

Exploring Hidden Connections Among Historical Images

Wu Zheng

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, MC-493, Champaign, IL 61820-6211
wuzheng2@illinois.edu

ABSTRACT

In this paper, an experiment is presented to explore hidden connections among historical images using literature-based discovery (LBD) method. The result of the experiment shows that LBD has the potential in finding implicit relations among these images.

Keywords

Literature Based Discovery, Information Retrieval

1. INTRODUCTION

Nowadays, more and more historical resources (old documents, photos etc.) are digitized and made available online together with metadata created by librarians. In the description of the resources, usually there is background information about the object, for example, a short biography of a person, or the history of an organization. Among historical images, there are numerous implicit pairwise connections that cannot be found using traditional keyword searches. Two seemingly unrelated images, when put together, might reveal some interesting patterns that were overlooked before. The background information in the description provides clues to these connections. An experiment was conducted to explore the metadata of historical images using literature-based discovery methods. Several implicit relations among the images were identified. The result of the experiment shows that LBD can help users better navigate and use historical image records.

2. INTRODUCTION TO LITERATURE-BASED DISCOVERY (LBD)

Literature-based discovery, first proposed by Swanson (Swanson, 1986), is a method to systematically explore implicit knowledge in literatures. Compared to traditional information retrieval, LBD aim at complementarities instead of similarities. Two branches of LBD are widely studied in different fields: one-node search and two-node search. In one-node search, a specific topic is represented by a set of A-literature that talks about it. The goal is to find another (usually disjoint) set of C-literature that is relevant to the topic. For example, A-literature can be a set of articles that discusses a problem. In that case, C-literature can be a set of papers that contains information which might contribute to the solution of the problem (Smalheiser & Torvik, 2008). In two-node search, concepts that connect two disjoint sets of literature are sought.

Previous studies on LBD have shown that it is effective in exposing “undiscovered public knowledge” (Swanson, 1986). Several computer-assisted LBD systems (Hristovski, Friedman, Rindflesch, & Peterlin, 2008; Smalheiser & Torvik, 2008; Srinivasan & Libbus, 2004; Yetisgen-Yildiz & Pratt, 2006) have been developed in biomedical field to help researchers generate hypothesis by analyzing scientific literature. Cory (Cory, 1997) applied LBD method on online humanities databases to discover hidden analogies. Gordon et al (Gordon, Lindsay, & Fan, 2002) extended the information sources of LBD from scientific literatures to information available on the World Wide Web. Luo et al (Luo, Tang, & Tian, 2007) developed a system that can answer relationships queries by analyzing web information based on the idea of LBD. Literature searches did not find any study applying LBD method on images.

3. THE EXPERIMENT

3.1 Data

The data used for experiment was selected from the Institute of Museum and Library Services (IMLS) Digital Collection and Content (DCC) project (<http://imlsdcc.grainger.uiuc.edu/>). The metadata schema used in the project for resource description is the Dublin Core Metadata Element Set (<http://dublincore.org/documents/dces/>). A single collection, the *King County Snapshots* (<http://imlsdcc.grainger.uiuc.edu/collections/FullDisplay.asp?cid=2384>), was chosen because of the high quality metadata provided. This is a historical image collection with more than 12,000 items portraying people, places and events in the county.

3.2 Method

The experiment followed the one-node search method mentioned above. The information source used to explore the implicit connections is the title and description field of the metadata associated with the images. For a specific topic (person, place or event), a set of A-records was retrieved using keyword searches. The following steps were performed on each A-record:

1. A-term identification. For each record, the subject of the image (A-term) was extracted from the title field.
2. B-term identification. The title and description field of the record were processed using the Stanford Named Entity Recognizer

(<http://nlp.stanford.edu/ner/index.shtml>). All the named entities (person, organization and location names) captured were potential B-terms that connected the A-record with another set of C-records. C-records contained images that had implicit relations to the topic represented by A-records.

3. Searching for C-records. New searches were conducted for each B-term identified. For each record retrieved, if it was not in the set of A-records, and the subject of the image in it was not the B-term, the record was a C-record. The image in an A-record had possible implicit connection with the image in C-record.
4. Evaluation. Manual evaluations were performed to ensure that the implicit connections identified were meaningful.

Further analyses were performed on the connections identified.

3.3 Findings

Following the steps described above, numerous hidden connections were identified. Some typical examples are presented here.

Ladies' Aid/Grace Presbyterian Church/ Juanita, Eleanor, and Eldred Martin Lewis



Figure 1: Ladies' Aid posing outside Grace Presbyterian Church, Seattle, 1922

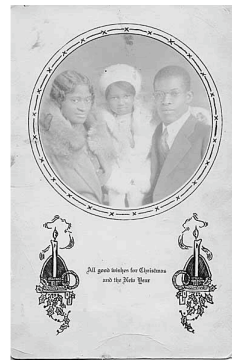


Figure 2: Juanita, Eleanor, and E. Martin Lewis, Seattle, ca. 1930

Two photos titled “Ladies' Aid posing outside Grace Presbyterian Church, Seattle, 1922” (Figure 1) and “Juanita, Eleanor, and E. Martin Lewis, Seattle, ca. 1930” (Figure 2) were connected through “Grace Presbyterian Church”. Ladies’ Aid was founded in 1912 for raising funds for the church. When the photo was taken (1922), Eldred Martin Lewis was the minister of the church. It appeared that Eldred received financial support from the organization and they might have other social connections. Juanita, Eldred’s wife, was also a famous lady (there are a lot of photos about her in this collection). It is possible that Juanita also had connections with the organization through her husband.

