Content-Based Document Image Retrieval in Complex Document Collections

G. Agam[a], S. Argamon[a], O. Frieder[a], D. Grossman[a], D. Lewis[b]

[a]Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616
[b]David D. Lewis Consulting, 858 W. Armitage Ave., #296 Chicago, IL 60614

ABSTRACT

We address the problem of content-based image retrieval in the context of complex document images. Complex document are documents that typically start out on paper and are then electronically scanned. These documents have rich internal structure and might only be available in image form. Additionally, they may have been produced by a combination of printing technologies (or by handwriting); and include diagrams, graphics, tables and other non-textual elements. Large collections of such complex documents are commonly found in legal and security investigations. The indexing and analysis of large document collections is currently limited to textual features based OCR data and ignore the structural context of the document as well as important non-textual elements such as signatures, logos, stamps, tables, diagrams, and images. Handwritten comments are also normally ignored due to the inherent complexity of offline handwriting recognition. We address important research issues concerning content-based document image retrieval and describe a prototype for integrated retrieval and aggregation of diverse information contained in scanned paper documents we are developing. Such complex document information processing combines several forms of image processing together with textual/linguistic processing to enable effective analysis of complex document collections, a necessity for a wide range of applications. Our prototype automatically generates rich metadata about a complex document and then applies query tools to integrate the metadata with text search. To ensure a thorough evaluation of the effectiveness of our prototype, we are developing a test collection containing millions of document images. This is in contrast to existing datasets for content-based image retrieval which normally contain only thousands of images. We believe that the formation of such large dataset is essential in understanding the problems associated with realistic applications.

Keywords: content-based image retrieval (CBIR), content-based document image retrieval (CBDR), complex document information processing, document image analysis, document image understanding, document recognition, text retrieval, test collections

1. INTRODUCTION

Complex Document Information Processing (CDIP)[1] involves the analysis of large masses of scanned paper documents, which often contain non-textual information such as handwriting, logos, signatures, figures, and tables. It is a critical problem in many key application areas, including litigation, intelligence analysis, knowledge management, and humanities scholarship. However, while specific solutions do exist for component problems such as OCR, handwriting analysis, logo recognition, and signature matching, no system to date has integrated extraction and analysis methods to enable users to pose queries that integrate these different forms of document information. Current practice, therefore, is to use separate processing systems independently, so that collating and cross-checking information items must be done by hand, underscoring the need for integrated systems. Furthermore, results can be degraded when individual processing modules are unaware of the larger context in which they operate; an integrated system will likely do better.

From an image analysis point of view the problem of indexing and searching large document image collections can be classified as a subproblem of the general content-based image retrieval (CBIR) problem. Nevertheless, there are several important characteristics to content-based document image retrieval (CBDR) problem that do not exist in the general CBIR context. Primarily, document images contain text and thus can be indexed and searched based on noisy OCR-ed textual data. Secondly, in contrast to the state of the art in CBIR in which datasets of thousands of images are being processed, a realistic CBDR search involves datasets of millions of documents.

In addition to textual elements, non-textual elements in a document contain important information that can be used for indexing. For example, signatures can be used for indexing document authorship, logos can be used for indexing document organizational unit, tables can be used for indexing document importance, and handwritten comments can be used to index document readership. Moreover, document layout analysis can be used to identify specific zones such as address or date zones, which can then be used to constrain interpretation and so improve OCR performance.

Given that both the textual and non-textual analysis of document images are inherently error prone, the effectiveness of these analyses can be improved by simultaneously solving both problems. For example, recognizing a name in a document through OCR can be used to improve offline signature recognition by constraining the set of candidates. Conversely, authorship determination through signature recognition can be used to assist OCR of names in a document. Similarly, recognizing an organizational unit through logo analysis can be used to assist authorship determination through signature recognition, and signature recognition can be simplified based on the recognition of an organizational unit through logo analysis. It is mainly in this synergy that the quality of both textual and non-textual information interpretation in document images can be improved to support the effectiveness of CBDR.

We describe a first research prototype for integrated Complex Document Information Processing (CDIP) we are developing to facilitate CBDR, and provide the details of a public test collection of millions of document images we are preparing to support CBDR evaluations. The prototype currently contains modules that extract and analyze text, signatures, and logos from complex documents, enabling integrated document image retrieval. The system contains a set of ingestion modules, which process different forms of document image data, importing them into a database schema which includes traditional document indices by keyword and relations between documents and their components (logos, signatures, and named entities). Retrieval operates via SQL queries on this unified database.

The remainder of the paper describes our current development effort in more detail. Section 2 describes the architecture of the research prototype we are developing. Section 3 details the test collections and evaluation procedures we are developing. Section 4 presents some preliminary results. Section 5 concludes the paper.

2. CDIP SYSTEM DESIGN

This section describes the CDIP research prototype we are developing. The prototype currently contains modules that extract and analyze text, signatures, and logos from complex documents, enabling integrated document image retrieval. The system contains a set of ingestion modules, which process different forms of document image data, importing them into a database schema which includes traditional document indices by keyword and relations between documents and their components (logos, signatures, and named entities). Retrieval operates via SQL queries on this unified database.

