Security Informatics in Complex Documents

G. AGAM[1a], D. GROSSMAN[2a] and O. FRIEDER[3a,b]
[a]*Illinois Institute of Technology*
[b]*Georgetown University*

Abstract: Paper documents are routinely found in general litigation and criminal and terrorist investigations. The current state-of-the-art processing of these documents is to simply OCR them and search strictly the text. This ignores all handwriting, signatures, logos, images, watermarks, and any other non-text artifacts in a document. Technology, however, exists to extract key metadata from paper documents such as logos and signatures and match these against a set of known logos and signatures. We describe a prototype that moves beyond simply the OCR processing of paper documents and relies on additional documents artifacts rather than only on text in the search process. We also describe a benchmark developed for the evaluation of paper document search systems.

Keywords: Information Retrieval, Text Mining, Document Segmentation

Introduction

Investigating terrorism on the web requires the analysis of documents that are often complex. Automated analysis of complex documents is, therefore, a crucial component. Consider for example the following case:

On September 29, 2002, during an episode of *60 Minutes*, the reporter, Lesley Stahl, broadcast a story called, “The Arafat Papers.” During the story the following dialog occurred:

> STAHL: (Voiceover) The Israelis captured tens of thousands of documents when they bulldozed into Arafat's compound in Ramallah in March. Now the Palestinian Authority's most sensitive secrets are stacked in a sea of boxes in an Israeli army hangar.
>
> Colonel MIRI EISIN: It's basically all of their files, all of their documents, everything that we could take out.

Clearly, searching this document collection is an example of a problem that involves searching text, signatures, images, logos, watermarks, etc. While it is true that search companies such as Google^{TM} and Yahoo!^{TM} have made search technology a commodity, they fail to support the searching of complex documents. A complex document, or informally a "real world, paper document”, is one that comprises of not only text but also figures, signatures, stamps, watermarks, logos, and handwritten annotations. Furthermore, many of these documents are available in print form only. That is, the documents must first be scanned so as to be in digital format, and their scan (image) quality is often poor.

Searching complex documents (such as those found in the Arafat Papers), involves the integration of image processing techniques such as, but not limited to, image enhancement, layer separation, optical character recognition, and signature and logo detection and identification, as well as information retrieval techniques including relevance ranking, relevance feedback, data integration and style detection. To date no such system is available. Searching such a collection often involves discarding all document components other than text and then searching the text with a conventional search engine.

Yet another problem is evaluation. Currently, even if a complex document search system did exist, it is not possible to scientifically evaluate it. The impossibility of scientifically evaluating such a system is a direct consequence of the lack of an existing benchmark. Search systems are evaluated using benchmarks, e.g., using the various NIST TREC data sets (see trec.nist.gov for details), and the lack of benchmarks prevents any meaningful evaluation.

To advance the state of the art of search in terms of complex documents, our effort focused on the development of a complex document information processing prototype and evaluation benchmark, the IIT CDIP data set.

[1] G. Agam, Illinois Institute of Technology, 10 West 31st Street, Chicago, IL, 60610 E-mail: agam@iit.edu
[2] D. Grossman, Illinois Institute of Technology, 10 West 31st Street, Chicago, IL 60610 E-mail: grossman@iit.edu
[3] O. Frieder, Georgetown University, Washington, DC 20057 E-mail: ophir@cs.georgetown.edu; He is on leave from IIT.

1. The CDIP Collection

For a data set to be of lasting value, it must meet, challenge, and exceed application domains. These application require a collection that:

- Covers a richness of input in terms of a range of formats, lengths, and genres and variance in print and image quality;
- Includes documents that contain handwritten text and notations, diverse fonts, multiple character sets, and graphical elements, namely graphs, tables, photos, logos, and diagrams;
- Contains a sufficiently high volume of documents;
- Contains documents in multiple languages including documents that have multiple languages within the same document;
- Contains a vast volume of redundant and irrelevant documents;
- Supports diverse applications, thereby, includes private communications within and between groups planning activities and deploying resources;
- Is publicly available at minimal cost and licensing.

