IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 25, 2018, accepted July 10, 2018, date of publication July 23, 2018, date of current version August 15, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2859052

Legal Decision Support: Exploring Big Data Analytics Approach to Modeling Pharma Patent Validity Cases

VIJU RAGHUPATHI[1], YILU ZHOU[2], AND WULLIANALLUR RAGHUPATHI[2]
[1]Koppelman School of Business, Brooklyn College, The City University of New York, New York, NY 10031, USA
[2]Gabelli School of Business, Fordham University, New York, NY 10458, USA

Corresponding author: Viju Raghupathi (vraghupathi@brooklyn.cuny.edu)

ABSTRACT This exploratory research examines the potential for applying a big data analytic framework to the modeling and analysis of cases in pharmaceutical patent validity brought before the U.S. Court of Appeals of the Federal Circuit. We start with two specific goals: one, to identify the key issues or reasons the Court uses to makes validity decisions and, two, to attempt to predict outcomes for new cases. The ultimate goal is to support legal decision-making with automation. The legal domain is a challenging one to tackle. However, current advances in analytic technologies and models hold the promise of success. Our application of Hadoop MapReduce in conjunction with a number of algorithms, such as clustering, classification, word count, word co-occurrence, and row similarity, is encouraging, in that the results are robust enough to suggest these approaches have promise and are worth pursuing. By utilizing larger case data sets and sample sizes and by using deep machine learning models in text analytics, more breakthroughs can be achieved to provide decision support to the legal domain. From an economic standpoint, the potential for litigation cost reduction is another objective of our study. Synergies are obtained in applying lessons to the computational field and vice versa, leading to acceleration in our understanding.

INDEX TERMS Big data analytics, Hadoop MapReduce, legal decision making, machine learning, pharma patent validity.

I. INTRODUCTION

The pharmaceutical industry is at a crossroads. With no new blockbuster drugs on the horizon, innovator firms are changing their strategies. Many are focusing on reducing operational costs by cutting R&D budgets, closing labs, and downsizing; some are consolidating through mergers and acquisitions; and yet others are forming loose partnerships with startups and biotech companies. In the meantime, generic manufacturers have entered the arena, creating a general perception that widespread availability of generic drugs will result in a drastic reduction in the cost of drugs [1]–[8]. In such a changing environment it is imperative for key players to examine the effect of current pharmaceutical patent law, such as the Hatch-Waxman Act, and the court's decisions on innovation in the pharmaceutical industry. This is important because the decisions of the court can have a major impact on drug discovery and innovation in terms of whether a drug patent is held as valid or invalid. And the question of validity has consequences for whether a drug comes to market; and when a patent expires, resulting in generic drug development. In shaping pharmaceutical patent policy on upholding or invalidating patents, the district courts and the federal courts along with the U.S. Federal Court of Appeals for the Federal Circuit play key roles [9], [10]. This suggests that regulatory agencies, patent law experts, and pharmaceutical companies should attempt to understand the decision-making processes of the courts. For example, they may examine a pattern of reasoning that often invalidates a patent. Generic drug companies can then carefully design their patent applications to enable enforcement, so as to increase the validity of the patents and prevent potential litigation in an uncertain environment [11]–[13]. That said, patent litigation analysis is complex, poorly structured, imprecise, and extremely costly, and a number of data sources and models impinge on the decision-making process (e.g. whether to litigate or settle). Computerized decision support addresses these hurdles with its potential to clarify the processes and guide stakeholders toward making more effective decisions. Much progress

has been made in exploring computerized approaches to legal decision support, including via basic case search systems, databases, discovery systems, artificial intelligence applications, machine learning, and other computational models [14]–[17]. But this progress is not without its own challenges. Legal decision-making is interdisciplinary and multidisciplinary by nature, and it's a highly complex domain with considerable amounts of free text or unstructured data, which does not lend itself well to computational modeling. However, in the past decade, two important advances have raised hopes that building meaningful, intelligent models of legal decision processes is viable. These developments include rapid advances in hardware and software that enable the storage and processing of large amounts of data, as well as increased understanding of artificial intelligence and machine learning models [18], [19]. Given progress in both text analytics and processing of unstructured data, this is an opportune time to combine the two activities and gain deeper insight into legal decision-making. Key objectives of computerized legal decision support research include the following: the ability to rapidly process large amounts of legal and other text data; examination of the steps involved in the decision-making process; identifying associations and correlations between and among legal facts, the law, cases, and outcomes; the ability to forecast court decisions; and moving from 'syntax-driven' approaches to 'semantics-driven' approaches [20]. The overall goal of computerized legal decision support is to reduce both the costs of litigation and uncertainty. As we see it, automated support can help parties make 'litigate or settle' decisions while mitigating costs, uncertainty, likelihood of failure in the courts, public burden, etc. [21].

The current research is exploratory and empirical in the application of big data analytics methods to the legal decision-making domain. In the context of pharma patents, a large number of cases are available at various levels of courts (e.g., District Court, U.S. Court of Appeals, etc.) that form a rich corpus for analysis. These cases can be retrieved quite easily using online legal databases such as Westlaw. In our research described here, we conducted preliminary analysis of the corpus of cases to gain knowledge into the decision processes and concepts involved in upholding or invalidating patent cases in the U.S. Court of Appeals. We start to identify the legal basis (keywords) for the decision and we attempt to model the outcome itself. Our research provides initial insight into the patent validity decision process, and forms the basis for future ongoing research. The ultimate goal is to provide legal decision support to parties in litigation so they can come to a settlement versus incurring uncertainty and costs, and overburdening the court system.

We apply the big data analytic framework Hadoop MapReduce and several machine learning algorithms to the analysis of a repository of pharmaceutical patent validity cases drawn from the U.S. Court of Appeals for the Federal Circuit in *Westlaw* for the period 2008 to 2011. After careful review of each case in data processing, some of the cases were removed from the data set since they were found to be outside the realm of this research (e.g., cases about medical devices etc.). We limited the case set to only drug patent cases. It is important to bear in mind that in the legal context, each patent case consists of a large body of text spanning several pages of rich textual information. Additionally, the content is domain intensive. Cumulatively, the case data exhibit the big data characteristics of complexity, volume and variety. More importantly, our objective is to demonstrate the potential and promise of big data techniques and applications in the legal domain, specifically, patent case analysis. Research has the potential to offer predictive decision support in the long run (that is, given the facts and laws of a new case, predict what the likely outcome would be), and analyze past patent decisions as an exploratory step so future research can build and incorporate additional elements (such as other qualitative and quantitative factors that include business considerations, court propensity, judges' demographics, and others).

The rest of this paper is organized as follows. Following this introduction we discuss the pharmaceutical patent case validity domain; third, we describe the current state of computerized support for the legal decision-making domain; fourth, we address the applicability of big data analytics to the problem domain; fifth, we discuss our methodology including results and analysis; sixth, we describe scope and limitations, as well as challenges; and finally, we offer conclusions and future research.

