In Proceedings of the CHI 2007 Workshop for Exploratory Search and HCI,
April 2007, San Jose, California.

Interfaces to Support the Scholarly Exploration of Text Collections

Georg Apitz
Department of Computer Science & HCIL
University of Maryland
College Park, MD 20742, USA
geapi@cs.umd.edu

Jimmy Lin
College of Information Studies & HCIL
University of Maryland
College Park, MD 20742, USA
jimmylin@umd.edu

ABSTRACT

The analysis of text collections forms the basis of scholarship in many disciplines in the humanities and social sciences. Despite the growing availability of electronic texts, automated techniques have not been effectively exploited to support the activities of scholars in these fields. We present a prototype search interface for exploring text collections that places equal emphasis on content, what the document is *about*, and metadata, the context that situates a piece of text. As a start, we focus on a selection of briefs and opinions from the U.S. Supreme Court to support legal scholars.

Author Keywords

User interfaces, exploratory search, creativity support tools

INTRODUCTION

Scholarship in many areas of the humanities and social sciences involves large collections of textual documents. Consider, for example, a legal scholar who combs through hundreds of court documents to find support for a particular hypothesis, or a literary scholar who analyzes countless pages of writings to ascertain stylistic influences among contemporaries. Traditionally, such studies have been laboriously carried out by hand, and despite the growing availability of these texts in electronic format, automated systems are not usually exploited to support scholarly endeavors.

At a broad level, our work aims to develop innovative interfaces to help scholars in the humanities and social sciences explore large text collections. We seek to develop creativity support tools [12] for scholarship in these various fields. The goal is not to supplant, but to augment the scholar in the creation of knowledge—by utilizing the tool to visualize relationships between different texts, to drill down into a text, to draw connections otherwise not apparent, and most important of all, to form hypotheses that provide the basis for further exploration. We take the view that computers by themselves do not generate knowledge—its creation falls within the purview of humans who interpret system output. Thus, the system's primary role should be to facilitate the human-centered processes of hypothesis formulation, evidence gathering, etc. As a start, we focus on a collection of legal briefs and opinions from U.S. Supreme Court cases.

It is apparent that the tools we're describing lie at the intersection of human-computer interaction and information retrieval. Information retrieval and related text processing techniques (data mining, linguistic analysis, text classification and clustering, etc.) can unearth characteristics of texts that may be of interest to scholars. However, these findings need to be synthesized and pre-digested into a form suitable for consumption by individuals who may not be experts in computational techniques. This requires, for example, visualizations that render various relationships apparent, explanation tools that help the scholar understand the findings of automated algorithms, and controls to subsequently affect the behavior of these algorithms.

This paper focuses on the interface aspect of such creativity support tools. Using off-the-shelf information retrieval technology, we demonstrate a prototype interface for exploratory search that emphasizes both the presentation of content and associated metadata. The emphasis on both aspects provides scholars with a rich environment to engage a text collection.

METADATA VS. CONTENT

A *document* (used broadly to encompass any piece of text) is characterized by its content, on the one hand, and its associated metadata, on the other hand. Content is comprised of the words that make up the text and define what the document is *about*. Metadata define important characteristics such as authorship, document type, time of creation, association with other documents, etc. In a sense, metadata define the context of a particular document.

Traditionally, the exploration of text collections is performed primarily through the manipulation of either metadata or content. In the legal domain, one might use metadata to explore briefs and opinions by case, by author (e.g., judge), by issue area, or chronologically. This is often facilitated by inserting metadata into a database, which allows users to issue arbitrary relational queries. For specific tasks, this can be a very useful and efficient method for accessing documents.

Free-text search on document content represents an alternative to exploration by metadata. Given the prevalence of Web search engines such as Google, this mode of information access is most familiar to users today. In response to a keyword query, retrieval engines return a list of "hits" that are likely to be relevant. However, metadata are not leveraged in the search algorithm for the most part, and when they are, results often conflate multiple factors (for example, Google rankings combine topical relevance, authority, popularity, etc.). In a general Web environment, one might argue that these issues are less important to the typical end user. However, this is not the case for scholarly collections, where the context of a particular document, as defined by metadata, is often the subject of exploration itself.

