Yuan Liu, Qiang Tan, and Kun Xu Shen. (1994). "The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing" (in Chinese). Qing Hua University Press and Guang Xi Science and Technology Press, page 36.

K.T. Lua. (1990). From Character to Word. An Application of Information Theory. *Computer Processing of Chinese and Oriental Languages,* Vol. 4, No. 4, pages 304--313, March.

Liang-Jyh Wang, Tzusheng Pei, Wei-Chuan Li, and Lih-Ching R. Huang. (1991). "A Parsing Method for Identifying Words in Mandarin Chinese Sentences." In *Processings of 12th International Joint Conference on Artificial Intelligence*, pages 1018--1023, Darling Harbour, Sydney, Australia, 24-30 August.

Automatic Link-Detection in Encoded Archival Descriptions

Junte Zhang
j.zhang@uva.nl
University of Amsterdam, The Netherlands

Khairun Nisa Fachry
k.n.fachry@uva.nl
University of Amsterdam, The Netherlands

Jaap Kamps
j.kamps@uva.nl
University of Amsterdam, The Netherlands

In this paper we investigate how currently emerging link detection methods can help enrich encoded archival descriptions. We discuss link detection methods in general, and evaluate the identification of names both within, and across, archival descriptions. Our initial experiments suggest that we can automatically detect occurrences of person names with high accuracy, both within (F-score of 0.9620) and across (F-score of 1) archival descriptions. This allows us to create (pseudo) encoded archival context descriptions that provide novel means of navigation, improving access to the vast amounts of archival data.

Introduction

Archival finding aids are complex multi-level descriptions of the paper trails of corporations, persons and families. Currently, finding aids are increasingly encoded in XML using the standard Encoded Archival Descriptions. Archives can cover hundreds of meters of material, resulting in long and detailed EAD documents. We use a dataset of 2,886 EAD documents from the International Institute of Social History (IISG) and 3,119 documents from the Archives Hub, containing documents with more than 100,000 words. Navigating in such archival finding aids becomes non-trivial, and it is easy to loose overview of the hierarchical structure. Hence, this may lead to the loss of important contextual information for interpreting the records.

Archival context may be preserved through the use of authority records capturing information about the record creators (corporations, persons, or families) and the context of record creation. By separating the record creator's descriptions from the records or resources descriptions themselves, we can create "links" from all occurrences of the creators to this context. The resulting descriptions of record creators can be encoded in XML using the emerging Encoded Archival Context (EAC) standard. Currently, EAC has only been applied experimentally. One of the main barriers to adoption is that it requires substantial effort to adopt EAC. The information for the creator's authority record is usually available in some

form (for example, EAD descriptions usually have a detailed field <bioghist> about the archive's creator). However, linking such a context description to occurrences of the creator in the archival descriptions requires more structure than that is available in legacy data.

Our main aim is to investigate if and how automatic link detection methods could help improve archival access. Automatic link detection studies the discovery of relations between various documents. Such methods have been employed to detect "missing" links on the Web and recently in the online encyclopedia Wikipedia. Are link detection methods sufficiently effective to be fruitfully applied to archival descriptions? To answer this question, we will experiment on the detection of archival creators within and across finding aids. Based on our findings, we will further discuss how detected links can be used to provide crucial contextual information for the interpretation of records, and to improve navigation within and across finding aids.

Link Detection Methods

Links generated by humans are abundant on the World Wide Web, and knowledge repositories like the online encyclopedia Wikipedia. There are two kinds of links: incoming and outgoing links. Substrings of text nodes are identified as *anchors* and become clickable. Incoming links come from text nodes of target files (destination node) and point to a source file (origin node), while an outgoing link goes from text node in the source document (origin node) to a target file (destination node). Two assumptions are made: a link from document *A* to document *B* is a recommendation of *B* by *A*, and documents linked to each other are related.

To automatically detect whether two nodes are connected, it is necessary to search the archives for some string that both share. Usually it is only one specific and extract string. A general approach to automatic link detection is first to detect the *global similarity* between documents. After the relevant set of documents has been collected, the *local similarity* can be detected by comparing text segments with other text segments in those files. In structured documents like archival finding aids, these text segments are often marked up as logical units, whether it be the title *<titleproper>*, the wrapper element *<c12>* deep down in the finding aid, or the element *<persname>* that identifies some personal names. These units are identified and retrieved in XML Element retrieval. The identification of relevant anchors is a key problem, as these are used in the system's retrieval models to point to (parts of) related finding aids.