2.1. Functional components

Our prototype comprises an integrated tool suite, based on several existing technologies, implementing three core CDIP functionalities: *document image analysis*, *named-entity recognition*, and *integrated retrieval*. This prototype tool will facilitate later the inclusion of a fourth core technology: *data mining*. As noted, specific attention is being paid to modular design, to ensure that the developed software modules will easily integrate into different task-level applications.

Document image analysis extracts information from raster scanned images such as the overall structure of the document,[2] the content of text regions,[3] the location of images/graphics, the location of logos and signatures, the location of signatures and handwritten comments,[4] and the identification of signatures.[5–7] It should be noted that OCR of machine printed text in real-world documents has limited accuracy (depending on the quality of the input documents) and so the textual features obtained are unavoidably noisy.

Named-entity recognition identifies meaningful entities such as people and organizations in textual components. Our prototype relies on ClearForest technology provided by Text Solutions. It is interesting to note that initial tests on real-world data show that the the effectiveness of entity extraction on noisy text obtained from OCR of a test collection is reduced to 70% of its performance on noise-free text.

Integrated retrieval from different kinds of data sources is the key high-level function. Such integrated retrieval is possible through the IIT Intranet Mediator technology .[8,9] The IIT Intranet Mediator is capable of integrating traditional data sources such as unstructured text, semi-structured XML/text data, as well as structured database querying. A rule-based source selection algorithm selects those data sources most relevant to an information request, enabling the system to take full advantage of domain-specific searching techniques, such as translation of a natural language request into a structured SQL query. Results are then fused into an integrated retrieval set.[10] Although the IIT Mediator is protected by an issued patent providing us with guaranteed unconstrained free use the technology, the mediator implementation technology that exists is only at the prototype level. Consequently, as we need a more robust framework by which to implement our CDIP prototype, we have built our prototype using the Claraview integration fabric.

Data mining which is not implemented in our current prototype, will be able to leverage text, metadata, and information extracted from complex documents. Our approach allows application of traditional data mining and machine learning methods to discover relationships between different data such as association rules[11] and document clusters.[12] We will further develop routines to find correlations in document descriptors (for example, possible relationships between the author of a document and particular language styles). Note that data mining is not targeted in our initial implementation of the system prototype but is a goal for follow-on efforts.

2.2. Software Architecture

The prototype's architecture is designed as a generic framework for integrating component technologies with appropriate APIs and data format standards through SOAP (Simple Object Access Protocol) to allow 'plugging in' different subsystems for performing component tasks. Our current effort integrates available components with little emphasis on the development of new ones.

The current system architecture is depicted in Figure 1. The workflow of the system consists of three main processes: a document *ingestion* process, a data *transition* process, and a document *querying* process. The document ingestion process is a straightforward pipeline that consists of:

- Low-level image processing for noise removal, skew-correction, orientation determination, and document and text regions zoning (using Abbyy's SDK[3] and the DocLib package[2]).
- OCR in text regions (using Abbyy's SDK[3]), recognition of logos (using the DocLib package[2]), and recognition of signatures (using CEDAR's signature recognition system[5,4,6]) and a signature warping module.[7]
- Linguistic and classification analysis of extracted information for annotation in the database: entity tagging, relationship tagging, and stylistic tagging in text regions (using Text Solutions).

At the end of the ingestion process, we have an operational data store in 3rd normal form (3NF). At this point, it would be complex to perform sophisticated roll-up or drill-down computations along various data dimensions. Hence, we transition the data from 3NF that has been ingested into a multidimensional star schema. This is a very common technique for analyzing structured data, and it is well known to dramatically improve decision support. Using this structure for complex document metadata results in a scalable query tool that can quickly answer questions like "How many documents do we have from Fortune 500 companies" and then quickly drill into different market sectors (e.g., manufacturing companies, IT companies, etc.)

At the center of this process are tools from Claraview. These tools use web services to access the point solutions and identify metadata about complex documents to populate the 3NF schema. Claraview tools also migrate the 3NF schema to a star schema using well known extract, transform, and load processing. Claraview is a startup dedicated to the application of well known structured data techniques such as a star schema and applying these to integrate structured data and text. As the analysis of document images involves errors which are inherent to the automated interpretation process, each attribute in the database is associated with a probability that indicates the confidence in this value as obtained from the corresponding point solution. Finally, following the ETL process, a query tool is used to access both an inverted index of all text and the star schema to integrate structured results.

