Advanced Information Systems for Archival Appraisals of Contemporary Documents

William McFadden, Kenton McHenry, Rob Kooper, Michal Ondrejcek, Alex Yahja and Peter Bajcsy
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Abstract

This work addresses the problem of designing a scalable framework for archival appraisals of contemporary PDF documents. The motivation for our work is to provide an e-Science solution that (a) fuses the independent research methodologies focusing on specific information types to one comprehensive analytical framework, (b) optimizes tradeoffs between computational requirements and preservation costs, and (b) bridges the small scale and large scale computational studies. The e-Science solution presented here consists of (1) a methodology for comprehensive comparisons of contemporary documents containing text, images and vector graphics, (2) a framework for including 3D and 3D+time data sets into the appraisal analyses, (3) interfaces supporting exploratory archival appraisal analyses with small scale data sets, and (4) infrastructure supporting the transition from small scale to large scale computations using commodity and high performance computing resources. The novelty of our work is in designing methodologies, mathematical frameworks and prototypes for comprehensive and scalable document appraisals that include text, images, vector graphics , and high dimensional data.

1. Overview

The objective of our work is to design a methodology, algorithms and a framework for contemporary document appraisal. The motivation for our work comes from the fact that the exponentially increasing amounts of electronic records have to be related to the existing permanent records, verified for integrity and authenticity, ranked and selected for preservation. Due to the heterogeneous content of electronic records and the volume of the records, specifically contemporary office documents, the challenges include (a) exploration of components in document containers, (b) comparative analyses and relationship extraction, (c) automation of labor intensive operations of the appraisal process, and (d) processing of large collections of documents with the appropriate software and hardware.

Our approach to the above challenges is (a) to enable exploratory document analyses by building a visual inspection framework, (b) to provide comprehensive comparisons of text, image, vector graphics and high-dimensional data objects forming document content by fusing content-based retrieval approaches, (c) to design integrity/authenticity verification by machine learning and modeling, (d) to evaluate computational and storage requirements for archival purposes by incorporating hardware specific tradeoff studies, and (e) to support automation and scalability of appraisal analyses by software parallelization for multiple hardware architectures. In order to address the aforementioned challenges, we decomposed the series of appraisal criteria[1] into a set of focused analyses, such as (a) to find groups of records with similar content, (b) to rank records according to their creation/last modification time and digital volume, (c) to detect inconsistency between ranking and content within a group of records, and (d) to compare sampling strategies for preservation of records.

In this work, we narrowed our focus to those electronic documents that contain primarily text, raster, vector graphics and 3D objects. Among the existing file formats, we selected the Adobe Portable Document Format (PDF) as the container that not only contains those aforementioned digital objects but also could be described as the most widely used format for exchanging documents in office environments.

In order to address the appraisal criteria, we adopted some of the text comparison metrics used in [1], image comparison metrics used in [2] and lessons learnt stated in [3]. Then, we designed a new methodology for grouping electronic documents based on their content similarity (text, image and vector graphics), and prototyped a solution supporting grouping, ranking and integrity verification of any PDF files and HTML files. First, text based, vector based and multi-image based comparisons are performed separately. Multiple images in each

[1] http://www.archives.gov/oig/reports/september-2005.html#challenges
POC: Peter Bajcsy, pbajcsy@ncsa.uiuc.edu, 217-265-5387.

document are grouped first and then groups of images across documents are compared to arrive to an image-based similarity score. The current prototype is based on color histogram comparison, line count in vector graphics and word frequency comparison. The image colors and word/ integers/ floating numbers can be analyzed visually to support exploratory analyses as shown in Figure 1. Subsets of the undesirable text and image primitives could be filtered out from document comparisons (e.g., omitting conjunctions, or background colors). The results of text, image and vector based comparisons are fused to create a pair-wise document similarity score. The matrix of pair-wise document similarity scores are used for grouping. The other appraisal criteria are approached by ranking documents within a group of documents based either on time stamps or on file name indicating the version number. The inconsistency between ranking and content within a group of records is based on frequency tracking, where the frequency of text, image and vector primitives is monitored over the time/version dimension of the grouped documents.

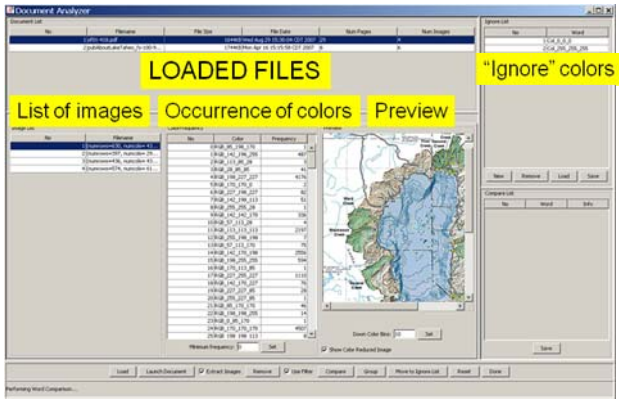

Figure 1: Exploratory framework supporting browsing PDF components and comprehensive comparisons.

In addition, we have explored the 3D file formats, data representations and 3D viewers that could be contained in office containers like PDF. Within the context of preservation, the most critical piece of 3D information is the 3D geometry followed by material properties of the model and the environment parameters. The plethora of 3D file formats (more than 140 formats found in our survey) poses challenges to comparative analyses, ranking, verification and preservation similar to the challenges arising from the volumes of individual documents and the number of documents to be analyzed. In order to address the scalability of computations, we have investigated and prototyped the use of Google's MapReduce and Yahoo!'s Pig frameworks. MapReduce is a programming model that allows programmers to focus on the tasks instead of the parallel implementation of the tasks. For the prototype, we leveraged an open source implementation of MapReduce called Hadoop which is available via the Apache Foundation. This approach to software parallelization aims at the use of commodity computer clusters in contrast to the use of high performance computing resources utilizing parallel programming paradigms based on message-passing interface (MPI) or open multi-processing (OpenMP). Based on our benchmarking results, MapReduce (Hadoop implementation) does not perform very well in heterogeneous environments as confirmed also by the most recent tech. report [4]

2. Summary

The novelty of our work is in designing a methodology for computer-assisted appraisal and in developing a mathematical framework for automation of appraisals based on image, vector graphics, text and high-dimensional types of information representation. Furthermore, our contribution is in prototyping a computer assisted appraisal system and investigating scalability approaches to address the computational requirements of such appraisal analyses. Although we selected to work with documents in PDF format, the framework is applicable to any file format as long as the information can be loaded from any proprietary file format.

3. Acknowledgement

This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019.

4. References

1. Salton, G., J. Allan, and C. Buckley, *Automatic structuring and retrieval of large text files.* in Communication of the ACM, 1994. **37**(2): p. 97-108.
2. Squire, D.M., et al., *Content-Based query of image databases: inspirations from text retrieval.* Pattern Recognition Letters, 2000. **21**: p. 1193-1198.
3. Marshall, J.A., *Accounting For Disposition: A Comparative Case Study of Appraisal Documentation at the NARA in the US, Library and Archives Canada, and the NAA,* in *Dep. of Library and Information Science*. 2006, Univ. of Pittsburg.
4. Zaharia, M., et al., *Improving MapReduce Performance in Heterogeneous Environments*. 2008, University of California at Berkeley.