Building a Test Collection for Complex Document Information Processing

D. Lewis
David D. Lewis Consulting
Chicago, IL 60614
sigir06poster@DavidDLewis.com

G. Agam, S. Argamon, O. Frieder, D. Grossman, J. Heard
Dept. of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
{agam, argamon, frieder, grossman, hearjef}@iit.edu

ABSTRACT
Research and development of information access technology for scanned paper documents has been hampered by the lack of public test collections of realistic scope and complexity. As part of a project to create a prototype system for search and mining of masses of document images, we are assembling a 1.5 terabyte dataset to support evaluation of both end-to-end complex document information processing (CDIP) tasks (e.g., text retrieval and data mining) as well as component technologies such optical character recognition (OCR), document structure analysis, signature matching, and authorship attribution.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.7.5 [**Document and Text Processing**]: Document Capture

General Terms
Measurement, Experimentation

Keywords
Corpora, metadata, queries, relevance judgments, TREC

1. INTRODUCTION
Analysis of masses of scanned paper documents is critical in intelligence, law, knowledge management, historical scholarship, and other areas. The documents are often complex in structure, include non-textual elements such as graphics and photos, and are produced by a variety of printing and handwriting technologies. What we call complex document information processing (CDIP) therefore requires combining evidence from document structure analysis, optical character recognition, signature and logo recognition, authorship attribution, named entity recognition, and other component technologies to accomplish information retrieval (IR) and data mining tasks.

SIGIR '06, August 6–9, 2006, Seattle, Washington, USA

Elsewhere we describe experience with a prototype CDIP system for retrieval and data mining on scanned documents [1]. One goal of our project is to evaluate the effectiveness of that system, and how that effectiveness responds to changes in the effectiveness of component technologies. This requires data sets that are annotated with desired outputs for end-to-end tasks (text retrieval and data mining, in particular) and, selectively, annotated for intermediate analyses (optical character recognition, document structure analysis, signature matching, authorship attribution, etc.), as well.

A good test collection should cover the richness of inputs CDIP faces: a range of document formats, structures, lengths, and genres, manifested with varying print and imaging quality. Documents should include handwritten text and notations, diverse fonts, and elements such as graphs, tables, photos, logos, and diagrams. The volume of documents, and the number of redundant or useless documents, should be large enough to stress the component technologies and the system as a whole. Finally, the data in the collection should be publicly available to researchers with minimal costs and licensing restrictions.

Existing document image collections are lacking in many of these dimensions, and this has hampered CDIP research. Consider the main meeting at the intersection of document analysis and information retrieval: the yearly IS&T/SPIE *Document Recognition and Retrieval* conference. Of the 170 papers presented at this conference between 2001 and 2006 only four contain effectiveness results from text retrieval experiments. No two of these studies use the same test collection, none of the papers indicate how to access their collection, the documents are largely homogeneous in genre and other characteristics, and the largest collection contains only 3000 documents. While Taghva and colleagues have conducted many larger studies [8], their data have been tied up with legal issues and are not publicly available.

2. THE TOBACCO DOCUMENTS
We are building a new test collection, the _____ Complex Document Information Processing Test Collection, to support the diverse needs of CDIP research, both in our project and in the IR and document analysis communities at large. Our collection is based on the MSA (Master Settlement Agreement) documents from the Legacy Tobacco Documents Library (LTDL), created and hosted by the University of California San Francisco (UCSF) [4,7]. These approximately seven million documents (roughly 40 million scanned pages in TIFF format) became public through legal proceedings against five US tobacco companies and two tobacco industry research institutes. The documents were scanned by the tobacco industry using diverse technologies.

Besides having the size and diversity necessary for CDIP research, the MSA documents have two unusual advantages over other materials we considered. The first is an active research community. Hundreds of peer-reviewed papers have been published using documents from LTDL and related sources [3], and the US National Cancer Institute has funded research using the documents [6].

The second advantage is the LTDL metadata records, one for each document. The tobacco industry created these records based on examination of the original paper documents, so they represent a huge amount of manual labor. While the records are of highly variable structure and quality (despite UCSF's normalization efforts), they do mean that every document has some retrievable content, even those whose images are beyond the capabilities of current document analysis technology.

We obtained a snapshot of the LTDL TIFF files, metadata, and optical character recognition (OCR) output from UCSF. The total size of the data set is about 1.5 terabytes. Figure 1 shows a typical document image, portions of its metadata, and a few words at the start of its OCR. The document shows several of the challenges typical in CDIP, including multiple fonts, poor reproduction quality, and important information in handwritten annotations.