II. PHARMACEUTICAL PATENT VALIDITY

There is consensus among healthcare policy experts, the pharmaceutical industry, and regulatory agencies that promoting innovation pharmaceuticals requires patent protection [22]. Historically, the pharmaceutical industry has researched and discovered drugs that are vital to supporting health and healing. However, the public perception of the industry is often less than satisfactory. This negative impression springs from a general sense that drug manufacturers are all about making huge profits, and that this explains why medications and treatments are so expensive [23]. The contrary view is that the high cost of drugs is driven, not by the profit motive, but by the costs involved with developing and bringing a drug to market. So how does pharma balance the costly and uncertain R&D necessary for discovering and developing new drugs, with making them more affordable around the world? In addressing this question within the context of the innovator-generic paradox, the patenting process plays a critical role. As a reaction to the 1984 Federal Court decision in Roche Products, Inc. versus Boar Pharmaceutical Co., Congress introduced the Drug Price Competition and Patent Term Restoration Act of 1984, referred to as the Hatch-Waxman Act. As a modification to the prior 1952 Patent Act, this created statutory exemptions for certain claims relating to patent infringement. The provision, in 35 U.S.C. §271(e) (1), stipulates that: ''It shall not be an infringement to make, use, offer to sell, or sell within the United States a patented invention... solely for uses reasonably related to the development and submission of information under a

Federal Law which regulates the manufacture, use or sale of drugs or veterinary biological products." As a result of this provision, generic manufacturers could work on unbranded versions of an approved drug at any time in the patent lifecycle, as long as the work complies with the FDA regulations. Since then there have been several amendments to the Hatch-Waxman Act, resulting in the Medicare Prescription Drug and Modernization Act of 2003. According to the newer act, if a certification is challenged within 45 days by a New Drug Application holder or patent owner, it will put into effect an automatic 30-month stay of FDA approval of the generic ANDA (Abbreviated New Drug Application) product. As an encouragement for generics to file certifications, the act automatically confers a 180-day period of marketing exclusivity to the first person to file an ANDA with a certification. Considering the scale and complexity of the drug patenting process, courts play a significant role in deciding the winners and losers. This study attempts to shed light on the process by looking at the decisions of the U.S. Court of Appeals for the Federal Circuit, where all appeals from the District Courts are heard. By understanding the decision processes of this Court, various stakeholders can better grasp the enforcement and validity of pharmaceutical patents. A key aspect in this initial research is to discover whether machine learning can elicit the key issues (keywords) used as a basis by the Courts to validate or invalidate a patent. We next provide a brief overview of the key issues.

A. KEY ISSUES IN DECISIONS OF U.S. COURT OF APPEALS FOR FEDERAL CIRCUIT

One research question we address in this study is: *When validating or invalidating a patent, what are some key issues the Court is ruling about?* We do this by applying a series of machine algorithms, including clustering, word count, and word- co-occurrence in a Hadoop MapReduce platform. We first describe the theoretical and legal reasons or issues [22] the Court considers in making decisions. The presence or lack of obviousness is the key issue, followed by the presence or lack of written description. Lack of enablement, doctrine of equivalents, anticipation, term extension, safe harbor provision, counter claim provision, and inequitable conduct are some other reasons cited by the courts. Overall, the top two reasons why patents are ruled invalid include the lack of obviousness, and the lack of written description.

1) OBVIOUSNESS (NONOBVIOUSNESS)

In order to be patentable, a pharma invention has to be adjudged novel and nonobvious. Novelty is when an invention is not wholly anticipated by prior art or public domain materials such as publications and other patents. Prior art pertains to materials, referred to as references, that include evidence of actual use or sales of a technology within the U.S., as well as documentary materials such as patents and publications [12]. Nonobviousness is when an invention is beyond the abilities of a person with ordinary skills in the art, in the appropriate field [12].

2) WRITTEN DESCRIPTION

A patent specification has to specify in a clear and concise manner the written description of the invention, the nature and process of making the invention, and of using it. The specification should be clear enough for a person who is familiar and skilled in the art, to be able to make and use it [22]. This is done in the context of what the inventor had in mind (e.g. with regard to the design). The courts have started using the written description as a key doctrine in evaluating the validity of a patent.

3) ENABLEMENT (LACK OF)

A relationship exists between the utility requirement and the enablement requirement of the law [22]. Clearly, one cannot describe how to use an invention if that invention is useless. Therefore, arguments about the lack of usefulness of the invention are often made in the context of both the utility requirement and the enablement requirement.

4) ANTICIPATION

In order to obtain a patent, an inventor must create something new. This basic concept is known as 'novelty.' The resolution of novelty questions under U.S. patent law involves a two-part analysis. The first determination is whether a single source of information— such as a journal article or earlier patent— fully describes the claimed invention. When an invention has been completely described in a qualifying source of information, it is said to be 'anticipated' and no patent can be issued. The standard of anticipation is strict. Each and every element of the claimed invention must have been disclosed. In addition, that source of information must enable persons of skill in the art to put the disclosed information into practice [22]. Second, assuming that a fully anticipatory source of information exists, it must be determined whether it is permissible for that source to be cited against a patent or patent application, a particular journal article, earlier use of the invention, or other source of information, for it to qualify under the law. Anticipation requires the presence in a single prior art disclosure of each and every element of the claimed invention [22].

5) DOCTRINE OF EQUIVALENTS

The doctrine of equivalents was created to prevent attempts at unscrupulous copying of patents by making unsubstantial changes to an invention that could take it outside the literal scope of the patent's claims [24]. Patent claims are composed of words, and the courts have recognized the fact that words may not always be the optimum medium for conveying inventive concepts. The doctrine reflects the attempts of courts to broaden the scope of patent protection sufficiently to incentivize inventors to publicly disclose their innovation [24].

6) COUNTERCLAIM PROVISION

The Hatch-Waxman Act provides for a narrow counterclaim by a pharmaceutical company to a product of a generic

manufacturer in an infringement action by the former. This counterclaim is possible only when the original patent claimed all approved methods of use of the tested drug. The counterclaim does not work if the original patent did not claim all possible uses of the drug.

7) SAFE HARBOR PROVISION

According to the law, if one application is filed for two or more independent and distinct inventions, the Director of the U.S Patent and Trademark Office (USPTO) can restrict the application to only cover one invention. If the other inventions are filed in a divisional application that complies with the legal regulations, then these inventions can retain the filing date of the original parent application [25]. The process of drug discovery intrinsically involves the use of materials or methods that are patented, thereby exposing the manufacturer to the possibility of potential infringement suits. The safe harbor provision, in this respect, provides a safety net for companies, from patent infringement actions, during the development of a pharmaceutical product. According to this provision, activities are not deemed infringing, even if they utilize patented materials prior to the expiration of the patents, provided they relate to testing for the purpose of introducing an equivalent generic substitute drug [25].

8) TERM EXTENSION

The term extension statute for patents aims to bring a balance between the sometimes competing interests of conducting research to introduce new drugs versus assisting in producing low cost, alternative, generic copies. To get FDA approval for a generic drug, the manufacturer can file an ANDA (abbreviated new drug application) in place of a full NDA (new drug application) that would require information on safety or efficacy of the drug. The ANDA relies on the safety or efficacy studies of the original manufacturer, once the equivalency of the generic drug is proven. According to the Hatch-Waxman Act, a pioneer manufacturer of a drug should notify the FDA of all patents that claim the proposed drug [11], [25].

9) JUSTICIABLE CONTROVERSY

Justiciability concerns the limits on legal issues over which a court can exercise its judicial authority. It includes, but is not limited to the legal concept of standing, which is used to determine if the party bringing the suit is a party appropriate to establishing whether an actual adversarial issue exists [27].

10) INEQUITABLE CONDUCT

An inequitable conduct arises when a patent applicant performs activities that amount to a breach of duty of candor and good faith to the U.S. Patent and Trademark Office. These activities include non-submission of known and relevant prior art; non-submission of explanations of references or incomplete submissions of pre-existing translations of references in a foreign language; misrepresentation of facts on patentability in affidavits; and inaccurate descriptions of authorship [28].

11) LACHES

This represents a legal doctrine that says that a claim or right will not be enforced if there has been a long delay in asserting it, and such delay has prejudiced the adverse party, making it a kind of 'legal ambush' [29]. The doctrine is a defense mechanism to prevent a party from resorting to legal ambush after he/she fails to file a claim in a timely manner [29]. The premise is that, with the passage of time, the opposing party's ability to produce witnesses or other evidential material wanes due to unavailability or fading/loss of memory. The justification for this defense is that law should not support those who do not act (sleep) on their rights. The only way for this defense to succeed is if the party invoking the doctrine can prove that they were forced to change the position due to the delay, and as a result were in a worse position now than at the time when the claim should have been filed [29].

