Generating Value from Textual Discovery

Peter Jackson

Chief Scientist, Thomson Corporation, 610 Opperman Drive, Eagan, MN 55123, USA
peter.jackson@Thomson.com

Abstract. Much of the information associated with legal, financial, medical and educational domains is stored in electronic document repositories and retrievable only by full text search. The ability of publishers to enhance such documents through traditional editorial processes struggles to keep pace with the volume and variety of textual data currently available in proprietary collections and on the Web. Fortunately, text mining tools that support the automatic classification, summarization, and linking of documents can be developed and deployed cost effectively. The outcome is a more flexible and dynamic approach to meta-data generation that does a better job of supporting the searching and browsing behaviors of information consumers. This paper describes the application of text and data mining techniques to legal information in a manner that enables powerful report generation and document recommendation services.

Keywords: Text Mining, Information Retrieval.

1 Introduction

Legal information is a primary example of a domain where language occupies a position of peculiar importance. Whether one thinks of the actual words of a statute, the exact phrasing of a judicial opinion, or the carefully constructed clauses of a contract, the linguistic details are extremely important, and constitute the true data of the document. Legal publishers such as Thomson West have traditionally been very scrupulous in the analysis, classification and annotation of documents, using highly qualified editors to read, sort and summarize them using a strict methodology.

As data volumes and document types proliferate, with everything from briefs to blogs being added to the mix, it is clear that the same care cannot be lavished upon the full range of information sources in a cost effective manner. Arguably, it ought to be possible for automation to aggregate, annotate and deliver documents of interest to users from myriad sources at a reasonable price. Furthermore, these processes should be able to support information services in addition to search, such as report generation and document recommendation.

There seem to two non-mutually exclusive views of what text mining is. To most writers, it is either (1) discovering novel patterns from text data, as data mining does for databases, or (2) a preliminary step to data mining, in which you first have to extract the relevant data from text sources. Here I propose a third view, namely that text mining technology can be used to create a 'metaverse' (a meta-data universe) that helps users perform their own acts of discovery and facilitates the implementation of discovery programs.

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 746–753, 2007.

In support of this thesis, I describe two different applications of text mining technology that support very different information offerings, but draw upon a common set of tools that has been assembled at Thomson. The first is a service that generates reports concerning the litigation profile of a company; the second is a document recommendation system that enhances search results.

2 Mining People Data from the Law

In the past, legal publishers have concentrated upon the systematization of legal concepts and the documents that discuss them. Terms such as 'negligence', 'liability', and 'nuisance' have relatively precise legal meanings that are nonetheless subject to continuous exegesis and interpretation. Giving researchers access to the latest rulings and other writings that adhere to these concepts is one of the services that any quality publisher should provide.

Publishers have been less concerned with mining what I shall call 'the people side of the law': who has sued whom, who has represented whom, or who works for which law firm. Yet this orthogonal stream of legal data is of great utility in certain situations. Information about judges, attorneys and expert witnesses can be extremely helpful to litigators who are trying to formulate a winning strategy for their clients.

2.1 PeopleCite and Profiler

In 2000, we began text mining in a small way with a project called PeopleCite [2]. The idea was a very simple one, namely to consider the occurrence of a judge or an attorney's name in case law as a citation to that person's record in a legal gazetteer. Entity extraction programs identified such occurrences in the front matter of cases, and then Bayesian entity resolution programs matched those occurrences to the right database records and created links.

A few years later, we went further, by mining Jury Verdicts and Settlements to create the first comprehensive database of expert witnesses, called Profiler-EW. Expert witnesses play an important role in the resolution of many legal conflicts, but information about who testified at what trials and in which capacities exists only in textual form. To address this problem, we created a comprehensive expert witness database of 110,000 records using text mining techniques [3].

In our initial implementation of Profiler-EW, we extracted 290,000 references to expert witnesses from 300,000 trial case documents, using a part-of-speech tagger and a cascaded set of finite state transducers, which parsed expert name, geographical location, and field of expertise from the text. After the extraction of the reference records, we merged them together to create a file in which each particular expert is listed only once. Finally, we linked the profiles to professional license records, *medline* articles, and newspaper articles, as outlined in [4].

2.2 Firm360: Looking at Litigation History

Understanding the relationships that hold among cases, companies, and law firms can be important to the business of practicing law. For example, a legal counsel to some corporation might want to know which law firm they should engage to handle a difficult

intellectual property lawsuit, or a law firm might want to know what their chances are of engaging a large corporation as a new client. In such situations, it would be very useful to have access to a litigation history report that chronicles what cases a company has been involved in, either as a plaintiff or a defendant, which law firms have represented the company, and what kinds of cases they typically handle. Similarly, it would be good to know which attorney typically acts for a law firm on, say, intellectual property cases, and how many such cases this person has been involved in.