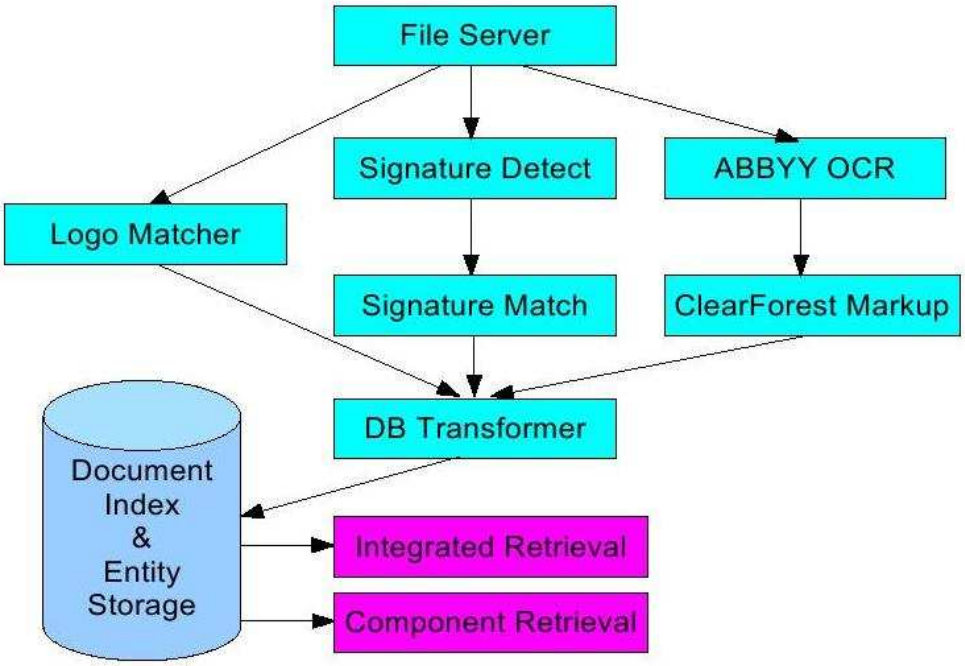


Figure 1. Architectural overview of the current CDIP research prototype.

A key component that is facilitated by our approach is a tight integration of the processes of document image interpretation, symbol extraction and relation, and information retrieval. This integrated approach could be used to increase reliability for all of these processes. Constraints on image interpretation, based on consistency with other data, can improve reliability of image interpretation. Similarly, gaps in the database can potentially be filled in at retrieval time, by reinterpreting image data using top-down expectations based on user queries. Due to its added complexity, this tight integration model is not followed in our current implementation of the system prototype.

A summary of the CDIP architecture is presented in Figure 1. Each component in this figure is a separate thread, so that processing is fully parallelized and pipelined. Image files are served to processing modules dealing different types of document image information. The Abbyy OCR engine is used to extract text from the document image. This text is fed to the ClearForest information extraction module, which finds and classifies various named entities and relations. Signatures are segmented and then fed to CEDAR's signature recognition system which matches document signatures to known signatures in a database. Logos are segmented and matched using the DocLib package. These three threaded processing paths are then synchronized, and the data extracted are transformed into a unified database schema for retrieval and analysis.

3. DATASETS AND EVALUATION

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 the CDIP project 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 as optical character recognition (OCR), document structure analysis, signature matching, and authorship attribution. One goal of our project is to evaluate the effectiveness of the CDIP research prototype and how that effectiveness

responds to changes in the effectiveness of component technologies. This requires datasets that are annotated with desired outputs for end-to-end tasks 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,[13] their data have been tied up with legal issues and are not publicly available.

3.1. The Tobacco Documents

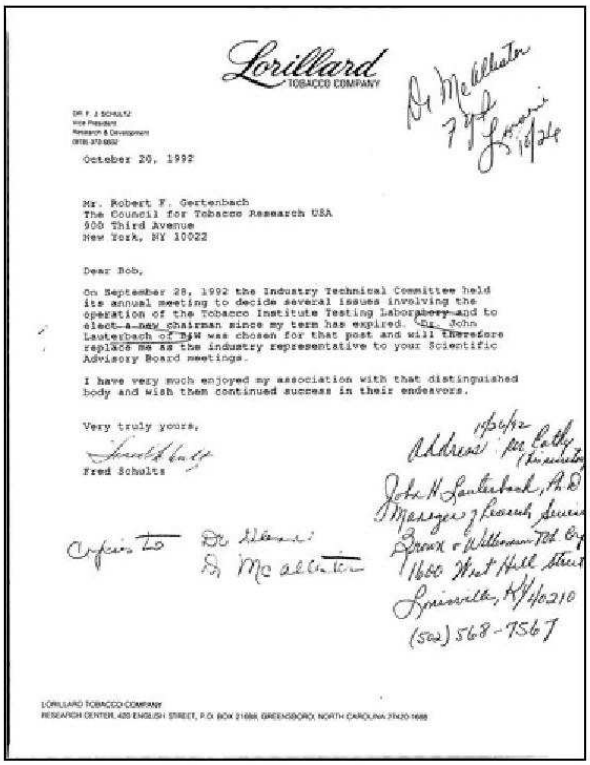
We are building a new test collection, the IIT Complex Document Information Processing Test Collection, to support the diverse needs of CDIP research, both in our project and in the IR and document analysis communities at large. Our collection is based on the MSA (Master Settlement Agreement) documents from the Legacy Tobacco Documents Library (LTDL), created and hosted by the University of California San Francisco (UCSF).[14,15] 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,[16] and the US National Cancer Institute has funded research using the documents.[17]

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 2 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 have cleaned up and combined two versions of the UCSF metadata (containing slightly different information) plus the UCSF-produced OCR to produce XML records for 6,878,327 documents in the IIT CDIP Test Collection, Version 1.0. The records are bundled in 650 files, totaling approximately 62GB. This data has been made available by FTP and DVD to TREC 2006 participants (see below). 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[18] 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.[19]



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 Lautenbach of BAT 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

