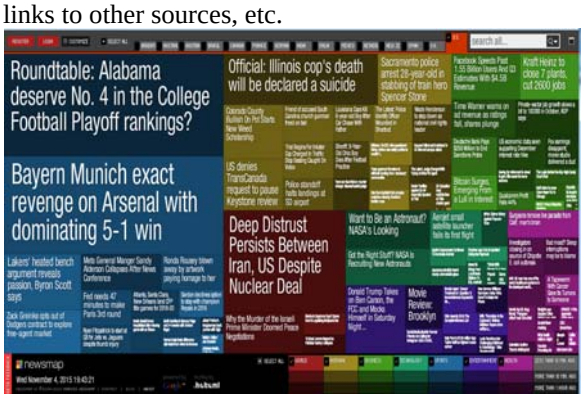


Figure 3. A sample treemap display

Treemaps [6, 4] display hierarchical (tree-structured) data as a set of nested rectangles. A rectangle is associated with a branch of the tree, and the smaller rectangular tiles within a rectangle are represented by sub-branches. A leaf node's rectangle has an area proportional to a specified dimension of the data. The leaf nodes are colored to show a separate dimension of the data. A tiling algorithm is used to determine a way to divide a rectangle into sub-rectangles of specified areas. The design of content targeted for mobile devices and the user interface for mobile content are very crucial in ensuring a pleasant user experience and effective consumption of content. Some of the major content-related issues that need to be addressed in mobile environments are presented in [7]. Several factors in the design of the user interfaces for mobile content/devices is are given in [8].

IV. PROPOSED SYSTEM

True to the adage "a picture is worth a thousand words", pictures and icons are known to convey a lot of information in a limited time. The proposed system provides a user interface to visualize news in the form of icons on a map– the graphic of icon reflecting the news category. The motivation for the proposed system is based on the recent technology trends, the changes in news content consumption, and changes in lifestyles and expectations of people.

A. Motivation

The technological advances described in Section I have resulted in explosive growths of digital news content from traditional news publishers as well as user generated content. The enormous and growing digital news content are available via Web sites and portals, and are accessible using fixed or mobile devices. The lifestyles of people are getting very busy and demanding, which leaves little time to keep track of all the news all over. Thus we have a situation where more and more digital news content is generated, and people have less and less time to read them. Thus the gap between digital (online) news production and consumption is widening. It may result in several news items being never brought to the fore. It is thus desirable to create a system which provides a way of presenting digital news content in a way which facilitates a very high level of efficiency in consumption – namely maximizing the amount of news 'perceived' in a limited time. The proposed system is expected to fill this need. This is illustrated in Figure 4.

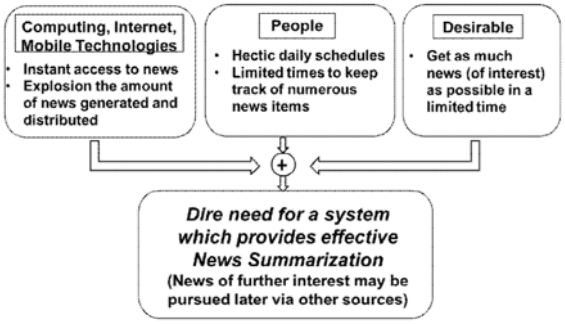


Figure 4. Need for a news visualization system

B. Expected Benefits of the System

The design and development of the proposed system is driven by the need to bridge the widening gap between as outlined earlier. The proposed system provides several benefits and a few major ones are given below.

- Intuitive interface via Icons, Text, and GUI – icons corresponding to news placed at the appropriate locations on a map facilitates more efficient consumption of a considerable amount of news in a short time.
- News summaries available anytime, anywhere – this helps in getting quick news updates as and when they happen via fixed or mobile devices.
- Choice of user–specified locations and topics – the news within a specified location (city / state / country) and specified topics (politics, sports, business & finance, etc.) can be obtained easily.
- Overview of lots of news items in a short time – provides a natural hierarchical view of news, with a quick glimpse via icons placed on a map, then through titles and summaries in text boxes, and pointers to news details via URLs.
- Overlay of news items in time windows – the news icons of a few days can be displayed in an overlay manner. This helps in getting a glimpse of areas on the map about what things are happening over time and where.
- Chain of news events over time – news which develops over time is chained and maintained in the system database. These can be listed and selected for

C. System Architecture

A high–level architecture of the proposed news visualization system is shown in Fig. 5. News items from a few online news sources are captured from their Web sites. The location, date, and the headlines and the first paragraphs of the main news items are extracted. This is passed on to JSOUP. JSOUP is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of CSS and JQuery-like methods. Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language (such as HTML).

