Finding Stories in the Archive through Paragraph Alignment

Esteva, Maria
maria@tacc.utexas.edu
Texas Advanced Computing Center (TACC), University of Texas at Austin, USA

Xu, Weijia
xwj@tacc.utexas.edu
Texas Advanced Computing Center (TACC), University of Texas at Austin, USA

We present research showing the possibility of finding stories in a digital text archive through computational methods. Referring to the concept of "archival bond", we define stories as formed by documents that relate to a target activity. We developed a method called *paragraph alignment* to find such documents and an interactive visualization to discover connected stories in context with provenance.

Our method was applied to the challenges presented by the digital archive of a multinational philanthropic organization who awarded grants to cultural, scientific, and social welfare activities (1985-2005). Over fifteen years, the staff members deposited their work documents in individual directories on a shared server without following any record-keeping rule. These documents reflect the organization's activities in the areas of Science and Education, Art and Humanities, and Social Welfare. They also reflect the staff members' records creation practices, afforded by the cut and paste function of the word processor and the possibility to collaborate through the network. These digital aggregations are sometimes perceived as chaotic, defined as ROT (redundant, outdated and trivial,) and deemed disposable (Henry, 2003; AIIM, 2009; Public Records Office, 2000). Yet they are ubiquitous in the networked servers of many organizations, so our goal was to find a method to make sense of the text records within.

1. Archival Bond

A fundamental concept in archival theory, known as archival bond, describes the relationships between documents in an archive as essential properties of the documents (Duranti, 1997). While all the documents in a collection are bonded through the collection's structure (McNeil, 2000), there are stronger relationships between sub-groups of documents that belong to the same function and/or activity. In the case of disorganized electronic text archives in which the structure is nonexistent or loose, we suggest that the relationships among documents be defined based on their content referring to a target activity. By finding trails of documents that narrate stories about activities in context with provenance, we aim to establish order, identify structure, and learn about the archive's creators.

2. Paragraph Alignment (PA)

We observed that in our archive, similar paragraphs about an activity are repeated across short - memos and press releases - and long documents - annual reports and board meeting minutes. As a group, these documents tell the story of an activity. We also observed that in many documents the same personal names, places, and institutions are mentioned in relation to different activities, and that documents that use similar terms may not be associated with the same activity. The traditional cosine similarity method measures global similarity between documents. Given the characteristics noted in this archive, we considered that calculating global similarity was not efficient to identify all the documents about a target activity. Instead, we draw from local alignment, a method used in bioinformatics to evaluate local similarity between sequences (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, 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. Efficient methods for computing sequence similarities often follow a framework in which sequences are broken into n-gram for similarity computations and then assembled to derive an overall similarity (Wu et al., 1990). Here we adapt a similar approach that we call paragraph alignment to determine archival bond between documents.

Our method contrasts with previous work on document segmentation (Hearst, 1994). Rather than measuring inter-paragraph similarity within one document to identify subtopic structure, our approach focuses on comparing the similarity between document segments to identify topics across a collection. Hence the primary goal of document segmentation is to minimize the variation of length between documents for subsequent similarity comparison.

3. Methodology

Figure 1 shows the workflow of our approach.

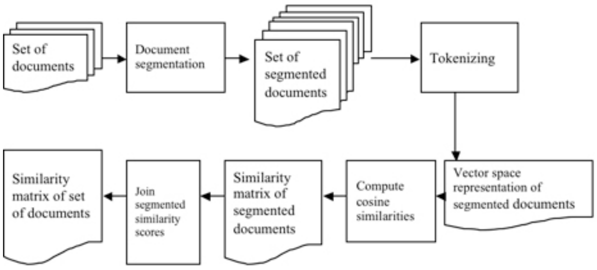
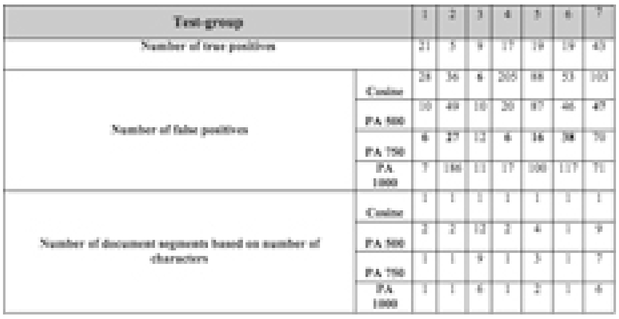

Set of documents
Document segmentation
Set of segmented documents
Tokenizing
Vector space representation of segmented documents
Compute cosine similarities
Similarity matrix of segmented documents
Join segmented similarity scores
Similarity matrix of set of documents

Figure 1

Each document in a set is broken into one or more ordered segments based on the paragraphs in the document. If the length of a segment (including spaces) is less than a pre-defined minimum number of characters threshold (MNCT), the segment is merged with the following segment. We used MNCT of 1000, 750, and 500 characters. For each set of document segments we create a matrix of TFIDF weighted term frequencies after stop-words removal (McCallum, 1996), and then calculate the cosine similarity between every other segment (Salton, 1988). We then process the resultant matrix to derive similarity scores between document pairs, which are defined as the maximum similarity score between their segments. For evaluation, we compare the results of the different MNCT with those obtained by calculating cosine similarity as a measure of global similarity between the documents.