Experiment: Name Detection

A specific name detection trial with the archive of Joop den Uyl (1919-1987), former Labor Party prime minister of the Netherlands, is done as a test to deal with this problem. This archive consists of 29,184 tokens (with removal of the XML markup and punctuation), of which 4,979 are unique, and where a token is a sequence of non-space characters. We collect a list of the name variants that we expect to encounter: "J.M. Den Uyl", "Joop M. Den Uyl", "Johannes Marten den Uyl", "Den Uyl", etc. We construct a regular expression to fetch the name variants. The results are depicted in illustration 1, which shows the local view of the Joop den Uyl archive in our *Retrieving EADs More Effectively* (README) system.

Illustration 1: Links detected in EAD

The quality of the name detection trial is evaluated with explicit feedback, which means manually checking the detected links for (1) correctness, (2) error, and (3) whether any links were missing. This was done both within finding aids, and across finding aids:

- First, the quality is checked within finding aids, by locating occurrences of creator Joop den Uyl in his archive. For detecting name occurrences within an archive, our simple method has a precision of (114/120 =) 0.9500, a recall of (114/117 =) 0.9744, resulting in an F-score of 0.9620. Some interesting missing links used name variants where the prefix "den" is put behind the last name "Uyl" -- a typical Dutch practice. Incorrect links mostly are family members occurring the archive, e.g., "Saskia den Uyl", "E.J. den Uyl-van Vessem", and also "Familie Den Uyl". Since these names occur relatively infrequent, few errors are made. The matching algorithm could easily be refined based on these false positives.

Table 1: Archive "Den Uyl"

	Link	No link
Name	114	3
No name	6	-

- Second, the same procedure to detect proper names of Joop den Uyl is applied across finding aids with the related archive of "Partij van de Arbeid Tweede-Kamer Fractie (1955-1988)" (Dutch MPs from the Labor Party). For detecting name occurrences across archives, we obtain a perfect precision, recall, and thus F-score of 1.

Table 2: Archive "PvdA"

	Link	No link
Name	16	0
No name	0	-

Concluding Discussion

In this paper we investigated how currently emerging link detection methods can help enrich encoded archival descriptions. We discussed link detection methods in general, and evaluated the identification of names both within, and across, archival descriptions. Our initial experiments suggest that we can automatically detect occurrences of person names, both within (F-score of 0.9620) and across (F-score of 1) archival descriptions. This allows us to create (pseudo) encoded archival context (EAC) descriptions that provide novel means of navigation and improve access to archival finding aids. The results of our experiments were promising, and can also be expanded to names of organizations, events, topics, etc. We expect those to be more difficult than personal name detection.

There are more uses for detecting cross-links in finding aids besides creating extra contextual information. Detecting missing links is useful for improving the retrieval of separate finding aids, for example, an archival finding aid with many detected incoming links may have a higher relevance. Links can also offer a search-by-example approach, like given one finding aids, find all related finding aids. A step further is to use the cross-links in the categorization of archival data. Concretely for historians and other users, who rely on numerous lengthy archival documents, new insights can be gained by detecting missing cross-links.

Acknowledgments

This research is supported by the Netherlands Organization for Scientific Research (NWO) grant # 639.072.601.

References

Agosti, M., Crestani, F., and Melucci, M. 1997. On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing and Management* 33, 2 (1997), 133-144.

Allan, J. 1997. Building hypertext using information retrieval. *Information Processing and Management* 33, 2 (1997), 145-159.

EAC, 2004. Encoded Archival Context. http://www.iath.virginia.edu/eac/

EAD, 2002. Encoded Archival Description. http://www.loc.gov/ead/

Fissaha Adafre, S. and De Rijke, M. 2005. Discovering missing links in Wikipedia. In *Proceedings of the 3rd international Workshop on Link Discovery.* LinkKDD '05. ACM Press, 90-97.

Huang, W. C., Trotman, A., and Geva, S. 2007. Collaborative Knowledge Management: Evaluation of Automated Link Discovery in the Wikipedia. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 2007.

INEX LTW, 2007. INEX Link The Wiki Track, 2007. http://inex.is.informatik.uni-duisburg.de/2007/linkwiki.html

ISAAR (CFP), 2004. *International Standard Archival Authority Record for Corporate bodies, Persons and Families.* International Council on Archives, Ottawa, second edition, 2004.

ISAD(G), 1999. *General International Standard Archival Description.* International Council on Archives, Ottawa, second edition, 1999.

Jenkins, N., 2007. Can We Link It. http://en.wikipedia.org/wiki/User:Nickj/Can_We_Link_It