Firm360 is a product that can answer these questions, based on a litigation history database that covers over 50,000 companies. This database was created by (1) mining entities and their relationships out of case law documents, (2) resolving these entities by matching them against records for people and companies in various authority files, and (3) classifying cases to legal practice areas. Entity extraction was performed using cascaded finite automata and n-gram language models, entity resolution relied upon support vector machines and Bayesian record linkage, and cases were classified using a proprietary system called CaRE, which we describe later.

The text mining programs had access to a number of data resources in making their judgments. For example, attorney names were matched against the *West Legal Directory*, a database of some 500,000 law firms and solo practitioners. Features for matching included the Levenstein distance and cosine similarity between law firm names as well as information about geographic location, also mined from case law documents. A sister company, Thomson Financial, provided us with an authority file containing over 250,000 public, large private and pre-IPO companies. To ensure high precision with respect to company identifications, we applied a very high threshold to the *tf-idf* match score, and reviewed near misses by hand. To boost recall, we manually reviewed names that occurred more than 10 times in cases but did not match our company authority files, even though their n-gram language model scores indicated that they might be companies. Items that were judged to be actual companies were then added to the authority file.

Cases were classified to a cut-down version of West's *Key Number System*, called *KeySearch*, which contains roughly 10,000 legal topics. For the purposes of Firm360, we want a comparatively gross classification that tells us what practice area or sub-area a proceeding belongs to, so that we can identify both the nature of the suit and the specialism of the participating attorney. Our document categorization system, CaRE, is a highly scalable, multi-algorithm ensemble of programs that has performed well on taxonomies of up to 200,000 nodes (see Section 3).

The first release of Firm360 mined entities and relationships from 3 million case law opinions and 40 million dockets published between 1990 and 2006. 579,000 attorneys and 9,700 judges were linked to West Legal Directory, while 153,000 law firms and 58,000 companies were linked to authority files. Precision and recall on the identification of relationships ranged from 88% to 97%. Automatic text mining software was the only practical way to extract this amount of data cost effectively with a high degree of accuracy.

2.3 The Challenge of Event Extraction

Mining people data from case law is a non-trivial exercise, but this kind of entity extraction does not exhaust the potential of text mining in the legal domain. For

example, Jackson et al. [7] describe a multi-year research project, called History Assistant, on the extraction of rulings from court opinions and the determination of which prior cases are impacted by a new ruling. This is an extremely challenging task, but an essential one for any publisher, and one that would greatly benefit from some degree of automation.

History Assistant combined partial parsing techniques with domain knowledge and discourse analysis to extract information from the free text of court opinions. The same parser was also used to extract references to courts, dates, parties, and dockets for the purposes of determining the prior case or cases impacted by the opinion. This information was then used to generate a query that would return prior case candidates from a database of 7 million past cases.

Our editors insisted on near perfect recall, but were willing to tolerate precision in the 50% range. Our event extraction modules failed to achieve this recall goal, and at the time it seemed unlikely that further research would yield significant returns. (On the other hand, the prior case retrieval module achieved its stated goals and was been deemed worthy of further development.) It is worth asking why this problem is so hard, and what it would take to truly solve it.

A major limitation of current extraction technology is the fact that the information sought must be explicitly stated in the text. It cannot, for the most part, be merely implied by the text. This lack of an inferential capability can pose significant problems when extracting from real documents, where the writer expects the reader to be able to draw simple conclusions.

For example, some bankruptcy cases posed special problems for History Assistant. A debtor moves to convert from Chapter 7 to Chapter 13, and a creditor files a complaint to oppose this. The judge decides the case by 'finding for the plaintiff.' The program would have to perform a number of steps of reasoning to identify the outcome correctly as 'conversion denied', i.e., that the plaintiff is the creditor, that the creditor is asking for a denial of what the defendant (the debtor) is asking for, and that the Judge grants the denial.

This kind of inferential capability is still beyond the state of the art in text mining, and seems to require domain-specific representations of knowledge.

3 Content-Based Document Recommendation

In this section, we look at a different application, one that might not be thought of as involving text mining in the first instance. The primary technology employed in our document recommendation system involves text categorization, which is something that has been identified as not being text mining, *per se* [5]. However, we will argue that the use of categorization with other meta-data generation techniques does constitute an instance of what can be accomplished in this area.

