Finding stories in the archive through paragraph alignment

Weijia Xu and Maria Esteva
Texas Advanced Computing Center, University of Texas at Austin, Austin, TX, USA

Abstract

Referring to the concept of archival bond, we define *stories* as formed by documents that relate to a target activity and developed a method called *paragraph alignment* to find these documents. The method computes archival bond by measuring the cosine similarity between document paragraphs. We tested the method in a chaotic case study collection created in a shared server by different authors. Results demonstrate that this method is more efficient to find stories than calculating the cosine similarity between entire documents. This research helps archivists make sense of collections that are considered inaccessible and whose stories may otherwise be lost.

Correspondence:
Weijia Xu, 10100 Burnet Road, ROC 1.101 (Bldg. 196) R8700, Austin, TX 78758-4497, USA.
E-mail:
xwj@tacc.utexas.edu

1 Introduction

As primary sources, archival collections are unique windows to people and organizations. Finding aids created by archivists point to the location of documents in a collection and show the relationships between them across time and provenance. This access tool allows researcher to find stories in the archive and provides the contextual information that enables making sound inferences and interpretations. And yet, due to the unorganized condition that characterizes many contemporary electronic document collections, their stories may remain buried in the archive.

We refer to collections such as those found on shared organizational servers, in which over time employees deposit work documents in idiosyncratic fashion (Henry, 2003). Ubiquitous in the modern workplace and rich in work experiences, these collective aggregations are perceived as chaotic, dexfined as ROT (redundant, outdated, and trivial), and deemed disposable by some archivists and records managers who find—and justifiably so—that they cannot make sense of them and consequently describe them for access (Public Record Office, 2000; Henry, 2003; AIIM, 2009). Considering the concept of *archival bond* we developed a method to recover the documents that form stories. We conceive stories as narratives about a specific activity that may span time, organizational areas, and creators. A story may include similar texts as well as documents that differ in length, style, and that may contain parts of other stories.

In archival theory, *archival bond* describes the relationships between documents in a collection as essential properties of the documents (Duranti, 1997). While all the documents are bonded through the collection's organization structure (McNeil, 2000), there are stronger relationships between sub-groups of documents that belong to the same activity. In controlled electronic record keeping systems, indexes, time stamps, metadata, and record-keeping rules are manifestations of archival bond, and thus of the relationships between documents. In the case of chaotic document collections in which the collection's structure is blurry, we suggest that archival bond be based on the documents' content referring to a target activity. We designed a

Literary and Linguistic Computing, Vol. 26, No. 3, 2011.

doi:10.1093/llc/fqr017 Advance Access published on 13 May 2011

method called *paragraph alignment* (PA) to compute archival bond and find the documents that narrate the story of an activity.

2 Methodology and Related Work

In information retrieval, a classic way to find related documents is to measure the cosine similarity between the documents in a vector model (Baeza-Yates and Ribeiro-Neto, 1999). Using cosine similarity to identify archival bond has specific limitations. Cosine similarity measure does not account for differences in length between documents, nor for documents that while related to the same activity are loosely similar. Instead, it will score high documents that share same words but are not part of a story. In general, these variations are present in archival collections.

We propose to compute the similarity between two documents locally instead of globally, a concept drawn from local alignment methodology in bioinformatics (Gusfield, 1997). While biological sequences evolve throughout time owing to constant mutation events, the parts of the sequences that directly participate in cellular activities remain relatively stable. Therefore, the global similarity between two sequences is often less important than the local similarity, which is defined by the highest similarity between any two substrings from two sequences. Here we adapt a similar approach to model archival bond as the most similar parts between two documents.

Figure 1 shows the methodology workflow. Each document in a set is broken into one or more ordered segments based on the number of paragraphs in the document. If the length of a segment (including spaces) is less than a predefined minimum number of characters threshold (MNCT), the segment is merged with the next one. We then remove stop-words and transform each segment into a vector to compute the similarity between every pair of segments. For a pair of documents, we derive a score based on the maximum similarity score found between their segments.

Rather than measuring inter-paragraph similarity within one document to identify subtopic structure as done by Hearst (1994), we compare document segments to identify related topics across a collection. Hence the goal of segmentation is to minimize the variation of length between the documents. Barzilay and Elhadad (2003) use a dynamic programming algorithm to align sentences and find documents that present similar information. Since we aim to find documents that convey different information about a target activity, we require less