III. COMPUTERIZED DECISION SUPPORT

Computational intelligence applications in legal decision-making include a wide range of applications (e.g., case prediction, legal research databases, compliance, contract analysis, document automation, Artificial Intelligence, machine learning, expert systems, neural networks that provide decision support, and others) [30]–[33]. All of these timesaving applications enable lawyers to devote less time gathering data, and more time applying the law. Intelligent systems can help in speeding up the process of mining documents during discovery and due diligence, addressing routine questions, predicting case outcomes with smart data searches, and drafting contracts. Additionally, they ensure that the work is cost-efficient and accurate. Lex Machina, an IP litigation research company, uses data mining and predictive analytics for forecasting the outcomes of IP litigation. Over the years, the company has expanded its data set to include court dockets and enhance its data insight and prediction capabilities [30], [32], [33].

The sustained application of computers and computational intelligence in legal decision-making has been going on for several decades. This is evidenced by the very existence of the journal, *Artificial Intelligence and Law* (http://www.springer.com/computer/ai/journal/10506) and the biannual International Conference on Artificial Intelligence and Law (http://www.iaail.org/?q=page/ai-law), as well as numerous review articles [14], [34]–[36]. Models and applications have focused on a wide range of topics, including argumentation, case-based learning, logic, representation, reasoning, text analytics, and others [37]–[41]. And they have been applied to a variety of legal domain areas [42]–[49]. The rationale for using computers to model legal decision-making processes is twofold: one, to drive a more cohesive and multi-disciplinary study of legal decision-making; and two, to advance our knowledge of computational methods. Understanding more about the ways legal decision makers reach their decisions and, as a result, providing computerized support for the process, can lead to

several practical benefits. These benefits include reducing legal costs via automation; resolving or settling issues without involving the court system; a more nuanced understanding of the dynamics, processes, multiple perspectives, stakeholder positions, bargaining, and negotiation of the law; and making decisions based on highly reliable outcome predictions. The renewed focus on computation in legal decision-making is driven by several factors, including technological progress in hardware, machine learning, natural language processing, and data science; more acceptance of technology in the legal profession; easy online access to large amounts of legal data; emerging success of Artificial Intelligence in general; and the recognition of the transformational role of technology [17], [21].

The focus of modeling legal reasoning is on cases. There is also continued focus on capturing reasoning with cases in rule based systems [43]. However, the legal domain possesses many characteristics that render this endeavor challenging. Some of these characteristics include diverse types of knowledge from theories, rules, cases, procedures, norms, hierarchy of authorities, and meta-rules. Cases can include precedents – situations that were litigated and decided at the trial court level and whose decisions have been appealed at other court levels [45], [50].

Recently, in the legal domain there is renewed focus on research that adopts a data centric approach. This refers to problem solving with knowledge gained from legal documents or other large data sets [38]. The influx of statistical techniques for analyzing large data sets facilitates pursuing this approach. By providing the ability to process large sets of legal data such as assortment of court cases or of statutes of a particular jurisdiction, these techniques offer new insight into perspectives such as graph topologies of citation networks, probability of court decisions and case outcomes, and evolution of legal doctrines [38], [51]–[53].

In the legal domain, retrieving and interpreting information from a very large case base poses a formidable challenge. In this context, information extraction tools can mine large amounts of text and identify relationships within and among cases. Additionally, as new cases are added to the case base, they become immediately available to legal researchers who are working on cases. Automated extraction, through identification of semantic properties or relationships, brings an element of structure to unstructured, machine-readable text [41].

Another focus of this study is case based approach. Case based AI models of legal reasoning generate reasonable arguments and counter-arguments based on precedents. The task of predicting the outcomes of new legal cases is both historical and empirical. It is historical in that there may be similar past cases that have been decided by courts. It is empirical in that statistical or symbolic machine learning can be applied to a database of classified cases containing features that strengthen or weaken the classification. Through such application, it can generate basic rules for classifying new cases and predicting outcomes [39], [40]. Drawing from existing literature, we next discuss big data analytics and its potential for analyzing large corpuses of patent validity cases.

IV. BIG DATA ANALYTICS

Big data analytics, with its storage and maintenance facilities for large data sets of structured or unstructured data, has the potential to model legal decision making [52], [53]. Big data typically encompasses an exceptionally large volume of collected, stored, and managed data that is assembled for the purpose of gaining insight and making informed decisions. This would typically include the large corpus of cases and statutes, among others. Not surprisingly then, big data analytics (with its array of architectures, platforms, technologies, programming languages and open source tools), has good application potential in the domain of legal decision making characterized by large repositories of narrative based cases [54].

Underlying this emerging sub-discipline is the application of distributed processing to handle the complexity, volume and real-time nature of analytics [18], [19]. Historically, the legal profession generates large amounts of textual data (and documents) that require special computational processing, namely, text analytics. A combination of large storage and speed, as well as advanced machine learning text analytics is required to model legal decision-making processes [51]. The availability of this combination has opened up opportunities for analyzing the ocean of data for modeling legal decision making.

Just as in other domains, the applications of big data analytics in law is characterized by three characteristics: volume, velocity and variety [56], [57]. (http://www-01.ibm.com/software/data/bigdata/). Obviously, data (e.g. cases) will continue to be created and amassed, resulting in an incredible *volume* of data. Meanwhile, this data is being accumulated in real-time and at a rapid pace, or *velocity*. Finally, there is a transition in the way in which data is collected and stored - from standard quantitative data stored in spreadsheets or relational databases, to multimedia data stored as unstructured text [55]–[57]. Therein lies *variety*. Analytics techniques have had to evolve as well, to keep pace with the volume, velocity and variety [55], [57] of data. Research has introduced a fourth characteristic called veracity or data assurance [58]. This indicates the error-free and credible nature of the data, analytics and the outcomes. Veracity also assumes that the infrastructure (such as architecture, platforms, algorithms, methodologies, and tools) scale up in performance to match the demands of big data [59]–[61]. An example is the way in which big data analytics is implemented. It uses distributed processing with multiple servers, and incorporates the architecture of parallel computing and the modular approach of divide-and-execute. There is a vast difference between traditional intelligence tools and analytic tools for structured and unstructured big data. Big data is characterized by robust and scalable architectures and tools [59]–[61]. Similarly, the models and techniques (such as data mining, statistical approaches, algorithms, and

visualization techniques) should in turn be cognizant of the characteristics of the big data analytics.

With the exception of the manner in which processing is executed, the conceptual framework for a big data analytics project resembles that of any traditional business intelligence project [62]. Unlike in a regular business intelligence project in which processing can be performed on a standalone computer, in a big data project, it needs to be distributed across multiple nodes [62]. Distributed processing, by itself, is not a new concept. What is novel is its application to large data sets in the legal domain to assist in legal decision making [52], [53]. The availability of open source platforms on the cloud, such as Hadoop/MapReduce, facilitates the application of big data analytics to several domains - law, being one of them [55]. There are differences in terms of the friendliness of the user interface between big data analytic tools and traditional analytic tools. While the traditional tools are becoming more user-friendly by the day, the big data analytic tools are extremely complex and programming-intensive. This primarily arises from the inherent complexity in the data itself - being that it comes from multiple sources (internal and external), in multiple formats (text, flat files, relational tables, ascii files), and resides in multiple locations (legacy and other applications) [55]–[59], [62]. For analytics purposes, this data, which includes cases, rules, statutes, doctrines, and so on, has to be pooled. The pooled data is then cleansed and prepared for processing, using the phases of extract, transform, and load (ETL). Based on the structured or unstructured nature of the data, appropriate data formats are then input into the Hadoop/MapReduce platform [18], [19].