The task of recommending documents to knowledge workers differs from the task of recommending products to consumers. Collaborative approaches [6], as applied to books, videos and the like, attempt to communicate patterns of shared taste or interest among the buying habits of individual shoppers. There are well-known problems with these approaches, e.g., when consumers temporarily shop for their children, but their effectiveness has been established in practice at many e-commerce sites.

Consumers of information typically rely upon more conventional classification schemes as an adjunct to search, such as browsing through tree-like structures, such as taxonomies and tables of contents, to narrow the application of queries. However, the problems with these approaches are also well known, primarily the inflexibility of purely taxonomical organizations of knowledge. Do the inheritance rights of children born out of wedlock belong under 'Wills' or under 'Infant Law', and how is the user supposed to know where to look in the tree?

Our approach to document recommendation leverages both user data and document meta-data. Legal researchers are typically confronted with an array of different sources: multiple document collections that are organized with respect to both document type and jurisdiction. When a user is searching a selected database, there are often relevant documents in other databases of which he or she may not be aware. Our document recommendation system, ResultsPlus, seeks to overcome this problem by matching the user's query, and other contextual information, against the indices and meta-data associated with such documents. Recommendations are then ranked by both relevance and popularity metrics based on user behavior.

3.1 Document Categorization for Content-Based Recommendations

The original version of ResultsPlus, released in 2002, used a blend of information retrieval and text categorization technologies to recommend secondary law materials to attorneys engaged in primary law research. Secondary materials include articles from legal encyclopedia, legal research papers, and law reviews. Primary law sources include cases, statutes, and regulations.

Research has long since demonstrated that superior automatic classification can be achieved through a combination of multiple classifiers. Both voting and averaging methods have been shown to improve performance. The rationale is that averaging reduces the classification score variance, which decreases the overlap between the scores of relevant and non-relevant documents, and therefore results in better classification. We decided to build a classification framework that allows the construction of multi-classifier systems that look at different document features along with custom meta-classifiers that determine how the results of multiple classifiers get combined and forwarded to a final decision making program. The resulting system, called CaRE (Classification and Recommendation Engine), is a generalization of CARP [1], a program that routed newly written case summaries to sections of American Law Reports for citation purposes. It has all necessary functionality for extracting features from documents, indexing category profiles, storing the profiles into databases, and retrieving them at run time for classification.

The ResultsPlus recommendation system relies on an innovative combination of traditional information retrieval, classification technologies, and editorial enhancements to generate recommendations. Candidate articles for recommendation are drawn from a pool of over 300 publications and document collections, representing over a million documents. All of these documents have been indexed by CaRE, and many of them reference, summarize or quote each other to form a web of legal facts and concepts.

Suggestions are generated in response to two different user actions on Westlaw:

1. The user searches a caselaw, analytical or statutes database.
2. The user views a specific document in the search result.

In the former case, the text of the search query is sent to ResultsPlus, along with meta-data from the top scoring search results. This is analogous to blind relevance feedback, where the initial search result is enhanced by another round of processing, except that we exploit meta-data from the first round of results, rather than just terms taken from these documents. In the second use case, the system uses selected text from the retrieved case or document as the ResultsPlus query text. In this case, we are using both text and meta-data to create an expanded query.

Since its release in 2002, ResultsPlus has been hailed as 'revolutionary' in the trade press and become a favorite of Westlaw users. It is powered wholly by machine learning technology, applied both to the content itself and to user behavior, as we shall see in the next section.

3.2 Optimizing Recommendations Based on User Behavior

The initial challenge for ResultsPlus was to provide a minimum number of relevant suggestions for a user's query. As the number of covered publications expanded and the pool of relevant documents for each query grew, the challenge evolved into presenting the best from among many good candidate suggestions in the limited real estate of the user's screen. We set out to leverage the significant amount of usage data collected daily to develop new ranking algorithms based on data mining approaches, and then test their performance on real users in controlled experiments.

In order to achieve this goal, we had to develop infrastructure that would perform the following essential tasks.

1. Collection of potentially useful data. The data mining algorithms we use depend on a rich data set that describes users, user groups, user sessions, and user actions.
2. Effective data mining. Our object was a system that discovers the relationships and features that drive optimal results, letting the data drive the tuning process.
3. Dynamic testing and rollout capability. To test each new ranking method, we use A/B split testing for a few days and compare the results to the current baseline.