We believe that a tool for exploring document collections must allow users to manipulate both content and metadata. Although some search engines provide mechanisms to search on metadata (through special query operators or in different fields), these capabilities do not go far enough and are not supported by corresponding interfaces that render the relevant relationships transparent. In our opinion, the traditional output of most search engines—a list of hits sorted by relevance—is not sufficient to support exploration. To address this, we focus on search interfaces as a starting point.

PREVIOUS WORK

We divide our discussion of previous work into two parts, one surveying the literature from information visualization, and the other reviewing work in information retrieval. Although visualization interfaces are clearly relevant to this work, most existing systems focus solely on metadata. For example, BiblioViz [11] employs a combination of different visualization techniques and clustering approaches to present bibliographical data (based on contributions to the InfoVis 2004 contest). Chen [2] also presents 3-D and 2-D visualizations of citation and co-citation networks as a summary of the InfoVis 2004 contest entries. The Hierarchical Clustering Explorer (HCE) [9] supports the user in exploring a dataset based on the metadata—the idea is that the user first explores the data in 1-D, then in 2-D, and then uses a rank-by-feature mechanism to expose more interesting patterns.

In the information retrieval literature, there is a substantial body of work on search interfaces. Some of these focus on query specification [14] or graphically convey the distribution of query terms in retrieved document sets [3, 15]. Others attempt to visualize the relationship between documents in the result set [1, 6]. Cat-a-Cone [4] represents an interesting attempt to combine search and browsing, but was primarily designed for hierarchically-categorized documents. The work of Nowell et al. [7] share the most similarities with ours, although their Envision system was designed as a tool for accessing digital libraries.

THE DIGITAL DOCKET

This work is situated in the context of the Digital Docket, a recently-funded NSF project that aims to apply automated content analysis to support legal scholarship. Previous research of judicial systems has faced a trade-off between large scale quantitative inquiries focused on readily-counted behaviors and smaller studies that allow closer examination of legal texts. This project, a multi-disciplinary collaboration between several units on the University of Maryland campus, aims to apply computational techniques to the study of the U.S. Supreme Court.

By viewing the legal system as an intricate and complex web of communication, discourse, and debate, the project aims to better understand the role and influences of various actors through analysis of written records. Those records include, for example, briefs written by litigants and other stakeholders, and opinions written by judges and justices. The application of computational techniques to model the U.S. judicial system represents an opportunity to overcome many of the bottlenecks associated with traditional labor-intensive methods in political science, and also provides a new environment for the advancement of text algorithms.

Propagation of "Legal Memes"

