Summarizing Textual Information about Locations In a Geo-Spatial Information Display System

Congxing Cai
Information Sciences Institute
University of Southern California
Marina del Rey, California, USA 90292
ccai@isi.edu

Eduard Hovy
Information Sciences Institute
University of Southern California
Marina del Rey, California, USA 90292
hovy@isi.edu

Abstract

This demo describes the summarization of textual material about locations in the context of a geo-spatial information display system. When the amount of associated textual data is large, it is organized and summarized before display. A hierarchical summarization framework, conditioned on the small space available for display, has been fully implemented. Snapshots of the system, with narrative descriptions, demonstrate our results.

1 Introduction

Geospatial display systems are increasingly gaining attention, given the large amounts of geospatial data and services available online. Although geospatial imagery and maps show geometric relations among entities, they cannot be used to present other kinds of knowledge about the temporal, topic, and other conceptual relations and entities. Given an entity on a map, a description of what happened there, in what order in time, when, and why, requires additional types of information, typically contained in text, in order to support varied search and decision tasks.

In this demo, we apply text summarization to a geo-spatial information display system with potentially large amounts of textual data. By summarizing the textual material linked to each location, we demonstrate the ways one can organize this material for optimal display and search.

Of the many different types of text-oriented resources available, some are structured and others unstructured. This textual data can be linked to locations based on different reasons (containing place names, addresses, real objects with geographical features, etc.). Appropriately grouping and presenting the different aspects of the textual information in summarization is a challenging task.

A second challenge stems from the huge amounts of web material related to some geographical objects. For example, one may find millions of pages for a famous place or event at a specific map location. Given the common limitations of display space in most geospatial display systems, one must also design the interface to support dynamic browsing and search.

All these challenges bring new problems to existing summarization techniques. In the following sections, we demonstrate a hierarchical summarization framework that reduces displayed text and fully utilizes the small display space available for textual information.

2 Related Work

Associating each news page individually to its location(s) may overwhelm the amount of information displayable at any point and thereby limit the scalability of the system. Existing systems presented in (Teitler et al., 2008) and GeoTracker (Chen et al, 2007) organize material (at the area level) by time instead of somehow aggregating over larger numbers of related content. Since frequently the associated news contents overlap at least in part, a natural solution is to aggregate the content somehow to remove duplication. Moreover, the aggregation of news provides a global view of the textual information about the specific

Proceedings of the NAACL HLT 2010: Demonstration Session, pages 5–8,
Los Angeles, California, June 2010.

location. Our system is the first available geospatial text aggregation system to our knowledge.

Within geospatial display systems, the space available to display textual information is often quite limited. We therefore need to summarize the most important and relevant information about each location, drawing from all the web pages linked to it. However, directly applying a multi-document summarization (Lin and Hovy, 2001) to the web pages will generate poor results, due to unrelated titles, duplicate articles, and noisy contents contained in web pages. When several different events have occurred at a location, more than one distinct summary may be needed. It is therefore important to deploy topic recognition (Lin and Hovy, 2000) and/or topic clustering (Osinski and Weiss, 2005) to identify and group relevant pieces of each text into single-topic 'chunks'. We develop a novel hierarchical summarization system to improve the interactivity and browsability.

3 Text Summarization

3.1 Content Extraction and Summarization

Multi-webpage summarization is different from traditional multi-doc summarization. First, most web pages are much more complex than pure text documents. Since the web contains a combination of types of information—static text, image, videos, dynamic layout, etc.—even a single page can be treated as multiple documents. Current linking functions are based on keywords, making the relevant content of each relevant web page only a limited block within the page. Second, our task is oriented to locations, and hence differs from general content summarization. Hence, we need to identify and extract the essential part(s) of the webpage linked to the geospatial imagery for summarization and display. In our work, we utilize two important features, layout and semantics, to identify and extract the relevant content.

By rendering each web page into a DOM tree, we segment the page into large blocks based on its layout, including header, footer, left bar, right bar, main block, etc. We implemented a rule-based extractor to extract the most relevant block from the web page based on the relevance to the location.

3.2 Clustering

Given a list of text blocks relevant to a local point of interest, one can employ traditional text summarization techniques to produce a short summary for each one. This solution may not be helpful, however, since a long list of pages associated with each point of interest would be very hard for users to browse. Especially when the space allocated to text display by the geospatial system is also limited, a high compression ratio is typically required for the summarization system.

The solution we adopt is to deploy cluster-based multi-document summarization. Clustering must observe two criteria: first, the location of interest, and second, the text topic. Different clustering methods can be employed. To delimit topics, a simple heuristic is to introduce as additional criterion the event/article date: when the difference in document dates within a topical cluster is (far) larger than the actual duration of the topic event, we are probably dealing with multiple separate events at the same location. Better performance is obtained by using a topic detection module first, and then clustering documents based on the topics identified.

Unfortunately, documents usually contain multiple locations and multiple topics. The problem of 'topic drift' can cause confusion in a short summary. As in (Hearst, 1997), we segment each document into several 'mini-documents', each one devoted to a single topic, and then to perform location- and topic-based clustering over the (now larger) set of mini-documents.

3.3 Hierarchical Summary Generation

Whatever the clustering approach, the result is a potentially rather large set of individual topics associated with each location. Since screen space for the summaries may be very limited next to the maps / imagery, they have to be formatted and presented for maximal interpretability. To address this problem, we adopt a hierarchical structure to display incrementally longer summaries for each location of interest. At present we have found three levels of incrementally longer summaries to be most useful.

Thumbnail: a very short 'topic' that characterizes the (clusters of) documents or segments associated with each location. We present essentially one or two single keywords -- the most informative

words for each cluster. We implemented a new version of our topic signature technology, one that uses tf.idf instead of the entropy ratio, as scoring measure to rank each cluster's words.

