A Complex Document Information Processing Prototype

S. Argamon[1], G. Agam[1], O. Frieder[1], D. Grossman[1] D. Lewis[2] G. Sohn[3], K. Voorhees[3]
[1]Dept. of Computer Science, Illinois Institute of Technology, Chicago, IL, USA
[2]David D. Lewis Consulting, 858 W. Armitage Ave., #296, Chicago, IL, USA
[3]Clarabridge Inc., 3130 Fairview Park Drive, Falls Church, VA, USA

ABSTRACT

We developed a prototype for integrated retrieval and aggregation of diverse information contained in scanned paper documents. 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. This is the first system to attempt integrated retrieval from complex documents; we report its current capabilities.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

CDIP, Document image processing, Integrated retrieval

1. INTRODUCTION

Complex Document Information Processing (CDIP) involves the analysis of large masses of scanned paper documents, which often contain non-textual information such as handwriting, logos, or signatures. 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.

We describe a first research prototype for integrated CDIP, to enable work on these important research problems (see also [4]). Elsewhere [2] we describe our work on building a testbed collection for CDIP, based on document images from the Legacy Tobacco Documents Library (http://legacy.library.ucsf.edu/).

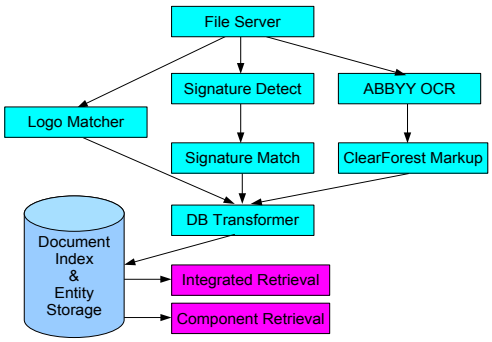


Figure 1: Architecture of the current prototype.

The prototype currently contains modules that extract and analyze text, signatures, and logos from complex documents, enabling integrated document image retrieval. The system (Fig. 1) 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.

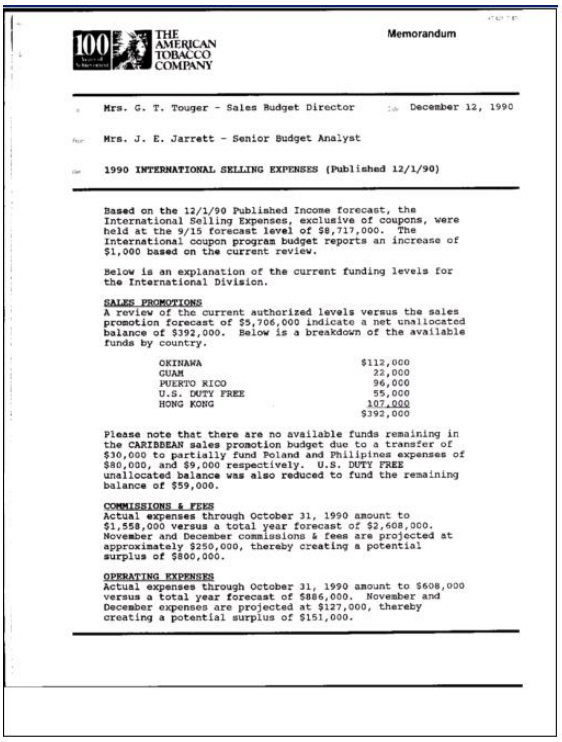
Each component in the figure is a separate thread, so that processing is fully parallelized and pipelined. Image files are served to processing modules dealing different types of document image information. (Future development will add preprocessing for noise removal, skew-correction, orientation determination, and region zoning.) The ABBYY OCR engine (www.abbyy.com) is used to extract text from the document image. This text is fed to the ClearForest (www.clearforest.com) information extraction module, which finds and classifies various named entities and relations. Signatures are segmented[1] and then fed to CEDAR's signature recognition system [3] which matches document signatures to known signatures in a database. Logos are segmented and matched using the DocLib package [1]. These three threaded processing paths are then synchronized, and the data extracted are transformed into a unified database schema for retrieval and analysis.

2. RETRIEVAL EXAMPLES

The rich collection of attributes our system associates with each document (including words, linguistic entities such as names and

[1]Signature segmentation is not yet integrated, and is currently simulated by hand-segmentation and labeling.

SIGIR '06 Seattle, Washington, USA



Memorandum

THE AMERICAN TOBACCO COMPANY

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.

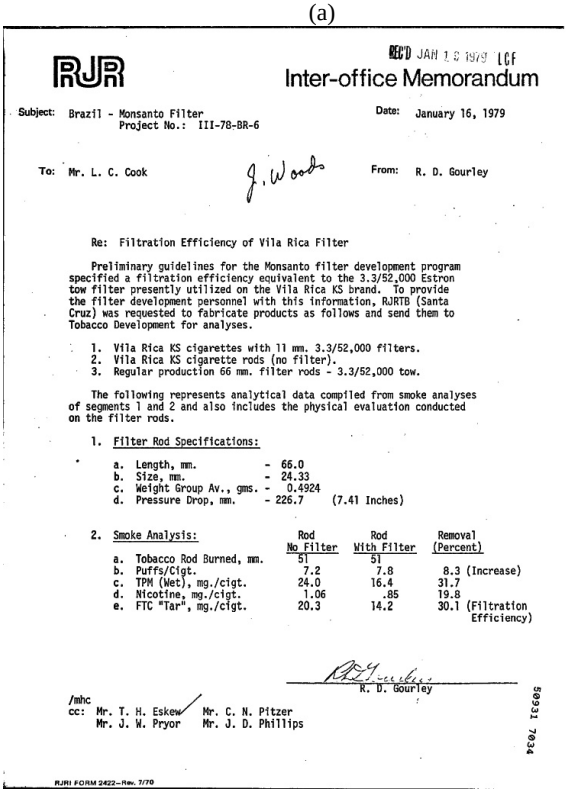
OKINAWA	$112,000
GUAM	22,000
PUERTO RICO	96,000
U.S. DUTY FREE	55,000
HONG KONG	107,000
	$392,000

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 Poland and Phillipines 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

RECD JAN 18 1979 LCF

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

RJRI FORM 2422--Rev. 7/70

(b)

Figure 2: 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 3: Prototype results showing signatures associated with the most total dollars (see text).

amounts, logos, and signatures) enables both novel forms of text retrieval, and the evidence combining capabilities of a relational database. 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 2 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 3 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.

Acknowledgment. Supported in part by an ARDA Challenge Grant.

3. REFERENCES

[1] K. Chen, S. Jaeger, G. Zhu, and D. Doermann. DOCLIB: A document processing research tool. In *SDIUT*, 2005.

[2] D. D. Lewis, S. Argamon, G. Agam, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *SIGIR-06*, 2006.

[3] S. N. Srihari, M. K. Kalera, and A. Xu. Offline signature verification and identification using distance statistics. *IJPRAI*, 18(7), 2004.

[4] S. S. Stein, S. Argamon, and O. Frieder. The effect of OCR on stylistic text classification. In *SIGIR-06*, 2006.