LORILLARD TOBACCO COMPANY
RESEARCH CENTER, 420 ENGLISH STREET, P.O. BOX 21688, GREENSBORO, NORTH CAROLINA 27420-1688

```
<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 2. A document plus selected metadata and OCR.

3.2. Task-Specific Data

Text retrieval experiments require, in addition to documents, both queries and relevance judgments. We are pursuing three avenues for producing these. First, the IIT CDIP Test Collection, Version 1.0 XML records are being used in the TREC 2006 Legal track.[20,21] A total of 46 queries simulate requests for document production of the sort that occur in legal cases. Relevance judgments will be produced by judging pooled retrieval results from diverse participant systems, as is usual in TREC. To increase the diversity of the judged document set, we have contracted with a tobacco document information specialist to do manual searches on these queries as well.

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 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 IIT 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

datasets for OCR, logo recognition, and other tasks. We welcome feedback and suggestions on how to maximize the usefulness of the IIT CDIP Test Collection.

4. PRELIMINARY RESULTS

The rich collection of attributes our system associates with each document (including words, linguistic entities such as names and amounts, logos, and signatures) enables both novel forms of text retrieval, and the evidence combining capabilities of a relational database.

We strongly believe that quantitative effectiveness evaluation, while difficult to perform, is essential in this project in several ways. Primarily quantitative effectiveness can be used to roughly asses the expected performance of the system. While this measure is of course data dependent, it can be used to compare the performance of systems should such become available in the future. Secondly, quantitative effectiveness evaluation is essential in measuring improvements within our system. Finally, such a quantitative evaluation will allow us to study the effectiveness of the the overall system as a function of the effectiveness of its individual components.

We have finished the initial implementation of our research prototype and are currently in the process of evaluating it quantitatively. A snapshot of the system configuration screen is presented in Figure 3. The evaluation includes using a subset of several hundred document images which were manually labeled for authorship (based on signatures), organizational unit (based on logos), and various entity tags based on textual information (such as monetary amounts, dates, and addresses). The evaluated tasks include authorship-based, organizational-based, monetary-based, date-based, and address-based document image retrieval, whereas in each experiment the precision and recall is recorded as a function of a decision threshold. This experiment is expected to be expanded in the near future to include a larger subset of several thousand document images. We realize that this testing methodology cannot be extended to higher order subsets as it requires complete manual labeling which is labor intensive. Consequently, effectiveness using larger subsets will be be evaluated by inserting document images containing unique labels into large subsets. These inserted documents will be manually labeled and their uniqueness will guarantee that documents with similar labels should not exist within the subset.