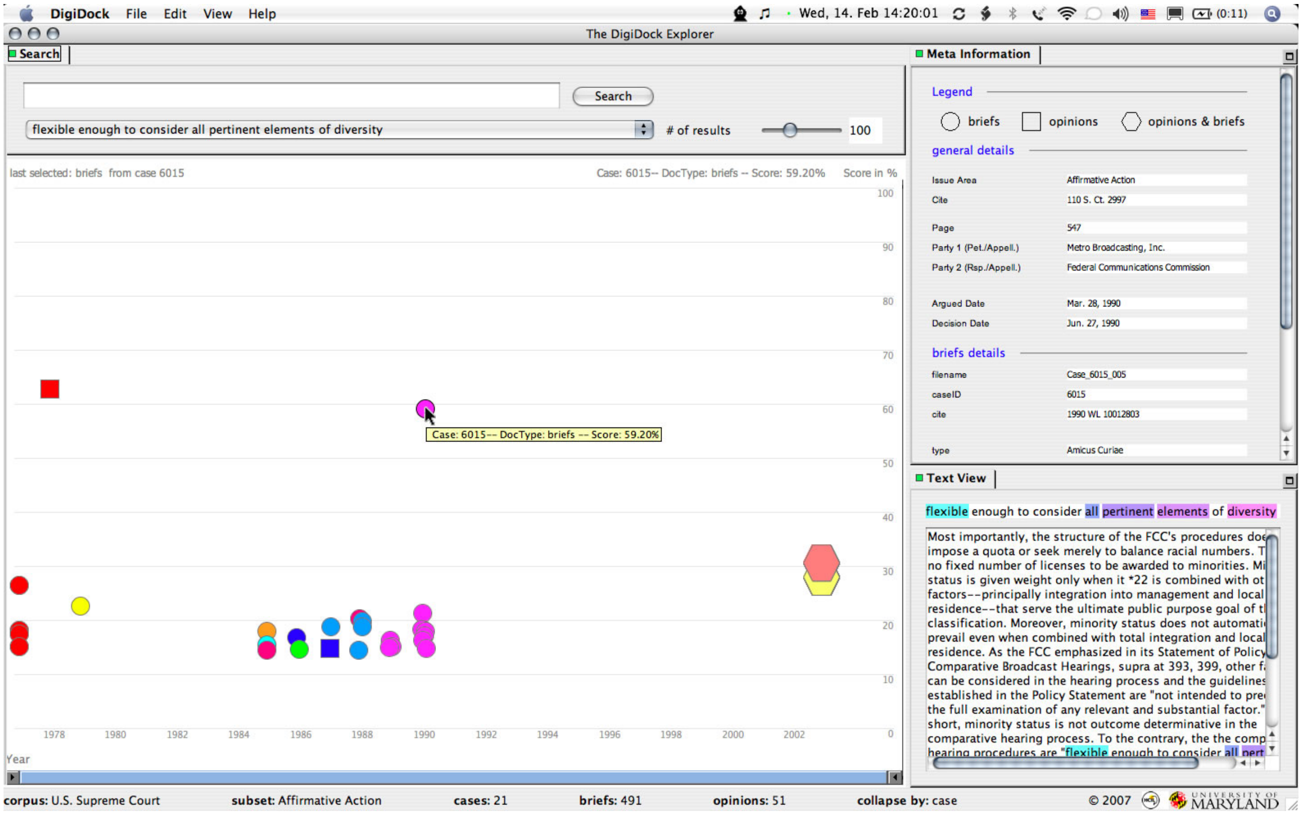
As a first goal, we are developing a tool that assists legal scholars in exploring the propagation of "legal memes". Consider the issue of affirmative action, a topic that has recently received substantial coverage in the popular press. In the landmark Supreme Court case *Regents v. Bakke (1978)*, Justice Powell wrote that an admissions program must be "flexible enough to consider all pertinent elements of diversity". Such a statement introduces a coherent concept into the "ecosystem" of the legal system—subsequent litigants read the opinion, future briefs may quote or cite this idea, law review articles may debate its interpretation, etc. A legal scholar may be interested in the following:

- Where did this "meme" come from? Did it originate from Powell, or from an even earlier piece of writing?
- How influential is this idea? Obviously, direct quotes and citations are one way to quantify this, but ideas are often paraphrased without explicit reference.
- Do references to this meme change over time? Is it currently "out of fashion" to discuss affirmative action issues in these terms?
- What are the subsequent interpretations of this idea? Do other justices have incompatible views?
- How does this idea affect the perception of the individual? Does one gain prominence as a result of widely-adopted ideas, or vice versa?

A tool for exploring court documents can be valuable to legal scholars in formulating hypotheses and in gathering evidence. Content is clearly important in this application, but so is metadata, since it provides the basis for answering many of the above questions.

A PROTOTYPE

We have developed a prototype exploratory search tool, the DigiDock Explorer, that allows legal scholars to explore documents associated with cases heard by the U.S. Supreme Court—it begins to support some of the activities discussed



flexible enough to consider all pertinent elements of diversity

# of results 100

last selected: briefs from case 6015

Case: 6015-- DocType: briefs -- Score: 59.20%

Case: 6015-- DocType: briefs -- Score: 59.20%

Issue Area	Affirmative Action
Cite	110 S. Ct. 2997
Page	547
Party 1 (Pet./Appell.)	Metro Broadcasting, Inc.
Party 2 (Rsp./Appell.)	Federal Communications Commission
Argued Date	Mar. 28, 1990
Decision Date	Jun. 27, 1990
filename	Case_6015_005
caseID	6015
cite	1990 WL 10012803
type	Amicus Curiae

