Searching Archival Finding Aids: Retrieval in Original Order?

Junte Zhang[1] and Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

Abstract. Archival principles as Provenance (keeping material from the same creator together) and its corollary Original Order (keeping the order of creation intact) could help improve access to the archival materials. We investigate the importance of relevance ranking and 'Original Order' when searching finding aids in EAD using XML Retrieval. Our experiment shows that relevance ranking is of paramount importance, although Original Order may help the retrieval of the first few results because these tend to cluster within the original order.

1 Introduction

Information in digital libraries and on the Web often has a rich internal structure – think of the document structure of books in a digital library. Such structure could be exploited to give direct access to relevant parts in these documents. A particular example of long and richly structured documents are archival finding aids created in Encoded Archival Description (EAD, [5]), which are structured in exactly the same way as the material they describe. First, by the principle of Provenance, all material created or received by the same individual, family, or organization is kept together. Second, by the principle of Original Order, all material of the same creator is stored in its original organization and sequence.

Archivists consider these principles crucial for archival access, though questioned [1], these have never been tested empirically [4]. The archival principle of Original Order corresponds to the preservation of the document hierarchy in an archival description. Physical re-arrangement (such as re-ordering by topic, time or geography), which could enhance user access to the archives, is rejected [3, 4]. Specifically, as stated in [4], even the re-arrangement of archives to suit the needs of historians is disallowed. In recent years, however, archival finding aids have been put online, giving it a new function as an information retrieval and discovery tool for users. Therefore, the actual impact of sticking to the original order of an archive when retrieving and presenting information needs to be examined.

Archives may span 100s or 1000s of meters of material, and the main purpose of the archival description is to help searchers identify the exact parts of the archive to consult. There is a direct and natural parallel between locating parts of the archival finding aids in EAD, and the focused access of other XML documents: XML Retrieval (XML IR) can be used to exploit the internal structure

of an EAD file. This structure could consist of elements that represent lengthy biographies, nested components, all the way down to the single item.

However, each and any of these elements can be returned in any order, either by respecting the Original Order, or returning them according to relevance only, or any other criterion. The retrieval effects of Original Order are not known, so *given the retrieval of any and arbitrary EAD/XML elements according to the relevance with a query, what are the retrieval effects of returning it in Original Order?* On the one hand, Original Order could enhance information access as relevant items may appear close to each other due to the intellectual organization by the archival creator. The would correspond with the Cluster Hypothesis, which deposits that 'closely associated documents tend to be relevant to the same requests' [2]. On the other hand, it may not be a useful feature to improve retrieval because the Cluster Hypothesis may not hold on archival finding aids.

2 Experimental Setup

The search requests were formulated as reference questions. The used queries consisted from 3 upto 13 keywords. Each of these reference questions was judged against a narrative in which the information need is clearly stated, including what is considered relevant. The relevance is determined by locating particular units of archival materials that will likely contain the sought answer. Descriptions of individual files and records tend to be very succinct – seldomly more than a single sentence. Additionally, a finding aid also contains contextual background descriptions of the archive, which may directly contain relevant information.

An assessment tool was created to facilitate the relevance assessments within an archive. For each search request, the most relevant archive was located – which may range up to a 1,000 pages and many thousands of XML elements, and all and only material in this archive was judged. This resulted in a relatively modest test collection (*qrels*) of in total 73 relevant elements in 5 archival finding aids in EAD, which we collected from the National Archives of the Netherlands.

The system used in our experimentation has been described in [7]. We indexed the collection without stopword removal, used the Dutch snowball stemmer, and standard parameters. For the retrieval of any arbritary elements, we employ statistical language models (LM) [6] , i.e. the probability distribution of all possible term sequences is estimated by applying statistical estimation techniques.

3 Results

3.1 Relevance versus Original Order

We first look at the whole run with 1,000 results in Table 1. In the top 1,000 results, both approaches obtain a reasonable recall of 62 out of 73 relevant elements ($R = 0.8493$). Considering that we look for very short descriptions (often a single sentence), the relevance ranking is performing quite well with a MAP of 0.1454 and a P@10 of 0.1600. What happens if we rank these 1,000 results in

Table 1. Retrieval Performance for first top N results for each topic.

Run	Top 1,000		Top 500		Top 100		Top 10		Top 5	
	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
Relevance	0.1600	0.1454	0.1600	0.1398	0.1600	0.1321	0.1600	0.0917	0.1400	0.0822
Original Order	0.0000	0.0296	0.0000	0.0260	0.0400	0.0517	0.1600	0.0660	0.1400	0.1031

Table 2. DOM Tree distances in the qrels.