The collection chosen is a subset of the Master Settlement Agreement documents hosted by the University of California at San Francisco as the Legacy Tobacco Document Library (see http://legacy.library.ucsf.edu). These data were made public via legal proceedings against United States tobacco industries and research institutes. For the most part, the documents are distributed free of charge and are free of copyright restrictions. (The sued parties did not own a few of the Legacy Tobacco Document Library documents included; hence, some of them are potentially subject to copyright restrictions.) The collection consists of roughly 7 million documents or approximately 42 million scanned TIFF format pages (about 1.5 TB). These documents are predominantly in English; however, there are some documents in German, French, Japanese, and a few other languages. A few of these documents also include multiple languages within a given document. As multiple companies at multiple sites using a diversity of scanners scanned the pages, the resulting image quality varies significantly.

As search benchmark data collections require queries with associated relevant documents indicated, we developed in excess of 50 such queries of varying complexity for the Legacy Tobacco Document Collection. This benchmark collection is, however, only in its "infancy" stage. It currently suffers from a rather limited coverage of query topics and a low number of relevant documents per query. None the less, the collection was successfully used for the NIST TREC Legal Track both in 2006 and 2007.

A complete description of the collection is provided in [1].

2. The CDIP Prototype

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 facilitates the later inclusion of a fourth core technology: *data mining*. As noted, specific attention was paid to modular design to ensure that the developed software modules are easily integrated 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, the location of images/graphics, the location of logos and signatures, the location of signatures and handwritten comments [3], and the identification of signatures [4, 5, 6, 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 Clarabridge^{TM} technology. Our initial tests on real-world data showed the effectiveness of entity extraction on noisy text obtained from OCR of our 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 of the technology, the mediator implementation technology that exists to date is only at the prototype level. Consequently, as we needed a more robust framework by which to implement our CDIP prototype. We built our prototype using the Clarabridge^{TM} integration fabric.

Data mining, a component not currently implemented in our current prototype, will 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 was 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 (Figure 1) 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^{TM}'s SDK and the DocLib package [2]).
- OCR in text regions (using Abbyy^{TM}'s SDK), recognition of logos (using the DocLib package [2]), and recognition of signatures (using CEDAR's signature recognition system [3, 4, 5]) and a signature warping module [6, 7].
- Linguistic and classification analysis of extracted information for annotation in the database: entity tagging, relationship tagging, and stylistic tagging in text regions (using Clarabridge^{TM} Software).

At the end of the ingestion process, we have an operational data store in third normal form (3NF). At this point, it would be too 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 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 Clarabridge^{TM}. These tools use web services to access the point solutions and identify metadata about complex documents to populate the 3NF schema. Clarabridge^{TM} tools also migrate the 3NF schema to a star schema using well known Extract, Transform, and Load (ETL) processing. Clarabridge^{TM} 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.

A key component that is facilitated by our approach is a tight integration of the processes of document image interpretation, symbol extraction and grounding, 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 with different types of document image information. The Abbyy^{TM} OCR engine is used to extract text from the document image. This text is fed to the Clarabridge^{TM} 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.

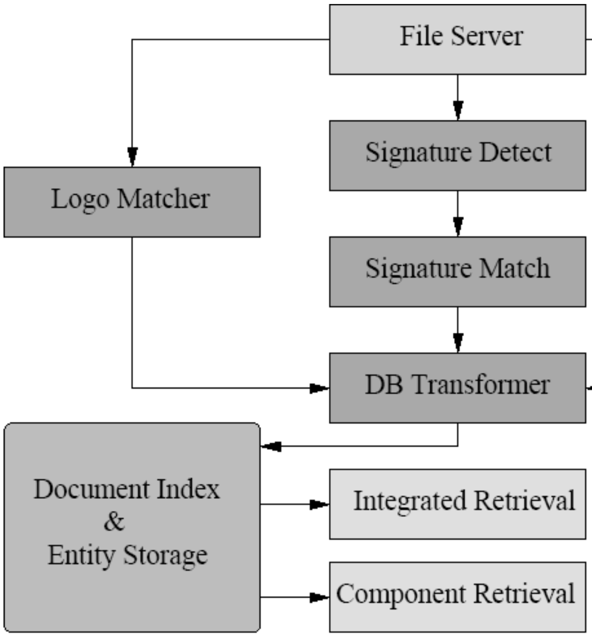

Figure 1: Architectural overview of the current CDIP research prototype.

