Papers

iConference 2010

FEBRUARY 3-6 • UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

What to Do with a Million Pages of Digitized Historical Newspapers?

Robert B. Allen
College of Info Sci & Tech
Drexel University
Philadelphia, PA, 19107
1-215-895-0460
rba@drexel.edu

Weizhong Zhu
College of Info Sci & Tech
Drexel University
Philadelphia, PA, 19107
1-215-668-4185
wz32@drexel.edu

Robert Sieczkiewicz
Hagerty Library and Drexel Archives
Drexel University
Philadelphia, PA
1-215- 895-1757
robs@drexel.edu

ABSTRACT
Newspapers are rich sources of evidence of history and literally millions of pages of historical newspapers have now been digitized. We aim to develop tools for effectively browsing this rich resource. However, the structure of newspapers is highly complex and a complete analysis will involve many interacting components. We demonstrate two approaches for extracting advertisements from other types of material in the newspaper. We also describe preliminary results of interviews with historians about features which they would find particularly useful for conducting research on collections of digitized historical newspapers.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries; H.5.2 User Interfaces

General Terms
Design, Human Factors

Keywords
Advertisements, History, Image Processing, Interface Design, Newspapers

1. DIGITAL HISTORY AND DIGITIZED HISTORICAL NEWSPAPERS
Along with several other areas of the humanities, information systems are greatly expanding the resources available at the desktop for historians. Literally millions of pages of historical newspapers have already been digitized and many millions more will be processed over the next few years.

While there are several digitized initiatives underway, we have worked with the files prepared by the NDNP (National Digital Newspaper Program). This is a joint project of LC and NEH which builds on the prior USNP (United States Newspaper Program). USNP had worked with state libraries to prepare archival quality microfilm of historical newspapers from their sates and that microfilm is now being digitized by NDNP. A notable aspect of the NDNP digitization is that public domain OCR files are delivered in the METS-ALTO format. In almost all other newspaper digitized projects the OCR is proprietary or otherwise unavailable. However while the OCR is available in NDNP, it is often of marginal quality and extensive text processing is necessary.

Figure 1: Sample page of the *Washington Times* for March 1904.

2. TEXT PROCESSING FOR EXTRACTING ADVERTISEMENTS
There are many approaches for text processing of the historical newspaper text. [2]. We have demonstrated extraction and categorization of articles is [3]. While those results are uneven, there are many additional constraints which can be considered. One technique which was explored by Allen and Hall [4] was to focus the first lines of text in a segment to find items which were repeated across days. This was used to identify features stories but it was also noted that after extracting the longer feature articles that many of the shorter items with repeated headings were advertisements. This is sensible since advertisements often run for many days while other items change from day to day.

Here, we study this observation in more detail. Specifically, we applied the article segmentation procedure described in [2] to OCR output for the *Washington Times* for March 1904. We then processed that to find repeated first lines of the text

segments. Though, these had to be heavily processed because of the large number of OCR errors.

J WILLIAM LEE

E M EARLE SON

THE STORE THAT SAVES YOU MONEY

NATIONAL BISCUIT COMPANY

ADVANCE SPRING STYLES

Table 1: Examples of repeated headings indicating advertisements.

We found three categories among the repeated items: feature articles, headings for advertisements, and headings for other special sections such as Vital Records. Most of the advertisements appeared on the right and bottom edges of the newspaper (see Figure 1). Moreover, they all had non-standard fonts which probably could have been helpful for identifying them.

While this technique does not capture all of the advertisements, many of them are clustered there. Furthermore, we can infer that other advertisements are located in the same region even if they don't have repeated first lines. However, as we noted in our earlier work [3, 4], there needs to be a design tradeoff between the complexity of automated inference and the amount of human knowledge to include but the simple rules of thumb described here caught the large majority of advertisements.

3. AUTOMATIC ANALYSIS OF IMAGE GENRES

As a second, very different type of analysis for detecting advertisements is to determine image genres[1]. Images with words are most often banners for advertisements on the other hand, portraits may accompany news stories. In addition to the news codes for text used by [3], the International Press and Telecommunication Council (IPTC) also specifies image genre types[2]. A complete system for processing newspaper images would itself be a major project. The goal here is not to extract the images from the page but to categorize then by the IPTC genre codes after they have been extracted.

