Developing a Knowledge-base to Improve Interaction with Collections of Historical Newspapers

Robert B. Allen
College of Information Science and Technology
Drexel University
Philadelphia, PA, USA
E-mail: rba@boballen.info

Meeting: **188 — *Newspapers in the Caribbean, Central and South America: production, distribution and conservation. Cultural heritage and news media in the digital age* — Newspapers Section**

Abstract:

Large quantities of historical newspapers have been scanned and OCR'd. We can now consider how those files should be indexed and presented to be useful for laypersons, scholars, and students. In this paper, we explore the extraction and indexing of names from the Washington Times for March 1904 using two Java programs. We then consider how those names can enrich a history knowledge-base we are developing. In turn, these knowledge-bases should enable us to improve retrieval from garbled or ambiguous passages in the text. Ultimately, the knowledge-base should also be useful for exploring history based on community models.

1. INTRODUCTION

Through projects such as the U.S. NEH/LC National Digital Newspaper Program (NDNP) a large amount of newspaper images and text is now available online. While basic search interfaces are already available, many advances interfaces features could be developed to enable users to effectively take advantage of this great volume of material. Such advanced interfaces may be based on graphical interaction coupled with rich indexing. With more than million pages of historical newspapers any processing of such a large corpus is a huge undertaking and robustness and scalability are significant considerations.

In Section 2, we discuss different approaches to interacting with collections of digitized historical newspapers. Section 3 discusses the knowledge-base that was built with information on historical government structure and officials and based upon the *Washington Times* of March 1904. We also describe a program to extract named entities from the OCR'd text. In the Conclusion, we describe how the knowledge-base may be extended and explore broader conceptual issues.

2. INTERACTING WITH COLLECTIONS OF HISTORICAL NEWSPAPERS

Newspapers are such complex information objects that supporting access and indexing can be much more complex than indexing a collection of homogeneous objects. A sophisticated user interface would enable users to explore people, places, and topics, mentioned in the article. For instance, a list of related stories would link to news stories related to the current one but there are many types of links. The relatedness of stories in newspapers is an amorphous concept. The stories may simply be part of a regular feature. Where there is continuity in events, the thread may be due to an explicit or implicit agenda. Events such as the sessions of Congress, sporting events, anniversaries, and planned travel are explicit. Other events have a partially predictable structure. For instance, a crime, arrest, and trial have a likely but not entirely predetermined sequence. These other event threads may be more like "scripts", which are expectations about sequences of events (Schank & Abelson, 1977). To help the user to understand the relationship of events, the interface could employ a focus-context view (e.g., Allen, 2006). Finally, a Control Panel might enable the user to add annotations, conduct new searches, or move to other newspapers.

Different types of users are likely to access the digitized historical newspapers and variations of the interface might be adapted to them. There may be laypersons interested in local history as well as genealogists, students, and historians. Through the interface, casual users might be essentially tutored in the history related to the articles while historians would have a broad range of options for exploring information on their own. Allen and Sieczkiewicz (2010), interviewed historians and explored features which might be useful in a "historian's workbench".

3. THE KNOWLEDGE-BASE

An effective user interface needs to be supported with richly indexed content. In this section, we describe an initial version of a knowledge-based of named entities. This could help in the disambiguation of names and, ultimately, in the identification of article topics. Eventually, we seek to go beyond name-lists to more complex models of people and their relationship to events. For example, when considering individuals mentioned in a newspaper, it may be helpful to know additional facts about them and then to develop a broad of them and they community context.

3.1 Historical Government Knowledge-base: Structure and Access

Newspapers, especially those from Washington, D.C., mention government officials frequently. Since the government is highly structured, references to government officials are readily captured in a knowledge-base. A knowledge-base was developed for the structure of the U.S. Federal Government. The roles associated with that structure, and the individuals who occupied those roles. The initial program was written in Java. A version of the program built on a MySQL database is under development. This knowledge-base is an example of the value of linked data applied to digital history.