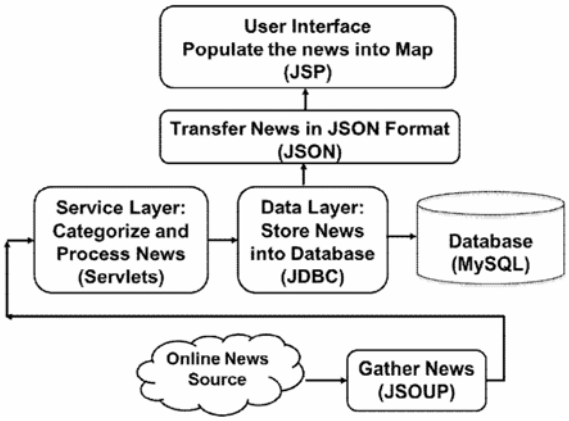


Figure 5. High–level architecture of the proposed news visualization system

The categorization and processing the captured news is done at the service layer by Java servlet. A Java servlet is a Java program most commonly implementing applications hosted on Web servers. (Such Web servlets are the Java counterpart to other dynamic Web content technologies such as PHP and ASP.NET).

The storage of news related data and the metadata is done at the data layer consisting of JDBC (Java DataBase Connectivity). JDBC is a Java database connectivity technology (Java Standard Edition platform) from Oracle. This technology is an API for the Java programming language that defines how a client may access a database. It provides methods for querying and updating data in a database. JDBC is oriented towards relational databases.

The queries in our system are handled by MySQL which is an open source relational database management system, and is a popular choice of database for use in Web applications. It is a central component of the widely used LAMP (Linux, Apache, MySQL, Perl/PHP/Python) open source Web application software stack (and other "AMP" stacks. The APIs in C# which support accessing MySQL databases are used, together with the JDBC driver for Java.

The news data from the data layer is converted to JSON format so it can be handled at the Web UI (User Interface). Although originally derived from the JavaScript scripting language, JSON is a language-independent data format. JSON (JavaScript Object Notation) is an open standard format that uses human–readable text to transmit data objects consisting of attribute–value pairs. It is the primary data format used for asynchronous browser/server communication (largely replacing XML, used by AJAX).

The uppermost layer of the system is the Web UI of the system facilitates population of news into the database, and the query and retrieval. This layer is implemented with the support of JSP (JavaServer Pages). JSP is a technology that helps software developers create dynamically generated web pages based on HTML, XML, or other document types. JSP is similar to PHP and ASP, but it uses the Java programming language.

D. Implementation and Results

A rudimentary prototype of the system has been implemented for daily news items for the US. A set of categories for the news items have been compiled such as political, business, sports, natural disasters, etc., and also subcategories for several of the top-level categories (for example, the sub categories for natural disasters are floods, earthquakes, hurricanes, etc.). For each of the categories and subcategories, an icon with an appropriate and intuitive visual effect is assigned.

There are two modes in the prototype – "overview mode" and "news–chain mode". In the overview mode, the user selects a given date, and the major news items on that given day (as captured by the system from online news sources) are displayed as icons on the map on the screen, where an icon is associated with a news category. When the mouse hovers over an icon, a text box with the title and the first few lines of the news item appears alongside for the user to peruse, and disappears when the mouse moves away from the icon. In the news chain mode, the user is given a choice of news items which are stored in the database. After user selection, the chain of events pertaining to that particular news item is displayed as a sequence of text boxes along a (scrollable) timeline.

The broad set of steps for news visualization in the overview mode is as follows:

1) Capture news from online news source(s)
2) Parse the headline and extract the title, the initial paragraph, keywords, location, date/time, and the URL
3) Assign a news category and the corresponding icon to each of the news items
4) Determine the location on the map on the screen (display) corresponding to the location of the news
5) Display the icon corresponding to the news item at the determined location on the map (For more than one item at the same location, a slight perturbation if the coordinates is performed before display)

6) When the mouse is moved over the icon, display the corresponding textbox containing the headline and first few lines of text. (the detailed news can be obtained by clicking on the provided URL)

A screenshot of the news visualization in the overview mode is shown in Fig. 6, on a particular date selected by the user. Icons corresponding to different news categories are displayed in corresponding locations on the map where the events have occurred. When the mouse is on an icon, a textbox with the news headline and the first few lines of news are displayed as shown.