Title: a headline-length phrase or short sentence (or two). The original titles of the web pages are often noisy or even unrelated to the current topic cluster. Sometimes, the title may be meaningless (it might for example contain the website's name "Pr Newswire"), or two different web pages may share the same title. We implemented a topic-related headline generator based on our previous work (Lin and Hovy, 2000) by incorporating a topic-based selector.

Snippet: a paragraph-length excerpt characterizing the cluster. To produce paragraph-length summaries, we implemented an extraction-based text summarizer. We built a new version of previously investigated technology (Lin and Hovy, 2001), implementing several sentence scoring techniques and a score combination function.

4 Demonstration

4.1 Geospatial Interaction

The hierarchical summarization service is built upon the geo-spatial information display system, GeoXRAY[1], a commercial product developed by Geosemble Technologies[2]. Figure 1 shows the system's display to support search and browsing of text content based on location of interest.

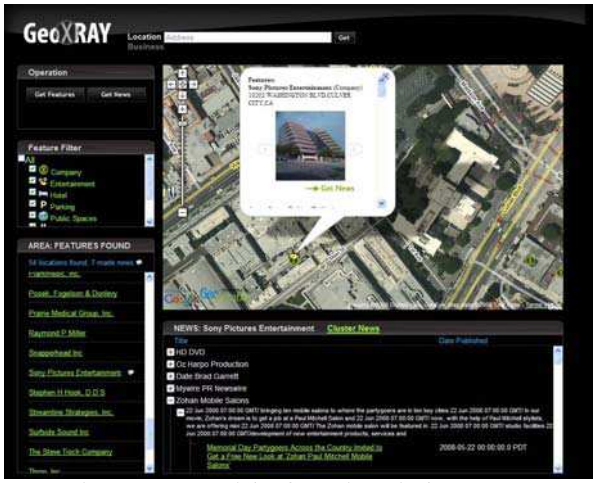

Figure 1. Geospatial Information Display System

[1]GeoXRAY: http://www.geosemble.com/products_geoxray.html
[2]Geosemble Technologies: http://www.geosemble.com/

The user can enter an address in the top search box, or search by business name. The system then centers the imagery at that address or business. Clicking on "Get Features" invokes the web services to get all features about the displayed image and displays the features in the "AREA: Features Found" list, and also draws them as points on the maps.

The user can explore the map using the navigation controller. On clicking the marker of an identified building, an information window pops up containing the associated structured web information (building name, business type, website, online images, and so on), as shown in Figure 2.

Figure 2. Navigating the Integrated Map

Clicking on "Get News" retrieves all news related to the displayed features; features with associated news show a small newspaper icon (see next to "Sony Pictures Entertainment" in Figure 4). Clicking on the icon displays the news that was linked with the feature, sorted by date.

The hierarchical summarization system, described in this paper extends the GeoXRAY system to show a summarized view of the news. The user can click on the "Cluster News" link. The results are displayed in a tree, showing the title of the cluster (thumbnail and title), under which appears a small summary of the cluster, under which appear links to all the news articles belonging to that cluster.

4.2 Summarization Example

We provide an example of our text summarization system performance in Figure 3. In this example, we have selected the location of Sony Film Studios in Culver City by clicking on the map. Figure 3(a) shows the titles and dates of some of

the 126 news articles that contain the words "Sony Pictures Entertainment". As described above, these documents are clustered based on topics. Using our current parameter settings, 20 multi-result clusters are formed, leaving 34 results unclustered. (The size of clusters, or the number of clusters desired, can be varied by the user.) As mentioned above, each cluster is presented to the users by a minimal length thumbnail summary consisting of a few characteristic keywords; a partial list of these is shown in Figure 3(b). Figure 3(c) shows the result of selecting the cluster labeled "solar electrical system" (second from the bottom in Figure 3(b)), which contains two results. The summary contains the 5 top-ranked sentences from the two documents, presented in document order. In addition, the summary includes two hyperlinks to the two full texts for further inspection.

(a) Partial list of the news articles linked to Sony Pictures Entertainment

(b) Clustering results relevant to Sony Pictures Entertainment

(c) Summarization from the news articles in cluster Solar electricity system

Figure 3. Document clustering and summarization for news relevant to Sony Picture Entertainment

The summary illustrates some of the strengths but also the shortcomings of the current system. It is clearly about a solar energy system installed in 2007 on top of the Jimmy Stewart Building by EI Solutions. This is enough detail for a user to determine whether or not to read the texts any further. However, two of the extracted sentences are not satisfactory: sentence 2 is broken off and sentence 3 should not be part of the news text at all. Premature sentence breaks result from inadequate punctuation and line break processing, which is still a research problem exacerbated by the complexity of web pages.

By showing the summary results, we merely demonstrate the improvement on browsability of the search system. We are relatively satisfied with the results. While the summaries are not always very good, they are uniformly understandable and completely adequate to prove that one can combine geospatial information access and text summarization in a usable and coherent manner.

Acknowledgments

Thanks to Geosemble Technologies for providing support of the geospatial information system.

References

Yih-Farn Robin Chen, Giuseppe Di Fabbrizio, David Gibbon, Serban Jora, Bernard Renger and Bin Wei. Geotracker: Geospatial and temporal rss navigation. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, 2007.

Marti A. Hearst. TexTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics*, 2000.

Chin-Yew Lin and Eduard Hovy. From single to multi-document summarization: A prototype system and its evaluation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2001.

Stanislaw Osinski and Dawid Weiss. Carrot2: Design of a flexible and efficient web information retrieval framework. In *AWIC*, 2005.

Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet and Jon Sperling. Newsstand: a new view on news. In *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, 2008.