The U.S. government's structure is mostly hierarchical which is easy to model; however, it changes over time. At the founding of the U.S. government, the Executive Branch included four Cabinet Departments. One hundred years later in 1889, there were 7 Departments and in 1989 there were 14. Moreover, there were reorganizations of the Cabinet and at sub-cabinet levels there were a great many reorganizations. Therefore, a flexible framework is needed for describing governmental structure. This dynamic structure was captured by creating a date range for each agent or Role. Individuals were then associated with those Roles. For instance, Abraham Lincoln was identified first as a Congressman and then as President. Additional biographical information is also included in the knowledge-base about each individual.

The names and relevant facts associated with the names in the current implementation of the knowledge-base were extracted from multiple databases and websites. The initial version of the program was loaded with the names about 1000 US government officials from the founding of the republic to the present. This list includes officials of each of the three branches of the government. Obtaining names beyond the initial set from on-line resources is more difficult. Moreover, names of state and local officials are not readily captured. Presumably, even if these data are unavailable online they should be available in official records, but those will be even more difficult to access for smaller locales. Similarly, records about religious and business leaders will be difficult to identify.

3.2 Populating the Knowledge-base: Extracting Personal Names from the *Washington Times* of March 1904

Named-entity extraction is by now a widely used text-mining technology. The historical newspapers contain many named entities that would be useful to extract and index. Indeed, Crane and Jones (2006) noted the wide variety of ways that named-entities were referred to in a 19th century newspaper and they discussed the complications that variety raises. However, they did not attempt characterizing a corpus of OCR'd news text as we do here. Moreover, many collections of historical newspapers have been processed with OCR and the quality of the resulting text is very uneven. Traditional named-entity extraction programs are based on determining grammatical structure and work poorly on text with errors.

We developed a relatively robust approach by looking for specific terms and focused on enumerating names of specific individuals. A name-mining Java program was developed that was seeded with (a) a list of 3000 common first and last names, (b) a list of role titles (e.g., Cardinal, Mayor, General), and (c) a list of place names. However, some ambiguous cases such as names with other meanings (e.g., Baker) were dropped. We used this program to explore OCR'd text from the *Washington Times*[1] for March 1904. There was ambiguity in the results of the automatic processing of names because of issues such as (a) "Father John" may or may not be the same person as "Father John Hurley" and (b) place names could also be personal names. Compounded by the large number of OCR errors in the extracted text, there may be as many as 30% errors in this list. Nonetheless, the results still suggest important possibilities for indexing the newspapers.

- More than 20 different Admirals and 10 different Bishops are mentioned.
- There are several references to historical figures such as Aaron Burr, Abraham Lincoln, and Andrew Jackson.
- 79 Senators were mentioned. There would have been 90 actively serving Senators at that time though some former Senators may also have been mentioned. In addition, there was some ambiguity with both a Senator Fry and Senator Frye appearing. (Authoritative sources suggest that Frye is correct.)
- There were about 50 baronet or knights (all identified as "Sir") referenced so the knowledge-bases might need to incorporate resources such as *Burke's Peerage and Gentry* and *Shaw's List of Knights.*
- Overall, more than 5000 distinct names were identified. The majority of them appeared without titles.

The names were indexed at the page-level of repetition. Although a name may appear several times on a page, it is counted as appearing only once. Thus, our counts of cumulative frequencies are actually counts of pages on which the name appears. These showed several interesting effects:

[1] This publication is not related to current newspaper called *The Washington Times*.

- After President Roosevelt, the most common name is French Simpson. Mr. Simpson is listed as an agent of the newspaper who received classified notices to be published.
- Peter Grogan was also a common name but that often appeared in the context of the "Peter Grogan Company Building". Presumably, a more elaborate program would have better distinguished this as a location rather than as a person.
- Another very common name was President Smith that referred to the President of the Church of Jesus Christ of the Latter Day Saints (Mormons). President Smith was active in the discussions about the admission of Utah at a State which were happening in this time period.