The next stage in the conceptual framework includes selection of the approach to data input, distributed design, choice of tool and selection of analytic data models. Following this is the incorporation of typical applications of big data analytics in terms of queries (discovery), machine learning (and statistics) and reports (summary of similar cases) [62]. The unifying theme for all these applications is visualization. In this research we offer a variety of techniques for aggregating, manipulating, analyzing and visualizing big data. Note that legal data is predominantly free-text or unstructured data.

Hadoop (Apache) is the largest open source distributed platform for big data analytics [62]. Hadoop has the capability to process vast amounts of data through data partitioning. This is a process in which the data is divided into partitions and each partition is assigned to a different server/node for processing. The processed pieces are then integrated to form the final solution (http://hadoop.apache.org). In this way, Hadoop lends itself to being both an organizing and an analytic tool. It enables businesses to tap into large repositories of data that, traditionally, were unmanageable and non-analyzable [62]. There are two important modules in Hadoop [61]. The first is the Hadoop Distributed File System (HDFS) that offers as a facilitator for the storage for the Hadoop cluster. Once data arrives in the cluster, the HDFS partitions it into parts/chunks and redistributes it across different servers in the cluster. Only a small portion/chunk of the entire data sits in each server/node, and gets replicated in other nodes. The second module of importance in Hadoop is the MapReduce. In Hadoop, just as the data are stored in a distributed fashion across multiple connected servers/nodes, the analytic tasks are also divided into sub-tasks and each sub-task is assigned to a node/server. After processing, all of the results from the various sub-tasks are then aggregated to produce a unified solution [61]. MapReduce is the module that provides the interface for the distribution of sub-tasks and assimilation of outputs. The distinct benefits of parallel/distributed processing include the ability of graceful degradation and of coping with probable failures [61]. In this context, the HDFS and MapReduce are both configured to be fault-tolerant. By continuously monitoring the server/nodes and storage devices, the HDFS, can detect an issue, and immediately reroute the data to an alternate functional node/server. Also, the fact that the data are replicated across nodes/servers confers an additional layer of redundancy and backup [61]. In the same way, MapReduce continuously monitors the tasks across servers/nodes and in the event of an anomaly (such as a hiatus, reduced speed or dead end), can instantly redirect the task to another node/server that holds the duplicate data. The synergistic combination of HDFS and MapReduce in the Cloud environment contributes to fault-tolerant support for storage and analytics [57], [61]. In the realm of big data analytics, Hadoop is typically used to find patterns in unstructured data using tasks such as correlation or cluster analysis [61].

Cloudera (https://www.cloudera.com/products/open-source/apache-hadoop.html) is an example of a platform that offers scalable and flexible integration interface, facilitating management of large volumes and varieties of data in an enterprise. Cloudera enables deployment and management of Apache Hadoop and related projects in terms of manipulating and analyzing, data and keeping it protected [63]. This is a reason why we use Cloudera in this study. Next, we describe our methodology and provide our preliminary results.

V. METHODOLOGY

A search was conducted in Westlaw for the period 1/1/2008 to 2/15/2011 to retrieve all pharmaceutical patent cases in the U.S. Court of Appeals for the Federal Circuit. From the corpus of cases generated from the search, cases that did not pertain to pharma drug patents were eliminated. For example, patents regarding medical devices were not relevant to our study. A case-by-case analysis was done to ensure relevance to the domain of pharmaceutical drugs. The research objective was to gain insight into the cases collectively to elicit keywords that help to characterize the 'reason' for patent validation or invalidation, potential correlations between relevant words that characterize the cases, and uncover the similarities between cases. Mindful of the challenges and difficulties in applying machine learning techniques to 'free text' or 'unstructured' data (narrative), our expectations were narrowed. We determined that it would be an achievement to demonstrate, in principle, the application of big data

analytics and machine learning concepts to legal decision-making. In future, we can build on current and past knowledge.

In terms of methodology, we first explored the potential of clustering and classification models by applying them to the set of cases. Clustering helped identify the keywords or reasons for patent validity, and classification helped classify new cases into these keywords. We used Python NLTK library to eliminate punctuations, numbers, and standard stop words. In order to increase the likelihood that the keywords would be more meaningful, additional stop words - "claim," "application," "issue," and others - were customized. The Snowball stemming algorithm was used to combine different words that have the same meaning. For instance, we stemmed "validity" and "validation" to "validity." Once the stemming process was completed, the text files with the stemmed words served as the input to the application of the K-means clustering, the word count, and word co-occurrence algorithms in Map Reduce in the Cloudera Hadoop system. In addition, the cases' pdfs were converted into sequence files to build the Document Term Matrix by TF-IDF index for clustering. Then, the Apache's Mahout library was used to apply the clustering method and transfer the outputs from the HDFS to the local machine. As Figure 1 shows, the keywords that represent the Court's reasons, (discussed in Section 2), are grouped into the clusters, along with the frequency of cases in each cluster. The model was able to elicit 'obviousness' as an important reason, and others such as written description, and inequitable conduct. The first cluster, with eleven cases, includes multiple keywords: term extension, prosecution of laches, doctrine of equivalents, etc. The clustering terms in figure 1 (marked with red circles) are representative of the issues summarized in section 2. This reflects the accuracy of the technique in identifying and addressing key elements in the domain.

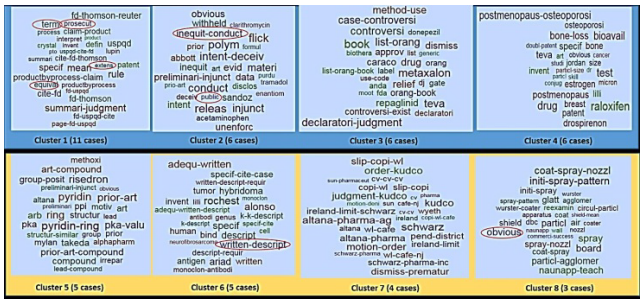


FIGURE 1. Keywords (patent validity/invalidity reason) in cases.

Next, we built the Naïve-Bayes classification model to classify the validity cases into three groups: invalid, partially valid, and valid. The training set consists of approximately 50 percent of the total number of cases. These included proportion of invalid, partially valid, and valid patent decisions. Likewise, the testing set included appropriate sample of invalid, partially valid, and valid patent decisions, in order to avoid the over-fitting problem and improve the model performance on test data.

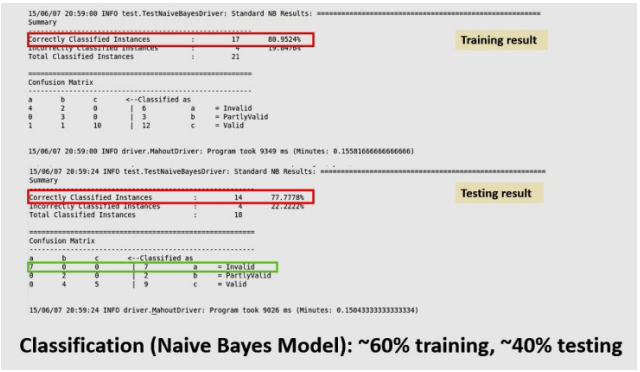


FIGURE 2. Classification using Naive-Bayes model.

As seen in Figure 2, the model has an 80.95% overall accuracy on the training data set and 77.78% overall accuracy on the testing data set. Meanwhile, the model works well at predicting invalid cases (7/7) but not so well with regard to valid cases (5/9). Since the court typically lists all the reasons for invalid cases, this could explain why there are more keywords that describe invalid cases than valid cases, resulting in higher accuracy for predicting invalid cases. With large samples and testing data sets and iterative refinement of the models, higher correct classification rates can be achieved.

