Complex Document Information Processing: Towards Software and Test Collections

G. AGAM, S. ARGAMON, O. FRIEDER, D. GROSSMAN
Illinois Institute of Technology,
Department of Computer Science
Chicago, IL 60616
{agam,argamon,frieder,grossman}@iit.edu

D. LEWIS
David D. Lewis Consulting
858 W. Armitage Ave., #296
Chicago, IL 60614
davelewis@daviddlewis.com

Keywords: Search & Retrieval, Fusion

I. THE CDIP PROBLEM

Complex document information processing (CDIP) is the analysis of combinations of handwritten notes, diagrams, and graphics in addition to printed and formatted text. Rapid analysis of collections of complex documents is increasingly critical to intelligence appraisals of hostile forces and other entities of interest. Such collections may be seized by military or law enforcement (recently in Iraq, Afghanistan, and elsewhere), or turned over by cooperating parties (e.g. defectors). This material must be scanned or otherwise input, understood, and exploited all in a short period of time if it is to be of maximum value.

Such collections may contain many types of complex documents and data, and ideally should be interpreted in the context of previously obtained collections and knowledge. Unfortunately, most current tools are specialized to one kind of data (text content or style, handwriting, graphics, database records, etc.) and sometimes only one genre of document.

Extracting intelligence from such collections therefore relies heavily on manual partitioning of data, coordination of analysis tools, and collation of results. Besides being slow, expensive, and demanding on the time of trained analysts, the current approach has two fundamental problems. First, the speed and accuracy of analyzing one data type can often be improved if partial analyses from other data types were available, but a collection of standalone tools does not support this. Second, manual collation, cross-checking, and consistency checking is slow and error-prone when dealing with large numbers of documents, further reducing the value of obtained information.

The operational problem, therefore, is the dearth of unified tools for CDIP in both commercial and research spheres. That in itself, however, is symptomatic of a second problem: the lack of standard test collections for CDIP research. Test collections of documents (and manual judgments/annotations) in text retrieval and categorization, handwriting recognition, and data mining have sparked intense research and commercial development. Conversely, CDIP research and development is stifled by the lack of a test collection of complex documents. Creating a CDIP test collection is thus an essential component of any serious CDIP effort.

II. BUILDING A CDIP SYSTEM

We are developing an integrated system for CDIP, which will support (when complete): (a) segmenting and analyzing complex documents into components, (b) importing analyzed documents into database structures supporting efficient data access, (c) answering queries against information spanning multiple data types, and (d) discovery of patterns and relationships spanning multiple data types and documents.

Our design is modular, with attention paid to allowing modules to be used both by each other and by existing external analytical tools. The goal is to enable users to access *all* the information contained in complex documents, including images, text, handwriting, graphics, and structured data. Our efforts are modularizing and adapting existing analysis techniques for each data type to our CDIP framework.

Image analysis is a critical component of CDIP since collections are often obtained in paper form and scanned. Image analysis modules are responsible for segmenting complex documents into components. Printed text will be separated from handwriting and graphics [1], [2], characters and words in printed and handwritten text will be segmented from each other and converted to the appropriate character set, and features characterizing these different components will be extracted for higher-level processing. Other modules determine spatial relations between document image components, to enable characterization of overall document structure, and recognizing tables, lists, and logos.

Text/symbol analysis augments text documents with class labels or structural interpretations. This allows queries to refer to combinations of classes from finite sets, rather than (or in additional to) unbounded sets of words and phrases. Data mining is similarly aided by providing sets of attributes that are smaller, less ambiguous, and more task specific than natural language. For text, we concentrate on text categorization by topic [11] and stylistic analysis [6], [7] to determine a partial profile of the text and its author, for genre, tone, and authorial characteristics. Handwriting and signature analysis can be used to identify the authors of handwritten notes (either by recognition based on known samples, or by clustering to determine that different samples are from the same hand). Too, diagram clustering [4] can provide symbolic labels for similar graphics in the corpus in some cases.

Integrated retrieval from different kinds of data sources is a key mediator function—our IIT Mediator currently supports a variety of traditional data sources such as unstructured text, semistructured XML/text data, as well as structured database querying [10]. 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. Results are then fused into an integrated retrieval set [8]. We build on this foundation to allow querying of information extracted from complex documents.