4. CONCLUSION

We have described the development of two Java programs. The first is a knowledge-base which organizes the names and roles of individuals with the U.S. Federal Government agencies. Having such a list of names should be useful in processing text from newspapers in which those names are mentioned. A second program extracts names from OCR'd newspaper text. These two programs could bootstrap each other to improve the scope and accuracy of name recognition from the newspaper texts. Potentially, the knowledge-base could then also improve the effectiveness of the interface for users to explore the newspaper articles.

4.1 Extending the Current Approaches for Building the Knowledge-base

Although we observed a high error rate, we have demonstrated several promising techniques for developing the knowledge-base. This proof-of-concept could be extended several ways.

For article-level tagging, we need better techniques for the segmentation and categorization of the articles, and a structure description for the newspapers to help restrict the indexing. The International Press and Telecommunications Council (IPTC, http://www/iptc.org) developed a categorization system for newspaper content. Allen et al., (2008) have developed the automatic assignment of categories to articles and Allen and Hall (2010) have explored the identification of newspaper sections. The identification of names itself should be helpful in article categorization and sections. For instance, it would be useful to know that the Peter Grogan mentioned above appeared only in the newspaper's masthead.

Thus far, we have primed the knowledge-base extraction program with only a modest amount of knowledge. Historical societies, archives, and records departments have extensive collections with materials that should be useful for populating the knowledge-base. Moreover, there is an increasing number of relevant, but stand-alone, databases for areas such as genealogy and building permits that could be tied in. Even records of historical baseball games will contribute to processing the sports pages.

Temporal order can also provide a useful constraint. Allen et al. (2008) showed that the frequency of words and names followed predictable temporal patterns. For instance, the term "drought" appeared mostly in the summer and the appearance of (Theodore) Roosevelt's name greatly increased in frequency after McKinley's assassination and Roosevelt became president.

Allen (2010) identified some significant events by coordinating across multiple newspapers with the same dates. That could also apply to identification of individuals. Moreover, coordinating across multiple newspapers enables incorporation of the perspectives of those different newspapers, which can be especially useful in providing a complete picture of a community. Notably, minority communities many not be well represented in the mainstream media.

4.2 Broader Conceptual Issues

The analysis of names that appear in newspapers and other historical resources would facilitate the development and understanding of historical social networks. For instance, we could determine for Washington, D.C. in 1904 who is mentioned with whom. While such social networks would provide additional constraints for processing the news articles, they are unlike the social networks we use today in social media.

Developing models of historical events should facilitate the indexing of newspapers along with many other historical documents. In addition, capturing and describing events might provide the basis for entries in timelines. Allen (2011) provides some suggestions and an example of how such timelines could be developed. More speculatively, user interaction may be mediated by computational conversational historical agents whose knowledge is based on an event-oriented knowledge-base. Indeed, indexing processes based on events could provide a new dimension to indexing many sorts of materials (e.g., oral histories) beyond simple taxonomic indexing.

As more detailed information is collected in the knowledge-base it will be increasingly helpful to organize the data into unified conceptual models. This may be done much more richly than was the case for the government knowledge-base described above. Presumably, some basic processes underlay many of the items that appear in a newspaper (Figure 1). Some of these are implicit social rules or even laws that would be familiar to individuals from that community but may not be apparent to casual browsers. We could try to summaries and organize these expectations as "community models" (Allen et al., 2007) (cf., Shopes, 1998) and successfully extracting and using them could greatly improve indexing, access, and derived historical summaries. Finally, a community annotation component could be applied to developing the historical model. This is similar to the suggestion of Anderson and Allen (2009) that members of the public could add annotations to the knowledge-base.