While we have, as yet, no quantitative evaluations to report, we give examples here of the kinds of capabilities that our prototype currently supports. The mini-corpus used for this consists of 800 documents taken from the testbed we are building. We consider integrated queries that our prototype makes possible for the first time. We apply conjunctive constraints on document image components to a straightforward document ranking based on total query-word frequency in the OCRed document text; in Figure 4 we show document images retrieved for two such queries. The first is the unique document found containing both of the words "income forecast" as well as the American Tobacco Company logo and a dollar amount (a recognized entity type) greater than $500K. The second example is the top-ranked document for "filtration efficiency" that also has the R.J. Reynolds logo and a signature. Note that neither of these documents would have been found just based on their printed text, as neither contains the company name explicitly. In Figure 5 we show a ranked retrieval results for a document component query which asks for the five signatures with the highest total of dollar amounts mentioned in documents with each signature. This shows another novel way of integrating useful information extracted from document images which is easily implementable in our framework.

Several examples that demonstrate the ability of the CDIP system to associate both textual and non-textual data are presented below. Figure 6 demonstrates the ability of the CDIP system to identify person association in a document collection. The left column lists the search person. Subsequent columns show associated dates, company logos, associated persons, and dollar amounts. For example, Dr. D. Stone was active during 1986, was associated with a company whose logo template is "liggett.tif", was associated with dollar amounts between $140K and $1.68M, and was associated with several other persons such as Dr. Calabrese. By clicking on the document ID, the system presents the user with the original documents for full examination. A similar example which demonstrates the ability of the CDIP system to associate persons with organizations is presented in Figure 7.

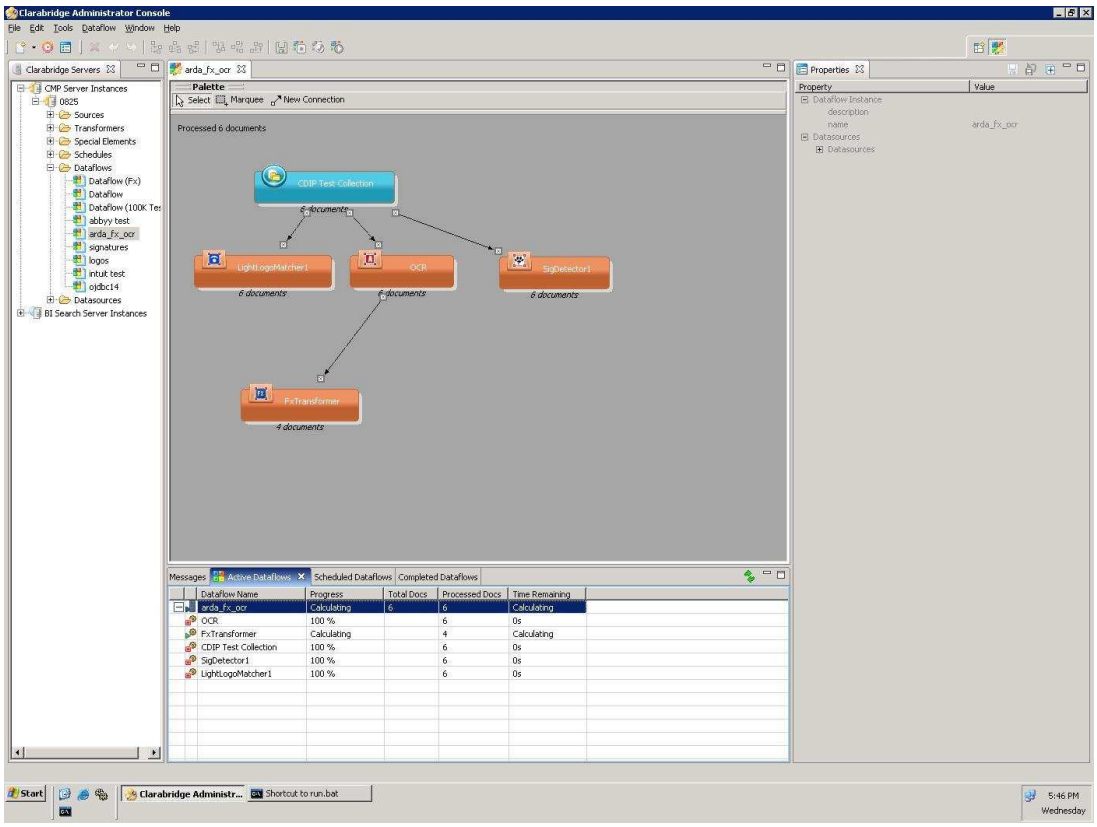


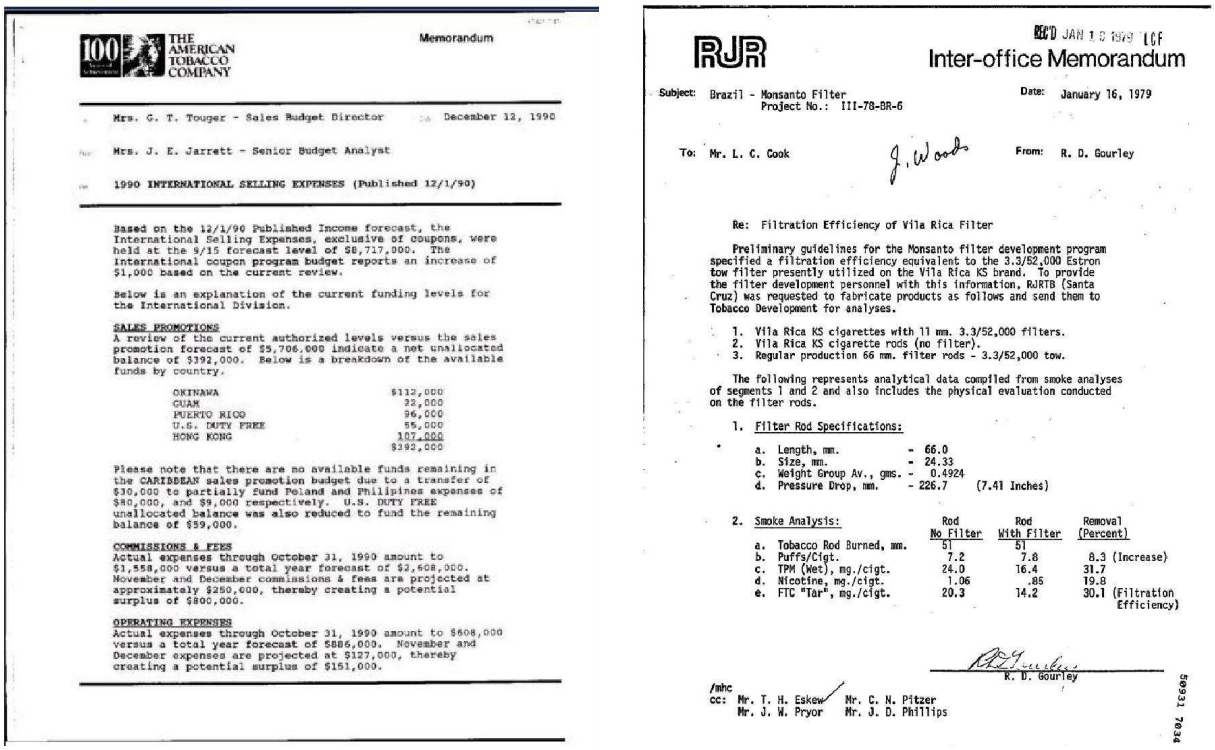
Figure 3. A snapshot of the CDIP system configuration screen.

5. SUMMARY

Content-based document image retrieval in complex document datasets is of increasing importance, and yet, little work has been done on developing integrated solutions to this problem. We have described the research prototype we are developing which unifies various component technologies and which is expected to give a clear roadmap for future research in this area. It is already clear that a key focus will need to be the estimation, propagation, and use of reliability estimates for data extracted and processed from document images; this information flow will likely need to be both bottom-up (from ingestion