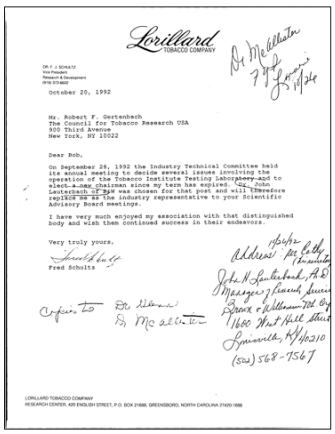
We are currently doing additional cleanup and formatting of the XML records in preparation for distribution to TREC 2006 participants (see below). Total size of the OCR and metadata is about 50 GB, making it a moderately large collection by current IR standards, but relatively easy to distribute. TREC 2006 will not use the 1.5 TB of TIFF files, and we are still investigating how to efficiently and cost-effectively distribute these files to researchers. The fact that the MSA [2] requires the public availability of the documents simplifies the legal issues in distributing the data, though some special treatment of material for which the tobacco industry organizations did not hold copyright is necessary, as mentioned at LTDL [5].

3. TASK-SPECIFIC DATA

Text retrieval experiments require not just documents, but queries and relevance judgments. We are pursuing three avenues for producing these. First, our revised MSA XML records will be used in the TREC 2006 Legal track [9,10]. Queries will simulate requests for document production of the sort that occur in legal cases, and relevance judgments will be produced by judging a pool of top retrieval results from diverse participant systems, as usual in TREC.

Second, we are working with tobacco document researchers to produce topics corresponding to their actual information needs. For example, Professor Robbin Derry of Northwestern University has collected documents on several topics relevant to teaching business ethics. The large numbers of documents already found by researchers will form our initial relevance judgments, followed by relevance feedback and further judging by tobacco experts.

Third, we are creating known item queries that seek particular documents. By choosing appropriate documents, we can more directly measure the impact on retrieval effectiveness of particular component technologies, e.g., signature recognition or OCR. Interestingly, known item queries are of intense interest to the tobacco document research community, which has often struggled


Lorillard
TOBACCO COMPANY

October 20, 1992

Mr. Robert F. Gertenbach
The Council for Tobacco Research USA
900 Third Avenue
New York, NY 10022

Dear Bob,

On September 28, 1992 the Industry Technical Committee held its annual meeting to decide several issues involving the operation of the Tobacco Institute Testing Laboratory and to elect a new chairman since my term has expired. Dr. John Lauterbach of BW was chosen for that post and will therefore replace me as the industry representative to your Scientific Advisory Board meetings.

I have very much enjoyed my association with that distinguished body and wish them continued success in their endeavors.

Very truly yours,

Fred Schultz

<LTDLWOCR>
<tid>rtv20a00</tid>
<bt>60033115/3115</bt>
<dd>19991020</dd>
<dt>LETTER</dt>
<au>SCHULTZ FJ, LOR</au>
<rc>GERTENBACH RF,CTR</rc>
<np>CATHY; GLENN; LAUTERBACH JH, BW; LORRAINE; MCALLISTER</np>
<np>TOBACCO INST TESTING LABORATORY; SAB; ITC</np>
<ot>ZTOBACCO COMPANY...
</LTDLWOCR>

Figure 1: A document plus selected metadata and OCR.

to find a particular important document known only through an indirect mention elsewhere. Indeed, while not of use for conventional IR experiments, we plan to create "unknown item queries" for documents of interest to scholars that are believed to exist in the MSA documents, but have not yet been found.

We intend the _____ CDIP Test Collection to support research in areas beyond text retrieval as well. One goal in our metadata cleanup work is to improve the usefulness of the data for social network analysis and other data mining tasks. Component tasks are also of interest. To support work on signature recognition, we segmented within document images 10 or more examples of the signatures of 66 distinct people. We are also developing data sets for OCR, logo recognition, and other tasks. We welcome feedback and suggestions on how to maximize the usefulness of the _____ CDIP Test Collection.

4. REFERENCES

[1] _____, et al. A Prototype System for Complex Document Information Processing. Submitted to SIGIR '06.

[2] *http://caag.state.ca.us/tobacco/resources/msasumm.htm*

[3] Hirschhorn, N. Research Reports and Publications Based on Tobacco Industry Documents, 1991-2006. Manuscript, January 2006.

[4] *http://legacy.library.ucsf.edu/*

[5] *http://legacy.library.ucsf.edu/legal.html*

[6] National Cancer Institute. Review and Analysis of Tobacco Industry Documents. Program Announcement PAR 01-063, March 7, 2001. *http://grants2.nih.gov/grants/guide/pa-files/PAR-01-063.html*

[7] Schmidt, H.; Butter, K.; and Rider, C. Building Digital Tobacco Document Libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D-Lib Magazine*, 8, 2 (2002).

[8] Taghva, K.; Borsack, J.; Condit, A.; Evra, S. The Effects of Noisy Data on Text Retrieval. JASIS 45(1), pp. 50-58 (1994).

[9] http://trec.nist.gov

[10] http://trec-legal.umiacs.umd.edu/