3.1 Image Feature Selection and Representation

There are several approaches to image analysis. One common strategy looks for specific features. We applied Eigenvector-based feature selection [5] and traditional retrieval/classification methods to differentiate images from the different genres. Models of individual features try to identify position, size, color, texture and relationships between these features but these models are insufficient to classify the multiple views such as multiple faces in one image, compared to the models of gestalt/pattern features. Because the task includes complex images such as images with multiple persons, Eigenvector based feature modeling seems to be more suitable for our project. The Eigenvector based feature indexing was originally developed for face recognition and includes the following steps:

- Take the M training image vectors and average them to find $\Psi = 1/M * \sum_{n = 1 \text{ to } M} \Gamma_n$ where Γ_n is the n^{th} image vector. Each image differs from the average by $\Phi_i = \Gamma_i - \Psi$.
- These are the eigenvectors of the Covariance Matrix: $C = 1/M * \sum_{n = 1 \text{ to } M} \Phi_n \Phi_n^T = AA^T$, where $A = [\Phi_1 \ldots \Phi_M]$. If $M << N^2$ then there will only be $M - 1$ eigenvectors which have non-zero Eigenvalues. So can solve an M x M matrix instead. Consider the eigenvectors, $\mathbf{v}_i$, of A^TA such that: $A^TAv_i = \mu_i v_i$ which yields $AA^TAv_i = \mu_i Av_i$ where Av_i are the eigenvectors of $C = AA^T$ (and μ_i are the Eigenvalues).
- To form the Eigenspaces u_i use the following equation: $u_i = \sum_{k = 1 \text{ to } M} v_{ik}\Phi_k$ for i = 1 ... M.

Each image in the Eigen indexing spaces is represented as a vector with at most M dimensions. If a new image is projected in the training Eigenspaces, the representation of that image is calculated by the following procedure:

- Given the set of M Eigenspaces, choose the M ' Eigenspaces that have the highest associated Eigenvalues.
- Take a new image, Γ, and project it into "Eigenspaces" by the operation: $W_k = u_k^T \cdot (\Gamma - \Psi)$ for k = 1 to M '.
- The weights (W_k) form a vector $\Omega^T = [W_1 \ldots W_k]$ which describes the contribution of each Eigen-space in representing the input face image.

3.2 Image Retrieval and Classification Methodology

To match the query image, the relevance between the query image and each of the training samples is defined by the Euclidian distance $\varepsilon_k = \|\Omega - \Omega_k\|$, where Ω_k is the weight vector describing the k^{th} training image. But to classify an image, the representation of the image is the vector Ω^T. The traditional classification methods include two types, unsupervised learning – clustering and supervised learning – machine learning. Two clustering algorithms: basic K-means [6] and Kohonen Mapping [7] and two machine-learning techniques, back-propagation [8] and simulated annealing [9] were tested.

3.3 Test Corpus of Images

A preliminary examination of the newspapers found that six of the IPTC image genres covered most of the cases. A research assistant obtained 60 newspaper images from the *Washington Times* for 1904 belonging to these six classes: Portrait (PRT), Text Characters (WORD), Exterior Views of Buildings (EV), Full Body (FB), Half Body (HL), and a Group of People (MP). There were more of the first three types than of the others. Examples of the images are shown in Figure 2. They were normalized to a fixed size before the analysis was conducted.

[1] The entire page scanned from the microfilm is also an image but here we are concerned with images extracted from within the page.

[2] http://www.iptc.org/cms/site/index.html;jsessionid=aiZ0zp rArDm8?channel=CH0089

	Kohonen	K-means
Purity	*0.55*	*0.42*
Entropy	*0.50*	*0.68*
NMI	*0.47*	*0.31*

Table 2: Comparison of clustering methods.

3.4 Clustering the Images

Two clustering techniques, Kohonen Mapping and basic K-means, are used to automatically classify the six categories of the images. In Table 2, three evaluation metrics, Purity, Entropy and normalized mutual information (NMI) are used to evaluate the performance of the two techniques. The values of the three measures range from 0.0 to 1.0. The higher the values of Purity and NMI are, the better the performance of the method is. The lower the value of Entropy is, the better the performance of the method is. The results in Table 3 indicate that for this dataset, Kohonen mapping out-performs basic K-means for every evaluation method.

Figure 2**: Examples of the six image genre classes (upper row: HL, WORD, EV, lower row: PTR, HL, MP)**

Clusters generated by the Kohonen Mapping are shown in 3. There are strong associations: Cluster 1 and PRT; Cluster 2 with MP; Cluster 3 with FB; Cluster 4 with ENV; and Cluster 5 with WORD; but Cluster 6 did not have a clear association and it might be deleted. Similarly, the half-body images might be merged with the full-body images.