processes) and top-down (in the form of expectations). We expect to flesh out this important issue more completely once we have a complex document test collection to systematically evaluate the performance of the system and its various components. As described, such a testbed is in development, based on the large collection of document images available in the Legacy Tobacco Documents Library; the many researchers interested in using this material makes it a prime candidate for a useful and feasible testbed for CDIP. We provide preliminary results to demonstrate the performance of our system. Generally, we expect that the prototype and test collection that we are developing will help in steering constructive research efforts toward the solution of this complex problem.

Acknowledgments

This work is supported in part by a Challenge Workshop grant from ARDA. We thank and acknowledge all the participants in the March 2005 ARDA CDIP Challenge Kickoff Workshop for their suggestions and help with our many follow-up questions. We also greatly thank our collaborators from Claraview, TSI, and the CEDAR and DocLib efforts. A large number of people have provided resources, information, and suggestions on this work, and we likewise thank them. We in particular acknowledge David Roberts for his guidance; Karen Butter, Albert Jew, Kirsten Neilsen, and Heidi Schmidt of the University of California San Francisco Library Center for



THE AMERICAN TOBACCO COMPANY

Memorandum

Mrs. G. T. Touger - Sales Budget Director — December 12, 1990

Mrs. J. E. Jarrett - Senior Budget Analyst

1990 INTERNATIONAL SELLING EXPENSES (Published 12/1/90)

Based on the 12/1/90 Published Income forecast, the International Selling Expenses, exclusive of coupons, were held at the 9/15 forecast level of $8,717,000. The international coupon program budget reports an increase of $1,000 based on the current review.

Below is an explanation of the current funding levels for the International Division.

SALES PROMOTIONS

A review of the current authorized levels versus the sales promotion forecast of $5,706,000 indicate a net unallocated balance of $392,000. Below is a breakdown of the available funds by country.

OKINAWA	$112,000
GUAM	22,000
PUERTO RICO	96,000
U.S. DUTY FREE	55,000
HONG KONG	107,000
	$392,000

Please note that there are no available funds remaining in the CARIBBEAN sales promotion budget due to a transfer of $30,000 to partially fund Peland and Philipines expenses of $80,000, and $9,000 respectively. U.S. DUTY FREE unallocated balance was also reduced to fund the remaining balance of $59,000.

COMMISSIONS & FEES

Actual expenses through October 31, 1990 amount to $1,558,000 versus a total year forecast of $2,608,000. November and December commissions & fees are projected at approximately $250,000, thereby creating a potential surplus of $800,000.

OPERATING EXPENSES

Actual expenses through October 31, 1990 amount to $608,000 versus a total year forecast of $886,000. November and December expenses are projected at $127,000, thereby creating a potential surplus of $151,000.

(a)

RJR

Inter-office Memorandum

Subject: Brazil - Monsanto Filter
Project No.: III-78-BR-6

Date: January 16, 1979

To: Mr. L. C. Cook

From: R. D. Gourley