Topic ID	Count	Mean Depth	Mean Distance	Total count <C$_n$>
1.04.02	11	10.000	3.545	17,184
2.19.123	37	10.811	1.492	5,661
2.19.124	7	6.000	1.286	1,491
2.03.01	7	10.143	1.429	14,017
2.21.286	4	9.750	0.750	2,035

their original order? The score plummets to almost zero; the relevance ranking is crucial. It should be noted that we are reranking the top N of results, and usually there are just a handful of relevant results (also see Table 2).

Given that set of results must contain many non-relevant ones, reranking the top 1,000 on original order may not fairly reflect the utility of the original order. What will happen if we rerank a smaller set of results? The remaining columns of Table 1 show the results for different sets. The MAP of the relevance ranking drops as expected for the shorter runs. As the cut-off level is decreased, we see that the precision of the original order ranking increases. Interestingly, we see that the MAP for Original Order is higher than the standard element ranking when the cut-off level for each topic is set to 5. This signals that although the relevance ranking is of paramount importance, there is also still potential value in the original order, because relevant results have a tendency to cluster.

3.2 Cluster Hypothesis Effects

We want to further investigate the Cluster Hypothesis – how near are the relevant results in the original order of the document? We do this by measuring the distance between the relevant elements in the DOM tree. We restrict our attention to the relevant elements in the hierarchical descriptions of the archive (i.e., the component <C$_n$> elements) in EAD. The components <C$_n$> are nested within each other in <ARCHDESC> given the n, where $n \in \{01, ..., 12\}$. A component can also be unnumbered. The results are shown in Table 2. The first topic has 11 results with a mean depth of 10 nodes in the DOM tree. For each pair of results, we look at the distance to a common ancestor, which could be at most the depth itself (i.e., 10). For the first topic the mean distance over all pairs is 3.5 – which signal that the results are somewhat scattered through the archive. However, for the other topics the mean distance is between 0.75 and 1.5, which shows that relevant results occur in close proximity within the archive, especially given the

large quantity of <C$_n$> elements per topic (see Table 2). For example, in topic *1.04.02* only 11 out of 17,184 (or 0.06%) <C$_n$> elements were seen as relevant.

3.3 Sparse Data on the Item Level

A challenge for effective XML IR is the sparse data, especially in the unit titles, which is a distinct property of the archival descriptions. When we analyze the selected relevant elements, we see that there are very short phrases, sometimes without the occurence of a keyword, e.g. "*Diverse stukken,* (Unit ID: 824)" (in English: "*Several pieces,*"). The sparse data on the item level can be attributed to idea of *inheritance of description* – each lower level inherits the description of the container [5], where this context is a crucial cue for assessing relevancy as related relevant items tend to be located in short distance from each other.

4 Conclusions and Future Work

We empirically examined the impact of the archival principle of Original Order on the ranking of search results by comparing it with a metadata retrieval system using XML IR techniques. Our results show that the relevance ranking is of paramount importance, but that the results have a (weak) tendency to cluster.

The Principle of Original Order is useful, because physical materials can only be ordered in a single way, and here the traditional archival practices make much sense. The question arises whether it will continue to be as useful in this digital age. With the advent of digital archives, we are no longer bound to the physical and practical limitation of before and we could construct multiple ordering of the same material inclusing those based on a search request or search profile at hand. This opens up a wealth of possibilities to improve archival access, and unleash the valuable treasures now hidden deeply inside the archive.

Acknowledgments We gratefully acknowledge the National Archives of the Netherlands, and Henny van Schie, for their support. This research is supported by the Netherlands Organisation for Scientific Research (NWO) under project #639.072.601.

Bibliography

[1] F. Boles. Disrespecting Original Order. *American Archivist*, 45:26–33, 1982.
[2] N. Jardine and C. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.
[3] H. Jenkinson. Reflections of an Archivist. *Contemp. Review*, 165:355–361, 1944.
[4] R. H. Lytle. Intellectual Access to Archives: I. Provenance and Content Indexing Methods of Subject Retrieval. *American Archivist*, 43(Winter):64–75, 1980.
[5] D. V. Pitti. Encoded archival description: The development of an encoding standard for archival finding aids. *American Archivist*, 60(Summer):268–283, 1997.
[6] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM.
[7] J. Zhang, K. N. Fachry, and J. Kamps. Access to Archival Finding Aids: Context Matters. In *ECDL '08*, pages 455–457, Berlin, Heidelberg, 2008. Springer.