Anna Louise Strong / Industrial Workers of the World (IWW)/John Henry Smith



Figure 3: Anna Louise Strong, possibly in Seattle, 1939



Figure 4: John Henry Smith, Seattle, ca. 1921

Two photos titled “Anna Louise Strong, possibly in Seattle, 1939” (Figure 3) and “John Henry Smith, Seattle, ca. 1921” (Figure 4) were connected through “IWW”. In the description of the first photo, it was mentioned that Strong “was elected to the Seattle School Board but later was recalled because of her socialist politics and her support of the Industrial Workers of the World”. On the other hand, John “became active in I.W.W. circles about 1913”. Given that both of them were living in Seattle at that time and their support of the IWW, it is possible that there were social connections between them.

George Frye/Henry Yesler/James Colman

Two photos titled “George Frederick Frye, Seattle, ca. 1888” (Figure 5) and “James Murray Colman, Seattle, ca. 1885” (Figure 6) were connected through “Henry Yesler”. In the description of the first photo, it is mentioned that George Frye worked with Henry Yesler at his mill from 1853. Later Frye established his own hotel and became famous in the hotel industry. James Colman, as described in the record, operated Henry Frye’s mill from 1872. He established a railway company later. Frye and Colman were famous in different industries, but it is very likely that they had some social connections through their common (former) colleague Henry Yesler.



Figure 5: George Frederick Frye, Seattle, ca. 1888

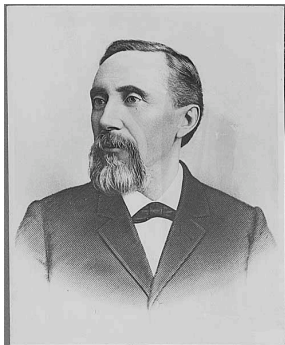


Figure 6: James Murray Colman, Seattle, ca. 1885

4. DISCUSSION

The large amount of connections identified in the experiment indicated that LBD has the potential in identifying implicit connections among images, and the metadata associated with the images are possible

source for exploration. In this experiment, a large proportion of the connections identified were social connections among people and organization. This is due to the coverage of the chosen collection (images of people, places and events) and the NER used to extract B-terms (identified person, organization and location names). One possible application is to explore the social networking among people covered in the collection. Applying the method on a different collection, across multiple collections or modifying the NER to capture other types of entities might identify other interesting connections. The method used in the experiment relied on the human-assigned metadata to explore the implicit connections, thus the performance of the method would be influenced by the quality of the metadata. For images with little or no description, especially background information on the subject, the connections are difficult to identified. The evaluation of the connections is another important issue. The relatedness in social science and humanities is very subjective, and depends heavily on the background and information need of the users. How to automatically evaluate and rank the connections will be focus of further research.

5. ACKNOWLEDGMENTS

The author is grateful to Vetle Torvik and Qin Wei for their review of this paper and the valuable suggestions.

6. REFERENCES

- Cory, K. (1997). Discovering hidden analogies in an online humanities database. *Computers and the Humanities*, 31(1), 1-12.
- Dublin Core Metadata Element Set. Retrieved December 5, 2008, from <http://dublincore.org/documents/dces/>.
- Gordon, M., Lindsay, R., & Fan, W. (2002). Literature-based discovery on the World Wide Web. *ACM Transactions on Internet Technology*, 2(4), 261-275.
- Hristovski, D., Friedman, C., Rindfleisch, T., & Peterlin, B. (2008). Literature-based knowledge discovery using natural language processing. In P. Bruza & M. Weeber (Eds.), *Literature-based Discovery* (pp. 133-152): Springer.
- Institute of Museum and Library Services Digital Collection and Content. Retrieved December 5, 2008, from <http://imlsdcc.granger.uiuc.edu/>.
- King county snapshots: a photographic heritage of Seattle and surrounding communities. Retrieved December 5, 2008, from <http://imlsdcc.granger.uiuc.edu/collections/FullDisplay.asp?cid=2384>.
- Luo, G., Tang, C., & Tian, Y. (2007). Answering relationship queries on the web, 16th

- International Conference on World Wide Web (pp. 561 - 570).
- Smalheiser, N., & Torvik, V. (2008). The place of literature-based discovery in contemporary scientific practice. In P. Bruza & M. Weeber (Eds.), *Literature-based Discovery* (pp. 13-22): Springer.
- Srinivasan, P., & Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20(Suppl. 1), i290-i296.
- Stanford Named Entity Recognizer. Retrieved December 5, 2008 from <http://nlp.stanford.edu/ner/index.shtml>.
- Swanson, D. (1986). Undiscovered public knowledge. *The Library Quarterly*, 56(22), 103-108.
- Yetisgen-Yildiz, M., & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, 39(6), 600-611.