One can map and reduce data based on a variety of criteria. A common example is the Java WordCount class. As the name suggests, WordCount maps (extracts) the words in the input and reduces (summarizes) the results with a count of the number of instances of each word. By reading the text files, Wordcount gives a count of how often a word occurs. It looks at the frequency of occurrence of words in a text file. The input and output are both text files. In the output file, each line contains a word and a number denoting the frequency of its occurrence, both separated by a tab. Every mapper looks at a line as input and parses it into words, printing out a key value of pair of the word and 1. Each reducer then totals the counts for each word and emits a single value containing the word and 1. The reducer is used as a combiner on the map outputs. This optimizes the amount of data transmitted across the network, by combining each word into a single record [64], [65].

The word cloud plot in Figure 3 displays the words most frequently found in the corpus of cases. The larger the size of a word, the more often it occurs in the corpus. From the plot, we can see that such words as district, product, invent, requirement, and evidence are quite distinct in the case data set, which confirms common sense understanding of the law cases. Such words as invalid, pharmaceutical, patent, and drug also have high frequencies. This affirms what we know via manual analysis and reading of the key implicit issues in the whole data set.

A co-occurrence matrix indicates how, while tracking an event in a given window of time or space, one can identify

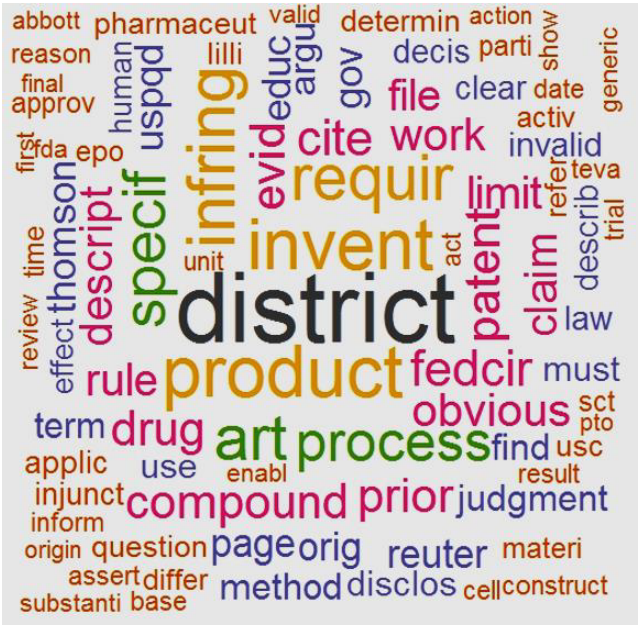


FIGURE 3. Word cloud.

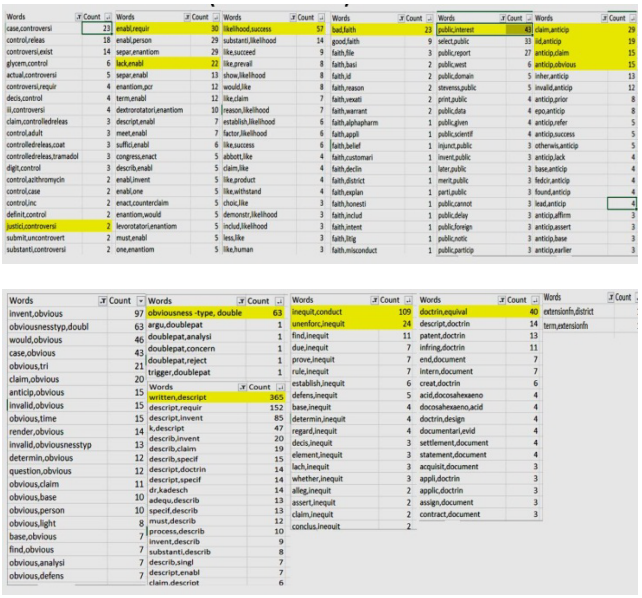
the occurrence of other events in the same window. In the current context, events are individual words found in the text. The matrix tracks what other words occur within the window that represents a position relative to the target word.

In linguistics, co-occurrence represents the likelihood of occurrence of two terms in a certain order alongside each other within a large corpus of data (textual) [66]. In this sense, it is used as an indicator of semantic closeness of terms [66].

Figures 4a, 4b and 4c show how many times two words come together in the sample. As we see from the above results, in Excel spreadsheet, the keyword 'obvious-type double' shows 63 times in the data set. Key word 'inequitable conduct' shows 109 times, 'Doctrine Equivalents' shows 40 times, and so on. On the other hand, some interesting words are shown together many times, like 'anticipate' and 'obvious' (15 times), 'bad faith' (23) and 'public interest' (43).

Term Frequency Inverse Document Frequency (TF-IDF) is a technique that searches through a corpus of documents and determines which words are favorable to use in a query. For each word in a document, a value is calculated as an inverse proportion of the frequency of occurrence of the word in the document, to the percentage of documents the word occurs in [67]. A high TF-IDF indicates a strong relationship with the document in which it occurs. This implies that if the word were used in a query, then the document would be relevant and of interest to the user [67]. Using this technique ensures efficient query retrieval with relevant words [67].

The row similarity analysis indicates the parallels between cases. Figure 5 shows groups of cases that are similar or related to each other. For example, the first file (key 0) appears to have some similarity with the files corresponding to keys 10, 9, 8, 7, 29, 4 and 25 that are shown in Figure 6.



Words	Count
case,controversi	23
control,releas	18
controversi,exist	14
given,control	6
actual,controversi	5
controversi,requir	4
decis,control	4
ii,controversi	4
claim,controlledreleas	3
control,adult	3
controlledreleas,coat	3
controlledreleas,tramadol	3
digit,control	3
control,azithromycin	2
control,case	2
control,inc	2
definit,control	2
justici,controversi	2
submit,uncontrovert	2
substanti,controversi	2

Words	Count
enabl,requir	30
enabl,person	29
separ,enantiom	29
lack,enabl	22
separ,enabl	13
enantiom,pcr	12
term,enabl	12
dextrorotator,enantiom	10
descript,enabl	7
meet,enabl	7
suffici,enabl	6
congress,enact	5
describ,enabl	5
enabl,invent	5
enabl,one	5
enact,counterclaim	5
enantiom,would	5
levorotatori,enantiom	5
must,enabl	5
one,enantiom	5

Words	Count
likelihood,success	57
substanti,likelihood	14
like,succeed	9
like,prevail	8
show,likelihood	8
would,like	8
like,claim	7
reason,likelihood	7
establish,likelihood	6
factor,likelihood	6
like,success	6
abbott,like	4
claim,like	4
like,product	4
like,withstand	4
choic,like	3
demonstr,likelihood	3
includ,likelihood	3
less,like	3
like,human	3

Words	Count
bad,faith	23
good,faith	9
faith,file	3
faith,basi	2
faith,id	2
faith,reason	2
faith,vexati	2
faith,warrant	2
faith,alphapharm	1
faith,appli	1
faith,belief	1
faith,customari	1
faith,declin	1
faith,district	1
faith,explan	1
faith,honesti	1
faith,includ	1
faith,intent	1
faith,litig	1
faith,misconduct	1

Words	Count
public,interest	43
select,public	33
public,report	27
public,west	6
public,domain	5
stevenss,public	5
print,public	4
public,data	4
public,given	4
public,scientif	4
injunct,public	3
invent,public	3
later,public	3
merit,public	3
part,public	3
public,cannot	3
public,delay	3
public,foreign	3
public,notic	3
public,particip	3

Words	Count
claim,anticip	29
id,anticip	19
anticip,claim	15
anticip,obvious	15
inher,anticip	13
invalid,anticip	12
anticip,prior	8
epo,anticip	8
anticip,refer	5
anticip,success	5
otherwis,anticip	5
anticip,lack	4
base,anticip	4
fedcir,anticip	4
found,anticip	4
lead,anticip	4
anticip,affirm	3
anticip,assert	3
anticip,base	3
anticip,earlier	3