Data mining can directly leverage indexing of heterogeneous data in a relational database model. This approach allows application of traditional data mining methods such as association rules [3] to combinations of structured and unstructured data.

III. Building a CDIP Evaluation Framework

A test collection for CDIP that enables both component tools and the integrated system to be tuned and evaluated requires a large set of scanned documents rich in alternative data types, as well as manual annotations and corresponding evaluation metrics and software.

We have researched publicly available collections of scanned documents, and for the initial test collection will be using the Legacy Tobacco Documents Library hosted at UCSF (http://legacy.library.ucsf.edu/) which comprises 42 million document images. Many of these documents include handwritten comments, logos, signatures, tables, and graphics in addition to text. Future phases of the project will include also non-English materials, starting with document images culled from Arabic-language online newspapers.

A test collection is not just a data set. We need to annotate documents with metadata indicating their provenance, context, and type. Selected samples must be manually annotated with correct answers to testbed retrieval, analysis, and mining tasks, to support tuning of components and evaluation of system effectiveness. We are in discussions with domain experts who have an intense interests in some of the above materials and are potential partners in annotation.

We will use standard metrics in evaluating individual components of a CDIP system. For instance, a query asking for documents on topic X having handwritten annotations by a person Y will produce a ranked list that can be evaluated by recall, precision, and other text retrieval metrics. Quantitative metrics, along with qualitative analysis and tracing errors to their component sources, will both be used to evaluate our initial prototype and to identify areas on which to focus in future attacks on the CDIP problem.

IV. Current Status

The IIT Mediator[1] is a prototype implementation of an integrated retrieval system, for document data represented in a structured format (as above). It accesses a variety of data sources, including free-text, semistructured XML, and structured databases. Complex document information will be integrated as lower-level processing modules reach maturity.

Our work on symbolic analysis has focused on topic and style based text categorization. We have emphasized both extracting informative features and combining them via machine learning. Of particular interest is a machine learning approach, Bayesian logistic regression [9], that uses both domain knowledge and training data. We have also developed an infrastructure for processing and representing linguistic annotations for texts in a database format, plus tools for extracting a wide variety of linguistic textual features. We are developing feature extractors based on new kinds of linguistic analysis, enabling more accurate and more insightful stylistic text classification [5].

Our work on document image processing is currently focused on text/handwriting separation, line segmentation, stroke segmentation, and feature extraction. Directional and regulated morphological operations [2] have proven useful for separating text/graphics as well as for processing mixed documents, maps, and line drawings. We are investigating issues such as parameter selection and scale invariance.

Our current efforts are funded in part by a subcontract to the Illinois Institute of Technology of an ARDA contract to Pacific Northwest National Laboratories. A two-and-a-half day workshop was held March 29-31, 2005 in Chicago to kickoff this effort. It was attended by over a dozen internationally-renowned experts.

[1]Demo version available at `http://www.ir.iit.edu/mediator`.

References

[1] G. Agam and I. Dinstein. Adaptive directional morphology with application to document analysis. In P. Maragos, R. W. Schafer, and M. A. Butt, editors, *Mathematical Morphology and its Applications to Image and Signal Processing*, pages 401–408. Kluwer Academic Publishers, 1996.

[2] G. Agam and I. Dinstein. Regulated morphological operations. *Pattern Recognition*, 32(6):947–971, 1999.

[3] Rakesh Agrawal, Hiekki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. pages 307–328, 1996.

[4] S. Aksoy and R. M. Haralick. Graph-theoretic clustering for image grouping and retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 63–69, 1999.

[5] S. Argamon and J. Dodick. Linking rhetoric and methodology in formal scientific writing. In *Proc. 26th Annual Meeting of the Cognitive Science Society*, August 2004.

[6] S. Argamon, M. Koppel, J. Fine, and A. R. Shimony. Gender, genre, and writing style in formal written texts. *Text*, 23(3), 2003.

[7] S. Argamon, M. Šarić, and S. S. Stein. Style mining of electronic messages for multiple author discrimination. In *Proc. ACM Conference on Knowledge Discovery and Data Mining*, 2003.

[8] 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), July 2004.

[9] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Journal of Machine Learning Research*, 2004. submitted.

[10] D. Grossman, S. Beitzel, E. Jensen, and O. Frieder. IIT intranet mediator: Bringing data together on a corporate intranet. *IEEE IT PRO*, January/February 2002.

[11] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, April 2004.