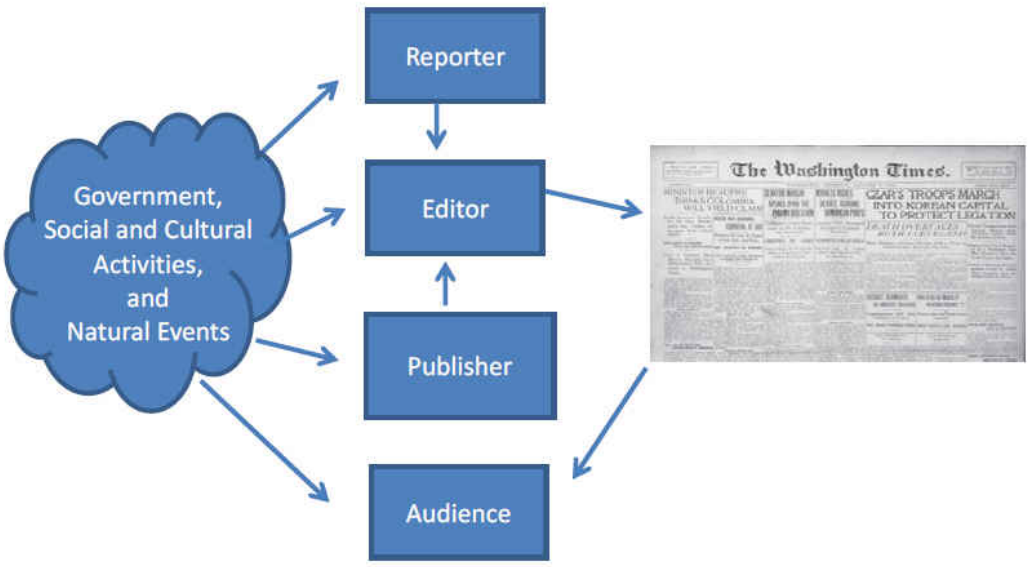

Figure 1: A rough schematic of the relationship between what happens in a community and what appears in the newspaper.

REFERENCES:

Allen, R.B., A Focus-Context Timeline for Browsing Historical Newspapers. *ACM/IEEE Joint Conference on Digital Libraries* (JCDL), 2005, 260-261.

Allen, R.B., Improving Access to Digitized Historical Newspapers with Text Mining, Coordinated Models, and Formative User Interface Design, *IFLA International Newspaper Conference*: *Digital Preservation and Access to News and Views,* 2010, 54-59.

Allen, R.B., Visualization, Causation, and History, *IConference*, 2011, DOI: 10.1145/1940761.1940835

Allen, R.B., & Hall, C. Automated Processing of Digitized Historical Newspapers beyond the Article Level: Finding Sections and Regular Features. *International Conference on Asian Digital Libraries (ICADL)*, 2010, 91-101. DOI: 10.1007/978-3-642-13654-2_1

Allen, R.B., Japzon, A., Achananuparp, P., & Lee, K-J., A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. *HCI International Conference*, 2007, DOI: 10.1007/978-3-540-73354-6_26

Allen, R.B., & Sieczkiewicz, R., How Historians use Historical Newspapers, *ASIST*, 2010.

Allen, R.B., Waldstein, I., & Zhu, W., Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres. *International Conference on Asian Digital Libraries*, Hanoi, Vietnam, 2008, 380-387, DOI: 10.1007/978-3-540-89533-6_49

Anderson, S., & Allen, R.B., Envisioning the Archival Commons, *American Archivist*, (2009) 72(2), 383-400.

Crane, G., & Jones, A., The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th Century Newspaper Collection, *ACM/IEEE Joint Conference on Digital Libraries* (JCDL), *2006*, DOI: 10.1145/1141753.1141759

Schank R C & Abelson R P., *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* Erlbaum, Hillsdale, NJ, 1977.

Shopes, L., Oral History and the Study of Communities: Problems, Paradoxes, and Possibilities. In: R. Perks and A. Thomson (Eds.), *The Oral History Reader* (2nd Ed.). Routledge, London, 1998, 261-270.