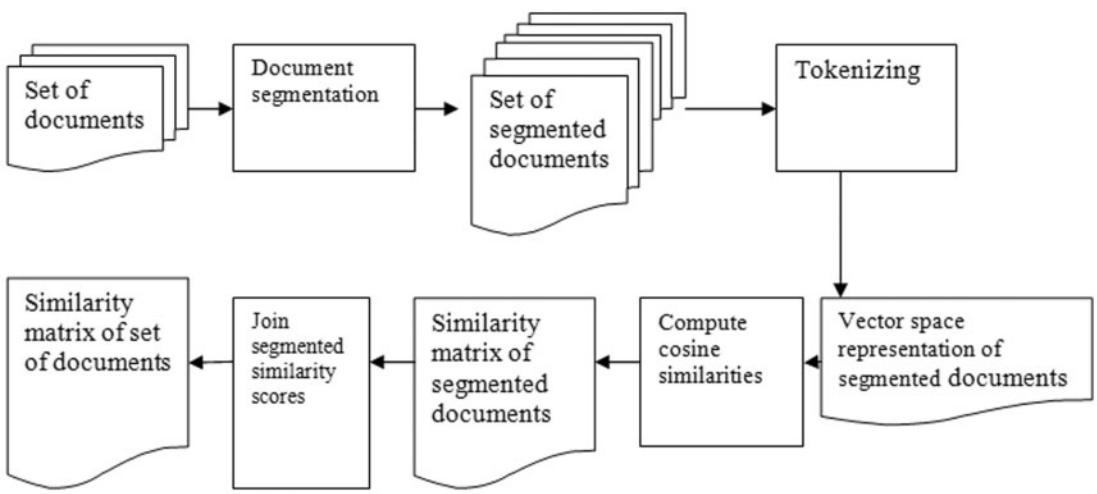

Fig. 1 Workflow to compute document similarity with PA

precision to identify similarity. Hence, our method is designed to identify related documents using the vector space model for words within paragraphs. Moreover, in contrast to PA, sentence alignment demands more computation. Currently, many research efforts are directed towards detecting duplicate and near duplicate webpages (Fetterly *et al.*, 2005; Henzinger, 2006). While PA recovers different versions of documents with high similarity, our interest is to find documents that, while related, may not have a high similarity score.

3 Implementation and Evaluation

3.1 Case study collection

We tested our method in a case study digital archive belonging to a philanthropic organization. Over fifteen years, the employees deposited their drafts and final documents in directories under their name on a shared server following their own record-keeping criteria. Observing the collection we noted that for periods at a time, duplicate or near duplicate paragraphs about an activity are present across long documents (annual reports, call for grants, plans, and board meeting minutes), while short documents about the activity (letters, memos, project summaries, and press releases) share some or many of the same terms. Short documents contain between 400 and 600 words and long ones 4,000 words or more. In turn, many long and short documents share the same terms, names, and places but in relation to different activities, and documents that have common business phrases are not about the same activity (Esteva, 2007). The collection reflects the employee's records creation practices, afforded by the cut and paste function of the word processor, the possibility to co-write documents, and the use of a common organizational language.

3.2 Test of the paragraph alignment method

The digital archive spans from 1996 to 2005 and contains a total of 16,404 documents and seventeen authors. Documents per year range from 500 to 3,000. We tested the method in all 714 documents from the year 1997 with eight authors and document types that are common to the entire archive. To sort the documents by author and by last modified date we used file management software. Documents were segmented using MNCT of 1,000, 750, and 500 characters. Year, author, and segment number were preserved in the documents' filename.

The test was based on assessing seven test groups. We selected five query documents, each corresponding to a target activity (test groups 1, 2, 4, 5, and 6), and two containing summaries of different activities (test groups 3 and 7). This selection aimed to return one story per each of the five test groups, and different stories for test groups 3 and 7. For each query document, we also identified a number of related documents as control groups, although other related documents could exist in the set.

For each test group, both the cosine similarity and the PA methods returned a list of documents ranked from more to less similar. For evaluation purposes and to determine the noise in the trail, we checked the results against the control groups, labeling the ranked document as a 'true positive' if it was related to the query document; otherwise the document was labeled as a 'false positive'.

3.3 Results evaluation

We compared the results of the different MNCT with those obtained by calculating cosine similarity as a measure of global similarity between the documents. Table 1 shows that the PA method with a MNCT of 750 characters returned the best results five out of seven times (test groups 1, 2, 4, 5, and 6). For test group 7, the best results were obtained with a MNCT of 500 characters. In this case, the query document contained multi-paragraph summaries of five different activities, which are also mentioned in other documents in the set. The PA method did not work for test group 3, which contains a long list of short sentences about different activities. The distinctive terms in these sentences correspond to proper names, which are not present in other documents in the set. In turn, the terms describing the activities are common across the recovered false positives.

Figure 2 shows a plot of the results of test group 1 in which the PA method with a MNCT of 750

Table 1 Results comparison of the PA and the cosine similarity methods

Test group	1	2	3	4	5	6	7
No. of characters (with space)	1,170	2,425	7,530	1,390	2,654	1,176	1,1356
No of true positives	21	5	9	17	19	19	43
No. of false positives							
Cosine	28	36	**6**	205	88	53	103
PA 500	10	49	10	20	87	46	**47**
PA 750	**6**	**27**	12	**6**	**16**	**38**	91
PA 1000	7	186	11	17	100	117	71

Best results are shown in bold.