Words	Count
invent,obvious	97
obviousnesstyp,doubl	63
would,obvious	46
case,obvious	43
obvious,tri	21
claim,obvious	20
anticip,obvious	15
invalid,obvious	15
obvious,time	15
render,obvious	14
invalid,obviousnesstyp	13
determin,obvious	12
question,obvious	12
obvious,claim	11
obvious,base	10
obvious,person	10
obvious,light	8
base,obvious	7
find,obvious	7
obvious,analysi	7
obvious,defens	7

Words	Count
obviousness -type, double	63
argu,doublepat	1
doublepat,analysi	1
doublepat,concern	1
doublepat,reject	1
trigger,doublepat	1

Words	Count
written,descript	365
descript,requir	152
descript,invent	85
k,descript	47
describ,invent	20
describ,claim	19
describ,specif	15
descript,doctrin	14
descript,specif	14
dr,kadesch	14
adequ,describ	13
specif,describ	13
must,describ	12
process,describ	10
invent,describ	9
substanti,describ	8
describ,singl	7
descript,enabl	7
claim,descriptor	6

Words	Count
inequit,conduct	109
unenforc,inequit	24
find,inequit	11
due,inequit	7
prove,inequit	7
rule,inequit	7
establish,inequit	6
defens,inequit	5
base,inequit	4
determin,inequit	4
regard,inequit	4
decis,inequit	3
element,inequit	3
lach,inequit	3
whether,inequit	3
alleg,inequit	2
assert,inequit	2
claim,inequit	2
conclus,inequit	2

Words	Count
doctrin,equival	40
descript,doctrin	14
patent,doctrin	13
infring,doctrin	11
end,document	7
intern,document	7
creat,doctrin	6
acid,docosahexaeno	4
docosahexaeno,acid	4
doctrin,design	4
documentari,evid	4
settlement,document	4
statement,document	4
acquisit,document	3
appli,doctrin	3
applic,doctrin	3
assign,document	3
contract,document	3

Words	Count
extensionfn,district	
term,extensionfn	

Words	Count
counterclaim,provis	32
assert,counterclaim	1
counterclaim,infring	5
enact,counterclaim	5
defens,counterclaim	4
counterclaim,defend	3
counterclaim,seek	3
file,counterclaim	3
ii,counterclaim	3
limit,counterclaim	3
ad,counterclaim	2
agre,counterclaim	2
counterclaim,amend	2
counterclaim,amgen	2
counterclaim,assert	2
counterclaim,avail	2
counterclaim,defendantapp	2
declaratori,counterclaim	2
entitl,counterclaim	2
hold,counterclaim	2

Words	Count
chevron,defer	19
entitl,chevron	
chevron,framework	2
chevron,skidmor	2
chevron,usa	2
give,chevron	2
see,chevron	2
accord,chevron	1
act,chevron	1
apart,chevron	1
applic,chevron	1
case,chevron	1
chevron,appli	1
chevron,applic	1
chevron,brand	1
chevron,domain	1
chevron,inquiri	1
chevron,nation	1
chevron,us	1
chevron,well	1

Words	Count
safe,harbor	16
safeti,efficaci	9
safe,effect	3
addit,safeguard	2
entitl,safe	2
occup,safeti	2
prod,safeti	2
reli,safeti	2
statutori,safe	2
address,safeti	1
comprehens,safeti	1
conclus,safe	1
deem,safe	1
demonstr,safeti	1
determin,safe	1
efficaci,safeti	1
evid,safeti	1
featur,safe	1
human,safe	1
less,safe	1

FIGURE 4. Word co-occurrence.

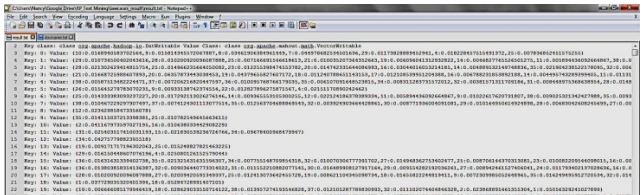


FIGURE 5. Row similarity.

This kind of analysis is also useful in grouping similar cases and further examining the commonalties between them.

VI. SCOPE AND LIMITATIONS

This exploratory study is not without limitations. First, this is an exploratory study that seeks to demonstrate the concept and feasibility of applying big data analytic methods and machine learning algorithms to the analysis of what is historically recognized as a challenging domain, namely,

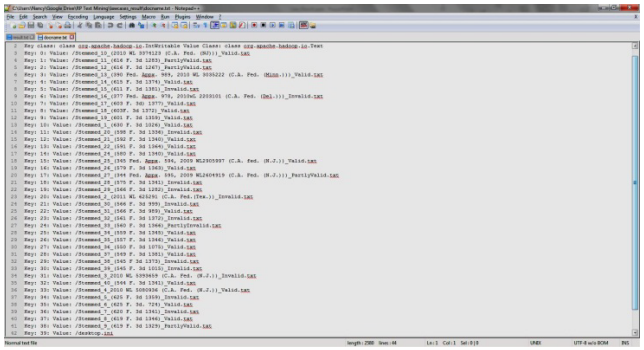


FIGURE 6. Document reference.

legal decision-making. Second, our study is limited to patent validity cases in general, and pharma patent cases in particular. Long-term research should address generalizability to all legal cases. Surmounting the domain-induced difficulties - such as large amounts of unwieldly free text and diversity of key issues, is difficult. It is only now that researchers are awakening to the prospect of mining large amounts of 'free text' data. Third, the sample size here is limited in terms of number of cases, but sufficiently large in terms of volume of data. Nevertheless, future research can be done on larger data sets (more cases). In addition, the number of pharma patent cases that go on to appeal in the U.S. Court of Appeals is few. Fourth, additional experiments and alternative machine learning and artificial intelligence techniques may be applied to improve predictability. Lastly, while our research focuses on analyzing historical patent cases with a view to providing insight on legal decision making, advanced predictions are possible by correlating former patent decisions with future new case facts.

VII. CONCLUSIONS AND FUTURE RESEARCH

Overall, results in this exploratory study are promising. A Hadoop MapReduce approach demonstrates the efficacy of the distributed approach to handling large amounts of unstructured text data. Likewise, the machine learning algorithms based off the Cloudera big data Hadoop MapReduce platform show promise. The model was able to surface out many of the key issues examined by the Courts such as obviousness, written description, etc. The classification was quite reasonable. Nevertheless, much work remains to be done.

We set out to demonstrate the feasibility of applying big data architectures and analytic techniques in conjunction with machine learning to what is essentially a complex domain, namely, legal decision-making. Our study affirms the promise of new and advanced methods in extracting insight from a corpus of pharma patent validity cases. We acknowledge that the study is limited in scope, and future studies need to be ongoing and extensive, building more comprehensive computational models of legal decision-making using alternative techniques that include artificial intelligence and neural networks. Generalizability and scalability to other patents will also challenge research going forward. The overall long-term goal is to support legal decision-making with automation, thereby improving quality and reducing costs of litigation.