3. Document Image Analysis Components

3.1. Document Image Enhancement

Given an image of a faded, washed out, damaged, crumpled or otherwise difficult to read document, one with mixed handwriting, typed or printed material, with possible pictures, tables or diagrams, it is necessary to enhance its readability and comprehensibility. Documents might have multiple languages in a single page and contain both handwritten and machine printed text. Machine printed text might have been produced using various technologies with variable quality. The approach we developed [13] addresses automatic enhancement of such documents and is based on several steps: the input image is segmented into foreground and background, the foreground image is enhanced, the original image is enhanced, and the two enhanced images are blended using a linear blending scheme. The use of the original image in addition to the foreground channel allows for foreground enhancement while preserving qualities of the original image. In addition, it allows for compensation for errors that might occur in the foreground separation.

The enhancement process we developed produces a document image that can be viewed in different ways using two interactive parameters with simple and intuitive interpretation. The first parameter controls the decision threshold used in the foreground segmentation, the second parameter controls the blending weight of the two channels. Using the decision threshold the user can increase or decrease the sensitivity of the foreground segmentation process. Using the blending factor the user can control the level of enhancement: on one end of the scale the original document image is presented without any enhancements, whereas on the other end, the enhanced foreground is displayed by itself. Note that the application of these two adjustable thresholds is immediate once the document image has been processed. The adjustment of the parameters is not necessary and is provided to enable different views of the document as deemed necessary by the user. For automated component analysis purposes the parameters can be set automatically.

3.2. Logo Detection

Our approach for logo detection is based on two steps: detection of distinct document zones and classification of the different zones detected. For efficiency reasons, some heuristics incorporating the expected location of logos are used to reduce the candidate set. We employed detection of distinct zones using the DOCLIB library [2]. Our approach uses automated means of training a classifier to recognize a document layout or set of layouts [14]. The classifier is then used to score an unknown image. For page segmentation, we use the Docstrum method for structural page layout analysis [15, 16]. The Docstrum method is based on bottom-up, nearest-neighbor clustering of page components. It detects text lines and text blocks, and has advantages over many other methods in three main ways: independence from skew angle, independence from different text spacing, and the ability to process local regions of different text orientations within the same image. Script identification [17] for machine printed document images can be used to increase reliability. This approach allows for classifying a document image as being printed in one of the following scripts: Amharic, Arabic, Armenian, Burmese, Chinese, Cyrillic, Devanagari, Greek, Hebrew, Japanese, Korean, Latin, or Thai. Script identification can also be retrained to focus on different language mixes. Once the zones are detected, logo detection works by identifying blocks with certain spatial and content characteristics including: relative position of the zone's center of mass, the aspect ratio of the zone's bounding box, the relative area of the bounding box, and the density of the bounding box. The features are tuned based on a training set of documents.

3.3. Logo Recognition

Logo recognition is performed by matching candidate regions against a database of known logos. While it is possible to match logos by extracting and matching feature vectors, it has been shown that direct correlation of bitmaps produces better results [18, 19]. To improve the correlation measure, we first normalize the logos to be of standard size and orientation and then sum the products of corresponding elements in the bitmaps. The computed correlation measure is the standard gray-scale correlation. For each candidate a score between 0 and 100 is generated corresponding to the degree of similarity. The best match is provided along with the score. To improve performance, the algorithm stops comparing against candidate logos when the best score is beneath a predefined threshold. Text that is associated with logos can be used in assisting the recognition of the associated logo, but is not currently used in our system.

3.4. Signature Detection

Signature detection is performed using algorithms for document zoning as described before, and analyzing the different zones for signatures [20]. In analyzing zones for signatures, line and word segmentation are necessary. The process of automatic word segmentation [21] begins with obtaining the set of connected components for each line in the document image. The interior contours or loops in a component are ignored for the purpose of word segmentation as they provide no information for this purpose. The connected components are grouped into clusters, by merging minor components such as dots above and below a major component. Every pair of adjacent clusters are candidates for word gaps. Features are extracted for such pairs of clusters and a neural network is used to determine if the gap between the pair is a word gap. Possible features are: width of the first cluster, width of second cluster, difference between the bounding box of the two clusters, number of components in the first cluster, number of components in the second cluster, minimum distance between the convex hulls enclosing the two clusters and the ratio between, the sum of the areas enclosed by the convex hulls of the individual clusters, to the total area inside the convex hull enclosing the clusters together. The minimum distance between convex hulls is calculated by sampling points on the convex hull for each connected component and calculating the minimum distance of all pairs of such points.