flexible enough to consider all pertinent elements of diversity

Most importantly, the structure of the FCC's procedures doe
impose a quota or seek merely to balance racial numbers. T
no fixed number of licenses to be awarded to minorities. Mi
status is given weight only when it *22 is combined with ot
factors--principally integration into management and local
residence--that serve the ultimate public purpose goal of tl
classification. Moreover, minority status does not automati
prevail even when combined with total integration and local
residence. As the FCC emphasized in its Statement of Policy
Comparative Broadcast Hearings, supra at 393, 399, other f
can be considered in the hearing process and the guidelines
established in the Policy Statement are "not intended to pre
the full examination of any relevant and substantial factor."
short, minority status is not outcome determinative in the
comparative hearing process. To the contrary, the the comp
hearing procedures are "flexible enough to consider all pert

corpus: U.S. Supreme Court **subset:** Affirmative Action **cases:** 21 **briefs:** 491 **opinions:** 51 **collapse by:** case

Figure 1. The DigiDock Explorer screen. The main portion of the screen shows the search results for the current search phrase displayed as time (x-axis) vs. relevance (y-axis). Several results have been aggregated by case (hexagons). The right side shows details about the selected item, these are: general and document type specific information as well as a text view. At the bottom a status bar shows details about the corpus and current settings.

above. The application is implemented in Java using the Prefuse [5] Information Visualization toolkit. A screenshot is shown in Figure 1.

Document Collection and Preparation

Twenty U.S. Supreme Court cases about affirmative action between 1978 and 2003 were selected for inclusion in our sample collection. For each case, we obtained all briefs from the litigants, all *amici* (third-party) briefs, and all opinions of the court. This totaled approximately five hundred documents, about half a million words. Although this represents a small collection, it is useful for legal scholars since its coverage of one specific issue area at the Supreme Court level is relatively complete.

Each document in our collection is associated with metadata mined from various sources. They include properties such as the document type, the author, date, the case association, etc. Different types of documents have additional specialized metadata; for example, opinions can be majority, dissenting, etc. In addition, we have metadata about the cases in general, e.g., what the outcome was, the vote, etc.

The entire collection was indexed with Lucene, an open source search engine that uses modified *tf.idf* weighting. Since the documents are very long on average, we segmented them into paragraphs and indexed each separately.

Interface

Like most retrieval systems, the starting point of the DigiDock Explorer is a search box at the top of the interface, where users can input a query string or select from predefined queries (that may be populated, for example, by text mining algorithms). Retrieved results (paragraph segments) are displayed in a scatterplot in which the *x*-axis shows the creation date of the document and the *y*-axis shows the similarity score returned by Lucene.

We have chosen this display format since feedback from legal scholars has indicated that time is the most important dimension to consider when examining cases. Since the legal system is causal in that previous cases shape the argumentation and outcome of subsequent cases, an explicit visualization of this property provides important cues. Different icons represent different document types (circles for briefs, squares for opinions). Users can group documents to reduce clutter—for example, a large hexagon is an aggregate representation of all segments belonging to a single case. With this display, scholars can get a broad overview of the information landscape and directly answer questions such as:

How many potential instances of this concept exist in the collection? How are they distributed temporally?

The right hand side of the application displays metadata associated with the retrieved segment and its contents. Keyword highlighting is employed to facilitate browsing of text. Arbitrary segments can be selected and used as "queries", thus allowing drill-down. Users can also use the metadata panel to control the display, for example, to show only certain types of documents, text by certain authors, etc.

NEXT STEPS

With our early prototype already in use by scholars we are using their feedback to extend the functionality of the DigiDock Explorer. We are planning to allow the user to have different visualizations for the search results, as well as very fine-grained control over how the search is conducted.

Furthermore, we are investigating ways of visualizing the propagation of legal memes. An initial 'river like' layout seems like a good starting point but we are envisioning representations that show the flow, as well as the kind of documents that are influences and the people who are influenced. This aims at the representation of several features in the propagation and should potentially support the user in ranking these features by their importance, similar to Seo and Shneiderman's Rank by Feature approach [10].