Figure 6. Screen-shot of the proposed news visualization system (Overview mode)

News items pertaining to a specific topic over days (or weeks) are linked together, called "news–chain". The broad set of steps for news visualization in the news–chain mode is as follows:

1) Display the news titles in the database
2) Based on user query, retrieve all the news items related (relevant) to the query
3) Sort the news items chronologically
4) Display icons with dates horizontally on a 'time line'
5) When the mouse is moved over the date, display the corresponding textbox containing the headline and first few lines of text. (the detailed news can be obtained by clicking on the provided URL)

In this mode, scrolling facility is provided to scroll left/right suitably. A screenshot of the news visualization in the News–chain mode is shown in Fig. 7, for a news item (story) selected by the user.



Figure 7. Screen-shot of a "news chain" in the proposed news visualization system (News–chain mode)

V. FUTURE DIRECTIONS

Some of the future directions of this project include:

1) Extension of the news visualization to other major countries (currently it is implemented for the US).
2) Provision of the news visualization interface on mobile devices (tablets and smartphones).
3) Automated identification and selection of other news items related to a given news item.
4) Development of enhanced user interface for a query–system for both the overview and news–chain modes. An example in the overview mode is to provide provisions to support queries for a combination of location–time–category.
5) Development of a "Telescoping" mechanism for news visualization which facilitates navigation starting from a high-level of headlines and summaries (course highlights) to a highly detailed level.

ACKNOWLEDGMENT

I wish to thank Gruha Lakshmi Veera and Lavanya Kumari, MSCS students at National University for their diligence in the implementation of the early prototype of the system.

CONCLUSION

The rapid advancements in computing, communications, consumer electronics, and mobile technologies have resulted in the tremendous growths in the generation, storage, transmission, exchange, and consumption of digital content. With respect to news content, there has been a steep surge in the generation and dissemination of both conventional and user–generated digital news content on the Web and on mobile platforms. Recent trends have shown the decreasing readership of conventional media and increased readership of digital news medium. Due to busy schedules, more and more people are turning to getting news updates on the Internet and on mobile devices. The above trends have necessitated the need for facilitating a quick glimpse of news, and options to follow in more detail the news items of particular interests.

This paper proposed a system for visualizing the news in the form of icons on a map–the graphic of icon reflecting the news category. Pictures/graphics/icons have been known to convey information in a limited time. A rudimentary prototype of the system has been implemented and has shown a proof of concept to fill the need for a quick glimpse of news in time-constrained situations with the use of icons, headlines, and a few sentences of the news story, placed at the appropriate positions on a map on the screen. It also provides a feature of coalescing news stories over time and to watch their development. The paper has also identified several future directions of this project.

REFERENCES

[1] Newsmap.jp.
[2] http://www.naa.org/Trends-and-Numbers/Newspaper-Websites/Newspaper-Web-Audience.aspx
[3] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, H. Pfister, "What Makes a Visualization

Memorable?", IEEE Transactions on Visualization & Computer Graphics, vol.19, no. 12, pp. 2306-2315, Dec. 2013.
[4] M. Bruls, K. Huizing, van Wijk, J. Jarke. "Squarified treemaps". In de Leeuw, W.; van Liere, R. Data Visualization 2000: Proc. Joint Eurographics and IEEE TCVG Symp. on Visualization, Springer-Verlag. pp. 33–42, 2000.
[5] J. Lankow., J. Ritchie, R. Crooks. Infographics: The Power of Visual Storytelling, New Jersey: Wiley, 2012.
[6] B. Shneiderman. Tree visualization with tree-maps: a 2d space-filling approach. ACM Transactions on Graphics, 11(1):92–99, Jan 1992.
[7] S.R. Subramanya and B.K. Yi, "Mobile Content Provisioning – Major Issues", Wireless World Research Forum, Paris, December, 2005.
[8] S.R. Subramanya and B.K. Yi, 'User Interfaces for Mobile Content', IEEE Computer, Vol.37, No. 4, April 2006, pp101–103.
[9] (AT1) A. J. Torget and J. Christensen, "Building New Windows into Digitized Newspapers", Journal of Digital Humanities, Vol. 1, No. 3 Summer 2012.
[10] (AT2) A. J. Torget and J. Christensen, "Mapping Texts: Visualizing American Historical Newspapers", Journal of Digital Humanities, Vol. 1, No. 3 Summer 2012.

★★★