REFERENCES

[1] C. Cookson. (Dec. 13, 2017). *Pharma Industry's Return on R&D Investment Falls Sharply*. [Online]. Available: https://www.ft.com/content/b020be56-e00a-11e7-a8a4-0a1e63a52f9c
[2] A. Jack, "Genzyme move eases way for Sanofi–Aventis," *Financial Times*, vol. 18, Feb. 2011.
[3] A. Jack, "New Pfizer Chief's remedy unlikely to cure longer-term ills," *Financial Times*, vol. 16, Feb. 2011.
[4] A. Jack, "Eli Lilly funds will aim to share costs and benefits of drug R&D," *Financial Times*, Feb. 2011.
[5] B. Kendall, "White house seeks to speed up generic drugs,' path to market," *Wall Street J.*, Feb. 2011.
[6] S. Neville and R. Atkins, "Roche ready to take its medicine over copycat drugs," *Financial Times*, vol. 14, Jan. 2018.
[7] G. Vina. (2016). *Returns on Big Pharma Research and Development Hit Six-Year Low*. Accessed: Dec. 12, 2016. [Online]. Available: https://www.ft.com/content/530b5626-c072-11e6-9bca-2b93a6856354
[8] D. Wilson, (2011). *Drug Firms Face Billions in Losses in '11 as Patents End*. Accessed: Mar. 6, 2011. [Online]. Available: http://www.nytimes.com/2011/03/07/business/07drug.html
[9] R. Schulman, "Is it harder to enforce pharmaceutical patents?" *Nat. Law J.*, Aug. 2006.
[10] H. Grabowski, C. Brain, A. Taub, and R. Guha, "Pharmaceutical patent challenges: Company strategies and litigation outcomes," *Amer. J. Health Econ.*, vol. 3, no. 1, pp. 33–59, 2017.
[11] Lex Machina (Apr. 26, 2016). *Pharmaceutical Patent Litigation Filings Have Risen Significantly Since 2014, According To Lex Machina's 2015 Hatch-Waxman/Anda Report*. Accessed: Jan. 5, 2018. [Online]. Available: https://lexmachina.com/media/press/pharmaceutical-patent-litigation-filings-risen-since-2014/
[12] C. Neumeyer. (2013). *Managing Costs of Patent Litigation*. [Online]. Available: http://www.ipwatchdog.com/2013/02/05/managing-costs-of-patent-litigation/id=34808/
[13] D. Scott. (2016). *Why an Obscure Supreme Court Case is a Big Deal for Prescription Drugs*. [Online]. Available: https://www.statnews.com/2016/04/22/supreme-court-patent-case-drugs/
[14] T. Bench-Capon, "A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law," *Artif. Intell. Law*, vol. 20, no. 3, pp. 215–319, 2012.
[15] D. Houlihan, "Ross intelligence and artificial intelligence in legal," Blue Hill Res., Boston, MA, USA, Res. Rep. A0280, Jan. 2017, pp. 1–2.
[16] S. Miller. (Oct. 5, 2017). *Artificial Intelligence and Its Impact on Legal Technology (Part IV)*. [Online]. Available: https://abovethelaw.com/2017/10/artificial-intelligence-its-impact-on-legal-technology-part-iv/
[17] J. Sobowale, (Apr. 2016). *How Artificial Intelligence is Transforming the Legal Profession*. [Online]. Available: http://www.abajournal.com/magazine/article/how_artificial_intelligence_is_transforming_the_legal_profession
[18] B. Schoenborn, *Big Data Analytics Infrastructure for Dummies*. Hoboken, NJ, USA: Wiley, 2014.
[19] A. Sathi, *Big Data Analytics*. Boise, ID, USA: MC Press, 2012.
[20] Owen Byrd. (Dec. 8, 2017). *What is Legal Analytics, and How is it Relevant to the Practice of Commercial law?* [Online]. Available: http://www.lawtechnologytoday.org/2017/12/legal-analytics-commercial-law/
[21] A. Marwaha, (Jul. 13, 2017). *Seven Benefits of Artificial Intelligence for Law Firms*. [Online]. Available: http://www.lawtechnologytoday.org/2017/07/seven-benefits-artificial-intelligence-law-firms/
[22] J. R. Thomas, *Pharmaceutical Patent Law*, 2nd ed. Arlington, VA, USA: BNA Books, 2010.
[23] R. Frank and P. B. Ginsburg. (Nov. 13, 2017). *Pharmaceutical Industry Profits and Research and Development*. [Online]. Available: https://www.healthaffairs.org/do/10.1377/hblog20171113.880918/full/
[24] J. R. Allison and M. A. Lemley, "The (unnoticed) demise of the doctrine of equivalents," *Stanford Law Rev.*, vol. 59, no. 4, p. 955, 2007.
[25] C. M. Davidson. (Jun. 2004). *The Safe Harbor Provision of Hatch–Waxman: Is There a Hole in the Safety Net? Drug Development & Delivery*. [Online]. Available: http://www.drug-dev.com/Main/Back-Issues/The-Safe-Harbor-Provision-of-HatchWaxman-Is-There-429.aspx

[26] U.S. Food and Drug Administration. (Oct. 23, 2014). *Types of Applications*. [Online]. Available: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/default.htm

[27] MacGuireWoods. (Oct. 31, 2006). *Generics Still Unable to Resolve ANDA Patent Issues by Declaratory Judgment, But is Supreme Court Resolution on the Way?* [Online]. Available: https://www.mcguirewoods.com/news-resources/publications/commercial_litigation/ANDA_patent_issues.pdf

[28] L. Petherbridge, J. Rantanen, and A. Mobiji, "The federal circuit and inequitable conduct: An empirical assessment," *California Law Rev.*, vol. 84, pp. 1293–1356, 2011.

[29] J. S. Welch, Jr., "Closing the laches: Does the split decision in the *Raging Bull* case finally bring some consistency to the doctrine of laches in copyright infringement?" *Southern Law J.*, vol. 26, no. 1, pp. 59–77, 2016.

[30] D. Garcia. (Jun. 7, 2017). *Preparing for Artificial Intelligence in the legal profession*. [Online]. Available: https://www.lexisnexis.com/lexis-practice-advisor/the-journal/b/lpa/archive/2017/06/07/preparing-for-artificial-intelligence-in-the-legal-profession.aspx

[31] P. Gunst. (Dec. 7, 2017). *Legal Tech: The AI Generation*. [Online]. Available: https://www.legalfutures.co.uk/blog/legal-tech-ai-generation

[32] S. Lohr. (Mar. 19, 2017). *A. I. is Doing Legal Work. But it Won't Replace Lawyers, Yet*. [Online]. Available: https://www.nytimes.com/2017/03/19/technology/lawyers-artificial-intelligence.html

[33] M. Mills. (2016). Artificial Intelligence in Law: The State Of Play 2016. Thomson Reuters, Legal Executive Institute. [Online]. Available: timecontent/uploads/2016/04/Artificial-Intelligence-in-Law-The-State-of-Play-2016.pdf

[34] B. G. Buchanan and T. E. Headrick, "Some speculation about artificial intelligence and legal reasoning," *Stanford Law Rev.*, vol. 23, no. 1, pp. 40–62, 1970.

[35] N. Love and M. M. Genesereth, "Computational law," in *Proc. Int. Conf. AI Law (ICAIL)*, Bologna, Italy, Jun. 2005, pp. 205–209.

[36] E. L. Rissland, K. D. Ashley, and R. P. Loui, "AI and law: A fruitful synergy," *Artif. Intell.*, vol. 150, nos. 1–2, pp. 1–15, 2003.

[37] V. Aleven, "Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment," *Artif. Intell.*, vol. 150, pp. 183–237, Nov. 2003.

[38] L. K. Branting, "Data-centric and logic-based models for automated legal problem solving," *Artif. Intell. Law*, vol. 25, no. 1, pp. 5–27, 2017.

[39] M. Curtotti, E. McCreath, T. Bruce, S. Frug, W. Weibel, and N. Ceynowa, "Machine learning for readability of legislative sentences," in *Proc. Int. Conf. AI Law (ICAIL)*, San Diego, CA, USA, 2015, pp. 53–62.

[40] C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij, "A method for explaining Bayesian networks for legal evidence with scenarios," *Artif. Intell. Law*, vol. 24, no. 3, pp. 285–324, 2016.