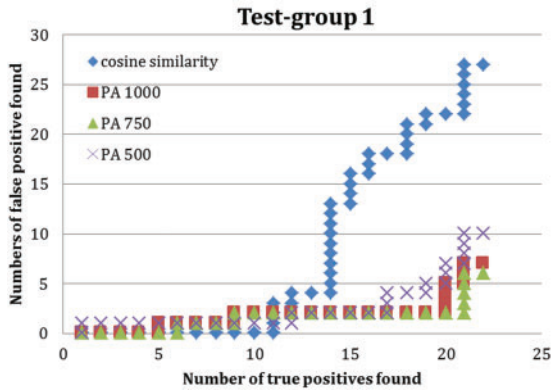

Fig. 2 Comparison of PA method and cosine similarity for test-group one

characters performed the best. The test group tells the story of a program to train young orchestra directors. The query document is a memo with a brief description of the project and estimated costs for lodging and travel for the participants. Returned true positives of five authors include: other planning documentation, correspondence with stakeholders, the call for grants, a press release, a list of participants, the musical program, and annual and reports in which the program is described among other programs in the organization. Since results are ranked by similarity, the chronological sequence of the story is found in the documents content.

For the results of the PA with 750 MNCT, before the last true positive is found, the graphic shows a peak with a sequence of false positives. After reading the documents we found that they are unrelated to the target activity but that some of its paragraphs share common terms with the query documents. This sequence is more pronounced for the cosine similarity results.

The results suggest that although related documents may not share similar global word distributions, they share similar word distributions in some of their segments. While the efficiency of the different MNCT depends on the particular word distribution of the documents that are being compared, in general, with a smaller MNCT, the PA method is more effective to retrieve related documents with less global similarity. However, if the MNCT is very small, the number of matching words shared by two paragraphs may decrease. Therefore, the method may return more false positives.

4 Conclusions

This research has implications for the long-term retention and access of chaotic archives. Using the concept of archival bond as a theoretical framework, we developed a method to find stories and make sense of a ROT archive that may be otherwise discarded. Results show that local similarity matters to identify documents that relate to the same story even though they may be different in size and convey somewhat different information. While PA still retrieves false positives, it is an improvement over the cosine similarity measure. Currently, to better identify the adequate MNCT, we are considering the distribution of number of characters in the collection. Research conducted in archives normally implies reviewing extensive holdings in an attempt to exhaust the possibilities of finding

evidence. PA allows archivists to find not just individual documents but a story. By recovering the documents that tell a story in relation to time and provenance, the archival bond between documents, and therefore the collection's structure, starts to show. Stories may then become points of access that embed contextual information.

Funding

This work was supported by a National Archives and Records Administration (NARA) supplement to the National Science Foundation (NSF) Cooperative Agreement (OCI-0504077).

References

AIIM. (2009). Best Practices for Information Organization and Access. http://www.aiim.org/infonomics/best-practices-for-IOA.aspx (accessed 29 October 2009).

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston: Addison-Wesley Longman Publishing.

Barzilay, R. and Elhadad, N. (2003). Sentence Alignment for Monolingual Comparable Corpora. *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*. Stroudsburg: Association for Computational Linguistics, pp. 25–32.

Duranti, L. (1997). The archival bond. Archives and museum informatics. *Springer*, **11**(3): 213–18.

Esteva, M. (2007). Bits and *Pieces of Text: Appraisal of a Natural Electronic Archive*. In *Conference Abstracts of Digital Humanities 2007, 19th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*. June 2007. Urbana-Champaign, Illinois, pp. 55–57.

Fetterly, D., Manasse, M., and Najork, M. (2005). *Detecting Phrase Level Duplication on the World Wide Web*. In *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval*. New York: ACM Press, pp. 170–77.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

Hearst, M. A. (1994). *Multi-Paragraph Segmentation of Expository Text. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 9–16.

Henry, L. J. (2003). Appraisal of Electronic Records. In Ambacher, B. I. (ed.), *Thirty Years of Electronic Records*. Maryland: The Scarecrow Press, p. 38.

Henzinger, M. (2006). Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms. In *Proceedings of the 29th Annual International Conference on Research and Development in Information retrieval*. New York: ACM Press, pp. 284–91.

McNeil, H. (2000). *Trusting Records: Legal, Historical and Diplomatic Perspectives*. Norwell: Kluwer Academic Publishers.

Public Records Office. (2000). Guidance for an Inventory of Electronic Records: a Toolkit. http://www.nationalarchives.gov.uk/documents/inventory_toolkit.pdf (accessed 29 October 2009).