3.5. Signature Recognition

Signature recognition works by obtaining feature vectors for signatures and measuring the similarity between feature vectors of compared signatures. Image warping techniques can be used to increase the similarity between signatures before comparing them. The approach for signature feature extraction we employ [4, 5, 6, 7], consists of taking the block of the image that is identified as a potential signature and partitioning it into rectangles such that the size of each rectangle is adapted to the content of the signature. Each rectangle is examined for multiple features (e.g., curvature of lines, principal directions, fill ratio, etc.). The obtained feature vector is then compared to a database of known signatures that are represented in a similar way. The vectors are matched and the signatures that match the closest are identified as possible candidates.

4. Performance Evaluation

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 have finished the initial implementation of our research prototype and are currently in the process of evaluating it quantitatively. 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. 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 tens of thousands of 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 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 illustrate here the kinds of capabilities that our prototype currently supports. The mini-corpus used for this consists of 800 documents taken from the IIT CDIP benchmark collection. 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.

Once the metadata are populated using logo and signature processing components, SQL queries easily associate both textual and non-textual data. One query involving currency amounts found in text showed that Dr. D. Stone, who 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.

5. Conclusion

As stated throughout, the complex document information processing area of research is only in its infancy. We have developed an initial prototype, but have yet to effectively evaluate it. We have, however, created a benchmark that should stress all foreseeable future complex document information processing systems. This benchmark was already used to evaluate some search systems in recent TREC activities; we can only hope that its availability will inspire further research into the design of complex document information processing systems.

Acknowledgements

We thankfully acknowledge and are grateful to Shlomo Argamon, David Doermann, David Lewis, Gene Sohn, and Sargur Srihari for their vast contributions to the CDIP effort. We also thank the March 2005 CDIP Workshop participants for their suggestions.

References

[1] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard, “Building a test collection for complex document information processing,” in *ACM Twenty-Ninth Conference on Research and Development in Information Retrieval (SIGIR)*, (Seattle, Washington), August 2006.

[2] K. Chen, S. Jaeger, G. Zhu, and D. Doermann, “DOCLIB: a document processing research tool,” in *Symposium on Document Image Understanding Technology*, pp. 159–163, 2005.

[3] 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.

[4] 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.

[5] 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.

[6] G. Agam and S. Suresh, “Particle dynamics warping approach for offline signature recognition,” in *IEEE Workshop on Biometrics*, pp. 38–44, 2006.

[7] G. Agam and S. Suresh, “Warping-based offline signature recognition,” *IEEE Trans. Information Forensics and Security*, 2007. Accepted for publication.

[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] G. Agam, G. Bal, G. Frieder, and O. Frieder, “Degraded document image enhancement,” in *Document Recognition and Retrieval XIV*, X. Lin and B. A. Yanikoglu, eds., *Proc. SPIE* **6500**, pp. 65000C–1 – 65000C–11, 2007.

[14] L. Golebiowski, “Automated layout recognition,” in *Symposium on Document Image Understanding Technology*, pp. 219–228, 2003.

[15] L. O’Gorman, “The document spectrum for page layout analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence* **15**(11), pp. 1162–1173, 1993.

[16] S. Mao, A. Rosenfeld, and T. Kanungo, “Document structure analysis algorithms: a literature survey,” in *Proc. SPIE Electronic Imaging*, **5010**, pp. 197–207, 2003.

[17] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, “Automatic script identification from document images using cluster-based templates,” *IEEE Trans. Pattern Analysis and Machine Intelligence* **19**(2), pp. 176–181, 1997.

[18] G. Zhu, S. Jaeger, and D. Doermann, “Robust stamp detection framework on degraded documents,” in *Proc. Intl. Conf. Document Recognition and Retrieval XIII*, pp. 1–9, 2006.

[19] D. S. Doermann, E. Rivlin, and I. Weiss, “Applying algebraic and differential invariants for logo recognition,” *Machine Vision and Applications* **9**(2), pp. 73–86, 1996.

[20] Y. Zheng, H. Li, and D. Doermann, “Machine printed text and handwriting identification in noisy document images,” *IEEE Trans. Pattern Analysis and Machine Intelligence* **26**(3), pp. 337–353, 2004.

[21] S. Srihari, H. Srinivasan, P. Babu, and C. Bhole, “Spotting words in handwritten Arabic documents,” in *Proc. SPIE*, pp. 101–108, 2006.