[41] A. Wyner, R. Mochales-Palau, M.-F. Moens, and D. Milard, "Approaches to text mining arguments from legal cases," in *Semantic Processing of Legal Texts*. Berlin, Germany: Springer, 2010, pp. 60–79.

[42] L. Al-Abdulkarim, K. Atkinson, and T. Bench-Capon, "Factors, issues and values: Revisiting reasoning with cases," in *Proc. 15th Int. Conf. Artif. Intell. Law (ICAIL)*, 2015, pp. 3–12.

[43] T. Bench-Capon and G. Sartor, "A model of legal reasoning with cases incorporating theories and values," *Artif. Intell.*, vol. 150, pp. 97–143, Nov. 2003.

[44] F. Bex, H. Prakken, B. Verheij, and T. van Engers, "Introduction to the special issue on artificial intelligence for justice (AI4J)," *Artif. Intell. Law*, vol. 25, no. 1, pp. 1–3, 2017.

[45] S. Bruninghaus and K. D. Ashley, "Predicting outcomes of case based legal arguments," in *Proc. 9th Int. Conf. AI Law*. New York, NY, USA: ACM Press, 2003, pp. 233–242.

[46] W. M. Campbell, L. Li, C. K. Dagli, K. Greenfield, E. Wolf, and J. P. Campbell, "Predicting and analyzing factors in patent litigation," in *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, Barcelona, Spain, 2016, pp. 1–6.

[47] J. G. Conrad and K. Al-Kofahi, "Scenario analytics: Analyzing jury verdicts to evaluate legal case outcomes," in *Proc. Int. Conf. Articial Intell. Law*, London, U.K., 2017, pp. 29–38.

[48] M. Surdeanu and S. Jeruss, "Identifying patent monetization entities," in *Proc. ICAIL*, Rome, Italy, Jun. 2013, pp. 201–205.

[49] M. Taddeo, A. Trombetta, D. Montesi, and S. Pierantozzi, "Querying data across different legal domains," in *Proc. IDEAS*, Barcelona, Spain, Oct. 2013, pp. 192–197.

[50] P. Jackson, K. Al-Kofahi, A. Tyrrell, and A. Vachher, "Information extraction from case law and retrieval of prior cases," *Artif. Intell.*, vol. 150, pp. 239–290, Nov. 2003.

[51] B. Marr. (Jan. 20, 2016). *How Big Data is Disrupting Law Firms and the Legal Profession*. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2016/01/20/how-big-data-is-disrupting-law-firms-and-the-legal-profession/#20839f6f7c23

[52] A. Maurushat, L. Bennett-Moses, and D. Vaile, "Using 'big' metadata for criminal intelligence: Understanding limitations and appropriate safeguards," in *Proc. Int. Conf. AI Law (ICAIL)*, San Diego, CA, USA, Jun. 2015, pp. 196–200.

[53] J. O. McGinnis and B. Stein, "Originalism, hypothesis testing and big data," in *Proc. Int. Conf. AI Law (ICAIL*, San Diego, CA, USA, Jun. 2015, pp. 201–205.

[54] W. Raghupathi and V. Raghupathi, "Big data analytics—Architectures, implementation methodology, and tools," in *Big Data, Mining, and Analytics*, S. Kudyba, Ed. New York, NY, USA: Taylor & Francis, 2014, pp. 49–70.

[55] B. Brown, M. Chui, and J. Manyika, (Oct. 2011). *Are You Ready for the Era of Big Data? McKinsey Quarterly*. [Online]. Available: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/are-you-ready-for-the-era-of-big-data

[56] J. Hagerty and T. Groves, *Unlock Big Value in Big Data With Analytics*. New York, NY, USA: IBM Redbooks, 2013.

[57] P. Zikopoulos, D. Deroos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles, *Harness the Power of Big Data The IBM Big Data Platform*. New York, NY, USA: McGraw-Hill, 2013.

[58] F. J. Ohlhorst, *Big Data Analytics: Turning Big Data into Big Money*. Hoboken, NJ, USA: Wiley, 2012.

[59] D. Loshin, *Big Data Analytics: From Strategic Planning to Enterprise Integration With Tools, Techniques, NoSQL, and Graph*. Waltham, MA, USA: Morgan Kaufmann, 2013.

[60] K. H. Pries, *Big Data Analytics: A Practical Guide for Managers*. Boca Raton, FL, USA: CRC Press, 2015.

[61] P. Zikopoulos, C. Eaton, D. Deroos, T. Deutsch, and G. Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY, USA: McGraw-Hill, 2012.

[62] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf. Sci. Syst.*, vol. 2, p. 3, Feb. 2014. [Online]. Available: http://www.hissjournal.com/content/2/1/3

[63] P. R. Chelliah, "The hadoop ecosystem technologies and tools," in *Advances in Computers*, Amsterdam, The Netherlands, 2017, doi: 10.1016/bs.adcom.2017.09.002.

[64] Apache Hadoop. (Nov. 18, 2017). *MapReduce Tutorial*. [Online]. Available: https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[65] DeZyre. (2017). *MapReduce Tutorial—Learn to Implement Hadoop Word Count Example*. [Online]. Available: https://www.dezyre.com/hadoop-tutorial/hadoop-mapreduce-wordcount-tutorial

[66] Y. Yi, L. Liu, C. H. Li, W. Song, and S. Liu, "Machine learning algorithms with co-occurrence based term association for text mining," in *Proc. 4th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Mathura, India, 2012, pp. 958–962.

[67] J. Ramos, "Using TF-IDF to determine word relevance in document queries," Dept. Comput. Sci., Rutgers Univ., Piscataway, NJ, USA, Tech. Rep., 2003.

VIJU RAGHUPATHI received the Ph.D. degree in information systems from The Graduate Center, The City University of New York. She is currently an Associate Professor with the Koppelman School of Business, Brooklyn College, The City University of New York. She has published in academic journals, including *Communications of the Association for Information Systems*, the *Journal of Electronic Commerce Research*, *Health Policy and Technology*, the *International Journal of Healthcare Information Systems and Informatics*, *Information Resources Management Journal*, and *Information Systems Management*. Her research interests include business analytics, social media, big data, innovation/entrepreneurship, sustainability, corporate governance, and healthcare.

YILU ZHOU received the B.S. degree in computer science from Shanghai Jiao Tong University, and the Ph.D. degree in management information systems from The University of Arizona. She was an Assistant Professor of information systems and technology management with the School of Business, George Washington University. She was also a Research Associate with the Artificial Intelligence Laboratory, The University of Arizona. She is currently an Associate Professor of information systems with the Gabelli School of Business, Fordham University. She has published in academic journals, including MIS Quarterly, the *Journal of the American Society for Information Science and Technology*, the IEEE Intelligent Systems, and *Decision Support Systems*. Her research interests include business intelligence, computational data analytics, text mining, multilingual knowledge discovery, and human–computer interaction.

WULLIANALLUR RAGHUPATHI is currently a Professor of information systems with the Gabelli School of Business, Fordham University, New York, the Program Director of the M.S. in Business Analytics Program, and the Director of the Center for Digital Transformation. He has published 40 journal articles and written papers in refereed conference proceedings, abstracts in international conferences, book chapters, editorials, and reviews, including several in the healthcare IT field. He was the Founding Editor of the *International Journal of Computational Intelligence and Organizations* (1995–1997). He also served as an Ad Hoc Editorial Review Board Member of the *Journal of Systems Management* of the Association for Systems Management, from 1996 to 1997. He is a Co-Editor for North America of the *International Journal of Health Information Systems and Informatics*. He has also guest edited (with Dr. Joseph Tan) a special issue of Topics in *Health Information Management* (1999) and a special section on healthcare information systems for the *Communications of the ACM* (1997).