Re: Filtration Efficiency of Vila Rica Filter

Preliminary guidelines for the Monsanto filter development program specified a filtration efficiency equivalent to the 3.3/52,000 Estron tow filter presently utilized on the Vila Rica KS brand. To provide the filter development personnel with this information, RJRTB (Santa Cruz) was requested to fabricate products as follows and send them to Tobacco Development for analyses.

1. Vila Rica KS cigarettes with 11 mm. 3.3/52,000 filters.
2. Vila Rica KS cigarette rods (no filter).
3. Regular production 66 mm. filter rods - 3.3/52,000 tow.

The following represents analytical data compiled from smoke analyses of segments 1 and 2 and also includes the physical evaluation conducted on the filter rods.

1. Filter Rod Specifications:

   a. Length, mm. - 66.0
   b. Size, mm. - 24.33
   c. Weight Group Av., gms. - 0.4924
   d. Pressure Drop, mm. - 226.7 (7.41 Inches)

2. Smoke Analysis:

	Rod No Filter	Rod With Filter	Removal (Percent)
a. Tobacco Rod Burned, mm.	51	51	
b. Puffs/Cigt.	7.2	7.8	8.3 (Increase)
c. TPM (Wet), mg./cigt.	24.0	16.4	31.7
d. Nicotine, mg./cigt.	1.06	.85	19.8
e. FTC "Tar", mg./cigt.	20.3	14.2	30.1 (Filtration Efficiency)

R. D. Gourley

/mhc
cc: Mr. T. H. Eskew Mr. C. N. Pitzer
Mr. J. W. Pryor Mr. J. D. Phillips

(b)

Figure 4. Retrieval results. **(a)** With ATC logo, the words "income forecast", and mention of than $500,000. **(b)** With R.J.Reynolds logo, the words "filtration efficiency", and a signature.

Signature Group	Me	Ent Dollars
	Gertenbach, RF	$37,454,447.88
	Schan, M	$30,885,327.00
	Boffa, JR	$17,420,705.00
	Nielsen, VG	$958,354.82
	Bergman, JI	$635,397.44

Figure 5. Prototype results showing signatures associated with the most total dollars (see text).

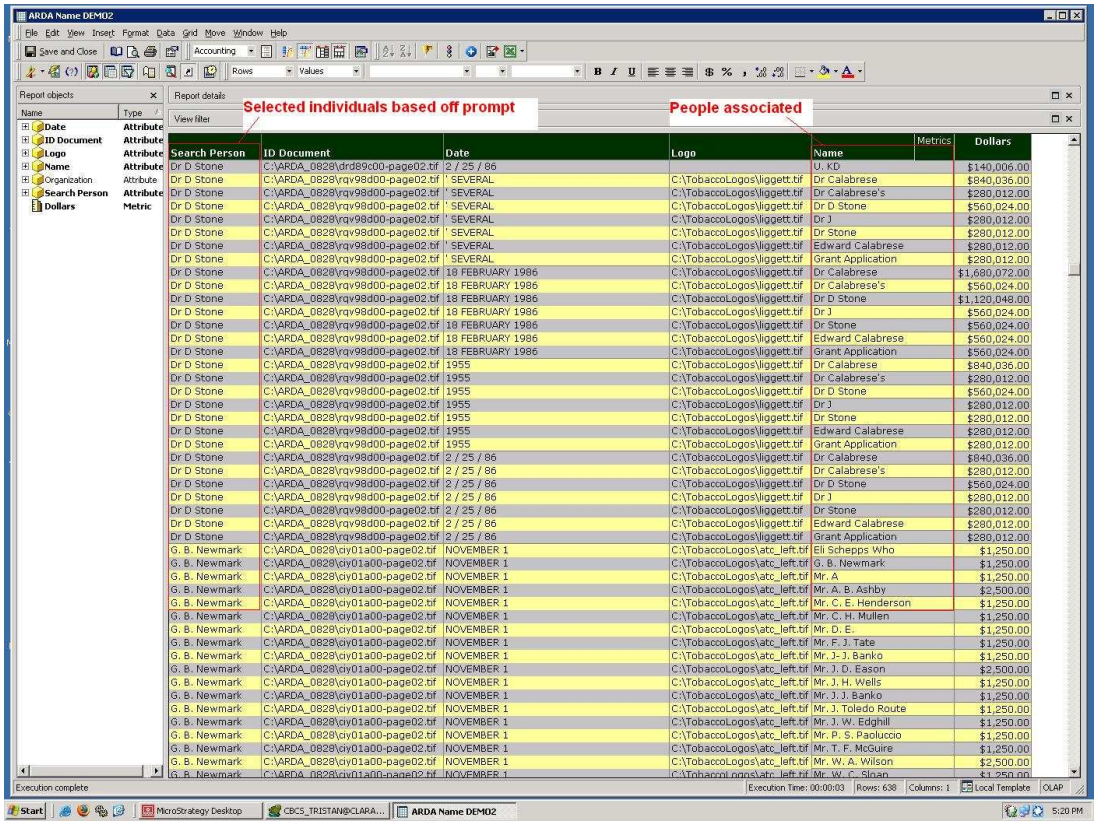


Figure 6. Example of the ability of the CDIP system to identify person association in a document collection based on both textual and non-textual data. The left column lists the search person. Subsequent columns show associated dates, company logos, associated persons, and dollar amounts.

Knowledge Management for providing access to and help with the Legacy Tobacco Documents Library; Michael Tacelosky and Keith Ivey of Tobacco Documents Online and smokefree.net for access to and help with their collection of tobacco documents; Donald Rubin and Cati Brown of the University of Georgia for access to and help with their collection of transcribed tobacco documents; and Robbin Derry of Northwestern University for access to and help with her research materials on the tobacco documents.

REFERENCES

1. G. Agam, S. Argamon, O. Frieder, D. Grossman, and D. Lewis, "Complex document information processing: System, test collection, and evaluation," in *Document Recognition and Retrieval XIII*, K. Taghva and X. Lin, eds., *Proc. SPIE* **6067**, pp. 0N–1 – 0N–11, 2006. Invited paper.
2. K. Chen, T. Drayer, L. Hernandez, S. Jaeger, S. Sampat, G. Zhu, and D. Doermann, "Doclib: A document processing research tool," in *Proc. Symp. Document Image Understanding Technology*, (Adelphi, Maryland), November 2005.
3. "Abbyy's SDK for OCR and document structure analysis." http://www.abbyy.com/.