In addition to "free association" exploration, we envision that scholars use our tool in *sessions* where they aim to accomplish a specific goal by performing several iterations of drill-down, refinement, and restriction of intermediate results with the tool. These *sessions* themselves encode information about a search strategy, the steps involved, and the criteria applied. Thus, we want to provide users with a recording functionality that allows them to store, replay, and share such sessions.

The legal domain represents one instance where scholars rely on large collections of text. We imagine a general purpose tool that can be applied to other disciplines in the humanities and social sciences. Comparative literature represents an interesting application—see, for example, [8]. Tools for exploring large text collections can be extended even further into domains such as biology and medicine.

The evaluation of interfaces for domain experts is a very specific challenge we must address. Traditional user studies, where subjects work 30 minutes or so with an application are not well suited for our purposes. Thus, we are planning on applying a technique called Multi-Dimensional, In-depth, Long-term, Case studies (MILCs) [13]. With this evaluation methodology, it is possible to observe expert users in their traditional environment for longer periods of time and gather information that only occurs in natural settings. Logging capabilities can deliver fine-grained information about user interactions that are difficult to observe. In addition, interviews, self reports, and think-aloud methods can provide more subjective insights from the scholars.

CONCLUSION

The exploration of text collections is a common task among scholars, yet so far, there is no solution that leverages information about the documents (metadata) as well as information on what the document is about (content). We aim to develop systems that combine the benefits from both with the domain expertise of scholars.

Our prototype demonstrates a new way for legal scholars to explore, research, and examine judicial systems. This application enables them to look at document collections in ways never before possible and to gain insights that traditionally would require much manual labor. We believe that this example illustrates a new way of interacting with text collections that enables scholars to more fully exploit the benefits of technology in combination with their knowledge.

ACKNOWLEDGMENTS

This project is funded by NSF under award BSC-0624067. Any opinions, findings, and conclusions expressed in this paper are those of the authors. We would like to thank Ben Shneiderman for valuable discussions. The authors would like to thank Anja, Esther, and Kiri for their kind support.

REFERENCES

1. M. Chalmers and P. Chitson. Bead: Explorations in Information Visualization. In *SIGIR 1992*.
2. C. Chen. Information Visualization Research: Citation and Co-Citation Highlights. In *INFOVIS 2004*.
3. M. Hearst. TileBars: A Visualization of Term Distribution Information in Full Text Information Access. In *CHI 1995*.
4. M. Hearst and C. Karadi. Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy. In *SIGIR 1997*.
5. J. Heer. Prefuse: Information Visualization Toolkit. *www.prefuse.org*, 2004.
6. R. Korfhage. To See, or Not to See—Is That the Query? In *SIGIR 1991*.
7. L. Nowell, R. France, D. Hix, L. Heath, and E. Fox. Visualizing Search Results: Some Alternatives to Query-Document Similarity. In *SIGIR 1996*.
8. C. Plaisant, J. Rose, B. Yu, L. Auvil, M. Kirschenbaum an M. Smith, T. Clement, and G. Lord. Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces. In *JCDL 2006*.
9. J. Seo and B. Shneiderman. Interactively Exploring Hierarchical Clustering Results. *Computer*, 35(7):80–86, 2002.
10. J. Seo and B. Shneiderman. A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data. *Information Visualization*, 4(2):96–113, 2005.
11. Z. Shen, M. Ogawa, S. Teoh, and K. Ma. BiblioViz: A System for Visualizing Bibliography Information. In *Proceedings of Asia-Pacific Symposium on Information Visualization*, 2006.
12. B. Shneiderman. User Interfaces for Creativity Support Tools. In *Proceedings of the 3rd Conference on Creativity & Cognition*, 1999.
13. B. Shneiderman and C. Plaisant. Strategies for Evaluating Information Visualization Tools: Multi-Dimensional In-Depth Long-Term Case Studies. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors*, 2006.
14. A. Spoerri. InfoCrystal: A Visual Tool for Information Retrieval & Management. In *CIKM 1993*.
15. A. Veerasamy and N. Belkin. Evaluation of a Tool for Visualization of Information Retrieval Results. In *SIGIR 1996*.