The outcome of our experiments is a system that recommends a personalized list of the most relevant documents in response to each user's request, taking into account meta-data and click-through statistics about the user's historic usage of the system, preferences for particular content types, search context (e.g., jurisdictions and databases searched), and the recommended document's global click-through rate. The use of personalized ranking has significantly increased the click-through rates of ResultsPlus, driving both usage and revenues. Meanwhile, the daily implicit feedback mechanism continually modifies document rankings to improve relevance and enhances overall system utility for all users.

4 Conclusions and Forward Look

I will end by making some general remarks about the generation and further leverage of meta-data.

In Thomson, we often use the term 'meta-data' rather loosely, to encompass all extraneous data associated with documents, including hand-written summaries,

programmatically-generated classifications, or citation patterns within and across collections. But meta-data is really *machine-readable* data about data, and not all of the so-called meta-data types enumerated above are in fact machine-readable, in the sense of being fully interpretable by a program. After all, a hand-written summary may be no easier for a program to understand than the original text.

Furthermore, an external data source, such as an authority file, only counts as meta-data if you can systematically relate it to the underlying texts to which it is relevant. *West Legal Directory* used to be a marketing tool; it was only after we linked judge and attorney names to it (as described in Section 2) that it became a useful meta-data repository. Once we were able to generate authority files from scratch, as with our expert witness database, we reached another level in our ability to go from text to data and back again, systematically relating people and documents.

In the course of our experiments, we have made four broad discoveries about meta-data. One is that even meta-data that are not machine-readable, such as summaries, can still be extremely valuable to a text mining program. These ancillary data can serve several functions inside a machine learning regimen, e.g., as document surrogates, as sources of normalized language, as objects for clustering, and so on. Another is that almost any meta-data is better than none. For example, if we are classifying a pool of documents to a new taxonomy, and these documents are already classified to another quite different taxonomy, the latter classifications can still assist in the learning of the former. Thirdly, we have shown that quite strong transfer effects take place between document types if training takes place on good sources of meta-data. Thus we managed to train a document classifier to sort law firm documents successfully using case summaries as the input to the learner, even though these do not closely resemble the types of document found in a law firm.

Fourth and finally, we have found that, constructing upon a firm foundation, one can automatically build a 'pyramid' of meta-data layers that deliver more and more value to the end user. Thus, we were able to construct CaRE profiles for legal concepts on top of West's *Key Number System* that were applicable beyond the usual case law documents. Subsequently, we could use these profiles to supplement secondary law materials with case law references in a continuous and timely fashion. These meta-data associations were then used for a range of other tasks, such as classifying law firm documents, sorting attorneys into practice areas, and driving document recommendations.

I see 'discovery' in the context of text mining as being not so much about finding novel patterns in data, as one would in data mining, but about building a web of relationships among documents of diverse types in a way that dissolves the usual content silos that a user has to contend with. Thus, ResultsPlus does not care where you are searching on Westlaw, or what kind of document you are currently looking at; it only cares about the associative paths between documents that it can find dynamically in pursuit of relevant information. These pathways include explicit citations within the universe of documents, but do not depend upon them exclusively, leveraging instead a 'metaverse' of relationships contained in the meta-data layers we have built.

In summary, I think that publishers have the opportunity to move away from a world in which editorial intervention, hand-built taxonomies, and domain knowledge are required for every advance. We can still avail ourselves of topical classifications,

citations, summaries, and the like, but the patterns these make in the data can now be processed by machines that improve search, make connections, offer recommendations, and the like. The resulting applications can take into account all the information that would be available to an omniscient searcher who knew every feature of the portal and possessed a panoramic view of the document collections. To me, that is the kind of innovation that text mining has the power to unleash.

References

1. Al-Kofahi, K., Tyrrell, A., Vachher, A., Travers, T. & Jackson, P. Combining multiple classifiers for text categorization. In *Proceedings of the 10th International Conference on Information and Knowledge Manage Management (CIKM-2001).*, New York: ACM Press. (2001). 97-104
2. Dozier, C. & Haschart, R. Automatic extraction and linking of personal names in legal text. In *Proceedings of RIAO-2000 (Recherche d'Informations Assistée par Ordinateur)*, (2000). 1305-1321.
3. Dozier, C., Jackson, P., Guo, X., Chaudhary, M. & Arumainayagam, Y.. Creation of an expert witness database through text mining. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL-2003)*, New York: ACM Press. (2003) 177-184
4. Dozier, C. & Jackson, P. Mining text for expert witnesses. *IEEE Software*, May/June, (2005). 94-100.
5. Hearst, M. A. Untangling text data mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, (1999) 3-10.
6. Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), (2004) 5-53.
7. Jackson, P., Al-Kofahi, K., Tyrrell, A. & Vachher, A. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150, (2003) 239-290.