4. S. N. Srihari, C. Huang, and H. Srinivasan, "A search engine for handwritten documents," in *Proc. Document Recognition and Retrieval XII*, pp. 66–75, SPIE, (San Jose, CA), January 2005.
5. S. Chen and S. N. Srihari, "Use of exterior contours and word shape in off-line signature verification," in *Proc. Intl. Conference on Document Analysis and Recognition*, pp. 1280–1284, (Seoul, Korea), August 2005.
6. S. N. Srihari, S. Shetty, S. Chen, H. Srinivasan, C. Huang, G. Agam, and O. Frieder, "Document image retrieval using signatures as queries," in *IEEE Intl. Conf. on Document Image Analysis for Libraries (DIAL)*, pp. 198–203, 2006.

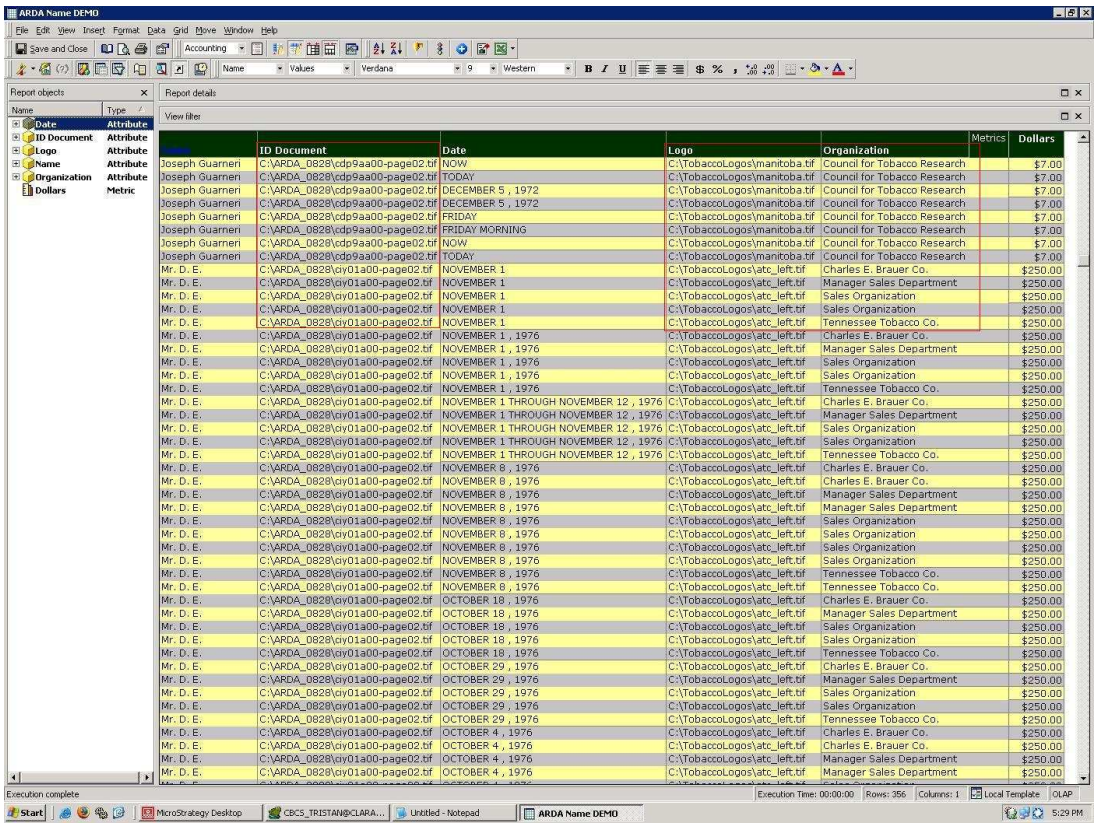


Figure 7. Example of the ability of the CDIP system to identify organizational association in a document collection based on both textual and non-textual data. The left column lists the search person. Subsequent columns show associated dates, company logos, associated organizations, and dollar amounts.

7. G. Agam and S. Suresh, "Particle dynamics warping approach for offline signature recognition," in *IEEE Workshop on Biometrics (part of CVPR 2006)*, pp. 38–44, (New York, NY), 2006.
8. D. Grossman, S. Beitzel, E. Jensen, and O.Frieder, "IIT Intranet Mediator: Bringing data together on a corporate intranet," *IEEE IT Professional* **4**(1), pp. 49–54, 2002.
9. J. Heard, J. Wilberding, G. Frieder, O. Frieder, D. Grossman, and L. Kane, "On a mediated search of the united states holocaust memorial museum data," in *Sixth Next Generation Information Technology Systems*, (Sefayim, Israel), July 2006.
10. S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, O. Frieder, and N. Goharian, "On fusion of effective retrieval strategies in the same information retrieval system," *Journal of the American Society of Information Science and Technology* **55**(10), 2004.
11. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in *Advances in knowledge discovery and data mining*, pp. 307–328, American Association for Artificial Intelligence, 1996.
12. M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. SIGKDD Workshop on Text Mining*, 2000.
13. K. Taghva, J. Borsack, A. Condit, and S. Evra, "The effects of noisy data on text retrieval," *Journal of the American Society for Information Science and Technology* **45**(1), pp. 50–58, 1994.
14. UCSF, "Legacy tobacco documents library," 2005. http://legacy.library.ucsf.edu/.
15. H. Schmidt, K. Butter, and C. Rider, "Building digital tobacco industry document libraries at the university of california, san francisco library/center for knowledge management," *D-Lib Magazine* **8**(2), 2002.

16. N. Hirschhorn, "Research reports and publications based on tobacco industry documents, 1991-2005," May 2005.
17. "Review and analysis of tobacco industry documents," June 1999. National Cancer Institute Program Announcement.
18. California Office of the Attorney General, "Tobacco master settlement agreement summary," 1999. http://caag.state.ca.us/tobacco/resources/msasumm.htm.
19. UCSF, "Legacy tobacco documents library," 2005. http://legacy.library.ucsf.edu/legal.html.
20. National Institute of Standards and Technology, "Text retrieval conference (trec)," 2006. http://trec.nist.gov.
21. National Institute of Standards and Technology, "Trec 2006 legal track," 2006. http://trec-legal.umiacs.umd.edu/.