Chapter 8

The Use of Text Mining Techniques in Electronic Discovery for Legal Matters

Michael W. Berry  
University of Tennessee, USA

Reed Esau  
Catalyst Repository Systems, USA

Bruce Kiefer  
Catalyst Repository Systems, USA

ABSTRACT

Electronic discovery (eDiscovery) is the process of collecting and analyzing electronic documents to determine their relevance to a legal matter. Office technology has advanced and eased the requirements necessary to create a document. As such, the volume of data has outgrown the manual processes previously used to make relevance judgments. Methods of text mining and information retrieval have been put to use in eDiscovery to help tame the volume of data; however, the results have been uneven. This chapter looks at the historical bias of the collection process. The authors examine how tools like classifiers, latent semantic analysis, and non-negative matrix factorization deal with nuances of the collection process.

INTRODUCTION

In the United States, litigation has a long history. When two parties are involved in a legal matter, the parties agree to a process of discovery wherein the particular facts of the case are uncovered. Discovery used to be a cordial exchange of legal papers and a few business documents. Today, discovery has moved into the electronic domain where the ease of digital duplication has caused volumes, and therefore costs, of documents to rise. To try and tame both cost and volume, companies and law firms have integrated information technology. When the documents became electronic (either through scanning or no longer being printed), tools from information retrieval and text mining were needed. These tools hold great promise but have had uneven results to date.
A BRIEF HISTORY OF LEGAL DISCOVERY

In countries that support discovery in the legal system, paper based discovery generally followed a simple pattern: identify the key people involved (initially referred to as witnesses and later as custodians), identify their support staff, get photocopies of the documents maintained by witnesses or from central filing systems, and send the boxes of documents to the legal team. The legal team would focus on the relevant custodians and thumb through the documents making judgment calls on the likely relevance of expense reports, budgets, and memorandums.

As volumes of documents grew, only larger law firms would continue making photocopies. In time, even those larger firms would be unable to keep up with the demands of photocopying. Without the means, ability, or desire to handle this manual job of photocopying, a market opportunity was created — the legal service provider. Specialists narrowed in on becoming litigation service providers. When a collection of paper-based documents was needed, the litigation service provider would step in and carry out the work that the law firm would not.

As self-motivated agents, the litigation service providers created a business driven by volume. Pricing for manual collections of paper documents were often priced by the page since the underlying cost was a combination of the staff required to handle the collection and the number of photocopies being made. This derivation of a service with revenues generated by volume set a pattern for future business models.

If a company found itself involved in a legal matter, it would often start by hiring a law firm. The law firm would then turn to their legal service providers to assist them in handling the matter. The litigation service provider welcomed advances in office technology that eased the creation of documents. With the gentle hum of the IBM Selectric on a secretary’s desk, the litigation service provider could count on more documents being printed and therefore more documents that needed to be photocopied.

Since the division of labor was cleanly split, law firms were not reviewing paper documents for relevance during collection. They would review the photocopies. Since the litigation service provider did not want to make legal judgments on document relevance, it behooved them to collect as much paper as they could and turn it over to the law firms for further evaluation. Thus a manual culling process carried out by legal experts was necessary. It was often too expensive to send people on repeated trips to make the physical copies of documents and they would collect and photocopy as much as possible. The drive to contain non-legal fees (like photocopying documents on repeated trips to the customer’s office) contributed to over-collection.

Information Technology Helps with Document Volume

Eventually, the volume of possibly relevant documents became cumbersome to transport and manage. Forward-looking law firms looked to information technology to help. Information technologists and vendors applied their efforts to litigation support. In 1976, “everything from PC-based database managers such as dBase and InMagic to mainframes such as Control Data’s Cyber 2000 were being used to manage portions of litigation support information. Control Data parleyed its mainframe work into Litigation Support Services (now Quorum Legal Services), one of the first companies to perform fee-based automated litigation support services.” (Griffin, 1996) Quorum began in 1966 and Kroll OnTrack eventually acquired Quorum in 2004 (Howie, 2009).

As information technology made inroads in litigation support, documents were scanned into images (stored as Tagged Image File Format or TIFFs), allowing hundreds of documents to be stored on floppy disks and whole cases on CD.
Related Content

XRecursive: Connecting XML with Relational Databases
[www.igi-global.com/article/xrecursive-connecting-xml-relational-databases/68376?camid=4v1a](www.igi-global.com/article/xrecursive-connecting-xml-relational-databases/68376?camid=4v1a)

Design of Information Spaces and Retrieval of Information Using Electrostatics in Virtual Spaces

Biometrics: Electronic Resources Management in Jaykar Library, University of Pune, India
[www.igi-global.com/chapter/biometrics-electronic-resources-management-jaykar/75910?camid=4v1a](www.igi-global.com/chapter/biometrics-electronic-resources-management-jaykar/75910?camid=4v1a)

Context-Aware Mobile Search Engine
[www.igi-global.com/chapter/context-aware-mobile-search-engine/64433?camid=4v1a](www.igi-global.com/chapter/context-aware-mobile-search-engine/64433?camid=4v1a)