We tested the method in a set of 714 documents from the year 1997 with eight authors. Date and authorship were preserved in the documents' file name. The evaluation was based on assessing seven document test-groups. A team member familiar with the archive selected five query documents, each corresponding to a different activity (test-groups 1, 2, 4, 5, 6) and two containing summaries of various activities (test-groups 3 and 7). For each query document, the team member also identified a set of related documents. For each test-group, both the cosine similarity and the paragraph alignment methods returned a list of documents ranked from more similar to less similar. The team member checked the results against the content of the corresponding document labeling the ranked document as a "true positive" if it was related to the query document; otherwise the document was labeled as "false positive". Results were checked until the last true positive was found.

4. Results



Test-group		1	2	3	4	5	6	7
Number of true positives		21	5	9	17	19	19	43
Number 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	70
	PA 1000	7	166	11	17	100	117	71
Number of document segments based on number of characters	Cosine	1	1	1	1	1	1	1
	PA 500	2	2	12	2	4	1	9
	PA 750	1	1	9	1	3	1	7
	PA 1000	1	1	6	1	2	1	6

Table1

The results show that the PA method with a MNCT of 750 characters returned better results five out of seven times (test-groups 1, 2, 4, 5, 6 and 7). For test-group 7, the best results were obtained with a MNCT of 500 characters. In this case the query document contained summaries of five different projects accomplished during 1997, each mentioned in other documents in the set. This suggests that although related documents in the set 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, the smaller the MNCT used the higher the documents with less global similarity are ranked by the PA method. The PA method did not work for test-group 3 which contains sentences about activities most of which are not mentioned in other documents in the set. Figure 2 shows a plot of the results of test-group 1 in which the PA method with a MNCT of 750 characters performed the best.

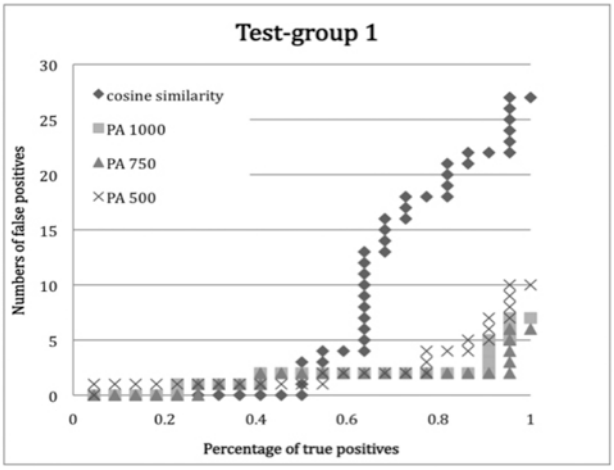


Figure 2

The test-group one (Figure 2) contains documents about a program to train young orchestra directors. The query document is a memo including a brief description of the project and estimated costs for lodging and travel. Returned true positive documents of five authors include: other planning documentation, correspondence with potential contributors, the call for applications, a press release, a list of participants, the musical program, and various reports.

5. Visualization

Through an interactive visualization (Figure 3) users can follow the connections between documents to identify stories (PREFUSE). Each document is labeled with a color corresponding to its author. The connectivity between the documents allows the identification of a) stories, b) the authors involved in a given activity, and c) connected stories. As the user interacts with the visualization, the structure of the archive takes shape. Below is a snapshot of the visualization interface showing stronger connections between a group of documents.

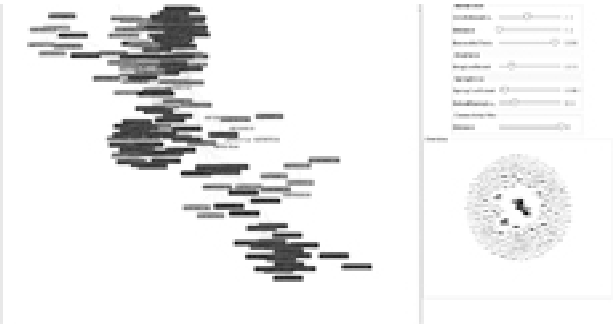


Figure 3

6. Conclusions

This research has implications for the retention of digital archives. Using the concept of archival bond as a framework we aim to make sense of ROT archives and to unveil their stories. The results show that for documents that share similar paragraphs, local similarity matters to identify an archival bond. The same characteristic is observed in the biological sequence analysis that inspired our method.

7. Acknowledgments

This work was supported through a National Archives and Records Administration (NARA) supplement to the National Science Foundation Cooperative Agreement (NSF) TERAGRID: Resource Partners, OCI-0504077.

References

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

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

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

Duranti, L. (1997). 'The Archival Bond'. *Archives and Museum Informatics*. 213.

McNeil, H. (2000). *Trusting Records: Legal, Historical and Diplomatic Perspectives.* Springer.

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

Wu, S. Manber, U., Myers, G. and Miller, W. (1990). 'An O (NP) Sequence Comparison Algorithm'. *Inf. Process. Lett.* **35(6)**: 317-323.

Hearst, M.A. (1994). 'Multi-Paragraph Segmentation of Expository Text'. *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics.* Las Cruces, New Mexico, pp. 6-9.

McCallum, A. K. (1996). *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, Computer software.* `http://www.cs.cmu.edu/~mccallum/bow` (accessed 10 May 2009).

Salton, G. and Buckley, C. (1988). 'Term-weighting Approaches in Automatic Text Retrieval'. *Information Processing & Management.* **24(5)**: 513–523.

PREFUSE, Interactive visualization software. `http://prefuse.org/` (accessed 10 May 2009).