3.5. Machine Learning of the Genre Categories

Two neural network-based machine learning methods, Back-propagation and Simulated Annealing were used to automatically classify the six categories of the images. Eighty percent of samples were randomly picked for training and the remaining 20% used for testing 10 or 20 times to calculate an average learning performance. In Table 4, PRT/HL/FB/MP/WORD/EV denotes that the training set includes six classes; PRT/HL/FB/MP denotes the four PERSON related classes; PERSON/WORD/EV denotes the three major categories, PERSON (PRT/HL/FB/MP), WORD and EV; PRT/WORD denotes the two classes, PRT and WORD. The results indicate that on this dataset, Back-propagation out-performs Simulated Annealing for every type of training set. Further, the two machine learning methods did better on the two training sets, PERSON/WORD/EV and PRT/WORD.

	Portrait	WORD	Environment	Full Body	Half Body	Multiple People
Cluster 1	*9	1	1	0	1	0
Cluster 2	0	0	1	0	0	*4
Cluster 3	4	0	1	*7	2	*4
Cluster 4	1	0	*6	1	2	0
Cluster 5	2	*9	0	0	1	0
Cluster 6	0	0	0	1	0	1

Table 3: Number of each type of image in the clusters. * indicates the best match.

Training Sets	Back-Propagation	Simulated Annealing
PRT/HL/FB/MP/WORD/EV	0.46	0.23
PRT/HL/FB/MP (Person)	0.33	0.27
PERSON/WORD/EV	0.64	0.40
PRT/WORD	0.75	0.63

Table 4: Comparison between Back-propagation and Simulated Annealing for categorization.

3.6. Image Retrieval with Query-by-Example

Image retrieval may use "query-by-example" compared to text retrieval. Our approach identifies the most relevant or similar images in the Eigen-indexing spaces to the query image by projecting it to the indexing spaces. In Table 5 one PRT image and one WOPRD image were randomly selected and then treated as queries and the top five most relevant images in the indexing spaces are listed. As can be seen, the query-by-example returns a high number of images from the same genre class.

WORD Image	***Similarity***
WORD_March_19_1904_Page9.jpg	*target*
FB_Jan_3_1904_Page3.jpg	*0.72*
WORD_March_13_1904_Page10.jpg	*0.63*
WORD_Jan_3_1904_Page12.jpg	*0.62*
WORD_March_18_1904_Page12.jpg	*0.58*
HL_Jan_18_1904_Page7.jpg	*0.57*
PRT Image	***Similarity***
PRT_March_3_1904_Page3.jpg	*target*
PRT_Jan__1904_Page2.jpg	*0.76*
PRT_Untitled.jpg	*0.76*
WORD_March_21_1904_Page5.jpg	*0.70*
MP_Jan_3_1904_Page8.jpg	*0.66*

Table 5: Two cases of query-by-example retrieval for images.

3.7. Image Genre Discussion and Conclusion

The clustering and machine learning methods categorize the three major categories, WORD/PERSON/VIEW effectively with Eigenspace indexing but differentiate the sub-categories of PERSON poorly. This is a small study, but it suggests that the image modeling of Eigenspace based features are good on image categorization with fewer classes, for instance, binary image categorization. For this data set, Kohonen mapping is a better clustering method than basic K-means, and Back-propagation is a better machine learning method than Simulated Annealing. In the future, a much larger collection will be tested for the proposed approach and integration with specific individual features and textual descriptions will be explored to improve the performance of the Eigenspace indexing model and automatic categorization.

4. INTERFACE REQUIREMENTS FOR SUPPORTING HISTORIAN'S ACCESS TO DIGITIZED NEWSPAPERS

There have been quite a few studies of the information needs of historians [10-13]. However, these studies do not provide clear guidance about what sort of interface tools historians would find most useful for interfaces with a collection of newspaper. The primary interface for searching the NDNP newspaper collection is Chronicling America[3] which is a basic search interface with few features.

There have been some ad hoc recommendations such as supporting searching for Facts, Trends, Searching for Details[4]. Tools could be based on the proposals of "Rachel"[5]

(1) Have a List,
(2) Find a thread of some kind,
(3) Don't just use one newspaper,
(4) Don't fall into the trap of only reading articles that your keywords throw up.

[3] http://chroniclingamerica.loc.gov/

[4] Rubenstein, A., Center for History and New Media, Unpacking Evidence, http://chnm.gmu.edu/worldhistorysources/unpacking/newshow.html (accessed November 2009)

[5] Rachel, A Historian's Craft: http://idlethink.wordpress.com/2009/06/16/on-newspapers-as-sources/ (accessed November 2009)

(5) Use existing secondary literature,
(6) Keep really, really scrupulous notes, and
(7) Don't neglect the letters and the advertisements.

This list suggests a number of potential features.

To explore the importance of these services by working historians in more detail, we conducted informal interviews with two historians. Here are some examples of notes from these interviews:

> The *Chicago Tribune* database is good for searching names, but broader topics are hard to research – e.g., race relations brings back too many results.
>
> A log of all searches – `this is a huge issue for me'. Editing a book manuscript recently, she found it `hugely taxing' to find items she hadn't cited.
>
> Searches lead to other searches, so she would like ways to see how searches are nested within each other and to get back to earlier search results. A visual map telling you where you are in your search would be especially helpful. A system that lets her easily use multiple windows.
>
> [The historian] used newspapers to fill in gaps in research and corroborate information from other sources. Exploratory searching included looking at larger issues and events such as elections and campaigns. She used newspapers to find public opinion about changes in liquor license laws – to get a sense of `the texture of the city… how the city was thinking'.

It is clear that the historians will benefit from an interface which would support richer ways of interacting with the collection than are currently supported by the Chronicling America web site. We are collecting more interviews and using those comments for an interface prototyping effort.

5. FUTURE OF ACCESS TO HISTORICAL NEWSPAPERS

There are a great many challenges to supporting effective access for the newspapers in NDNP collection. Yet, that is just the beginning. It will have to be scaled up to several fold to incorporate newspapers from around the world. Moreover, that can be scaled that much more again, to support links to other historical resources such as books, images, and manuscripts.

We are exploring techniques for coordinating among historical resources. As a first step, multiple newspapers from one town or region can show synergies which improve the text processing of each. In the same vein, other historical resources such as biographical and building databases can be cross referenced with the newspapers. Finally, we also envision novel tools for describing and interacting with "threads of history" and we are actively developing formalisms for describing chains and historical events and timeline interfaces for visualizing them (e.g., [14]).

6. ACKNOWLEDGMENTS

We thank faculty from the Drexel University Department of History & Politics for interviews. We also thank Mike Zarro for assistance. Weizhong Zhu completed his doctorate and is now employed at WebLib Inc.

7. REFERENCES

[1] Schudson, M., (1978). *Discovering the News: A Social History of American Newspapers.* Basic Books, New York.

[2] Allen, R.B., Japzon, A., Achananuparp,P., and Lee, K-J., (2007). A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. *HCI International Conference*.

[3] Allen, R.B. and Hall, C., (in preparation) Automated Processing of Digitized Historical Newspapers beyond the Article Level: Finding Sections and Regular Features.

[4] Allen, R.B., Waldstein, I., and Zhu, W., (2008). Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres. *ICADL*, 380-387.

[5] Turk, M.A. and Pentland, A.P., (1991) Face recognition using Eigenfaces," (1991) *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 586-591.

[6] Dubes, R.C., and Jain, A.K. (1988). *Algorithms for Clustering Data*, Prentice Hall, New York.

[7] Kohonen, T. (1990). The Self-organizing Map. *Proceedings of the IEEE 78*(9), 1464–1480.

[8] Rojas, R., "The Backpropagation Algorithm", (1996) *Neural Networks – A Systematic Introduction*, Springer-Verlag, Berlin, New York, ISBN 9783540605058.

[9] Kirkpatrick S., Gelatt C. D., and Vecchi M. P., (1983) "Optimization by simulated annealing", *Science*, vol. 220, 671–680.

[10] Tibbo, H.R., (2002). Primarily History: Historians and the search for primary source materials. *ACM/IEEE Joint Conference on Digital Libraries*, 1-10.

[11] Bradshaw, J. (1984). The Use of Newspapers in Historical Research, *Chronicle* (East Lansing) (04409426), 20(2), 18-19.

[12] Case, D. (1991). The Collection and Use of Information by Some American Historians: A Study of Motives and Methods. *Library Quarterly*, 61(1), 61-82.

[13] Dalton, M., and Charnigo, L. (2004). Historians and their Use of Information Sources. *College & Research Libraries*, 65(5), 400-25.

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