Задача поиска неточностей в электронных коллекциях судебных актов

© А.А. Рогов, Ю.В. Сидоров, И.Л. Бурлак

Петрозаводский государственный университет
rogov@psu.karelia.ru, sidorov@onego.ru, burlak@psu.karelia.ru

Аннотация

В данной работе рассмотрены предпосылки возникновения ошибок, несоответствий и неточностей в электронных хранилищах судебных актов и их влияние на потребителей данной информации. Изучается конкретная задача автоматического определения категории арбитражного спора.

1. Введение

В настоящее время очень большое внимание руководством страны уделяется открытости судебной системы Российской Федерации. Одним из этапов, направленных для достижения этой цели, является законодательно закрепленная с 1 июля 2010 года публикация всех судебных актов (за некоторым исключением) в сети Интернет. Таким образом, можно говорить о том, что в сети Интернет появился новый специализированный ресурс, представляющий собой большой массив текстовой структурированной информации, который, несомненно, будет востребован как специалистами в области юриспруденции, так и простыми гражданами и общественными организациями.

Так, например, специализированные на обработке юридической информации компании ("Консультант +", "Кодекс", "Гарант" и т.д.), на базе своих технологических платформ производят обработку данной информации, реализуя различные сервисы навигации по судебной практике, моделирования правовых ситуаций и т.п., которые доступны их подписчикам за определенную плату.

Кроме того, современные поисковые системы индексируют размещаемые на сайтах судов судебные акты, и в ответ на пользовательские запросы выдают наряду с обычными документами по данной тематике (например, научными или аналитическими статьями, нормативными актами и т.п.) также и ссылки на судебные акты. Ясно, что это требует определенных усилий от пользователя для последующего анализа и обработки получаемых результатов.

С другой стороны, в электронных коллекциях такого масштаба практически неизбежны некоторые ошибки или неточности, в исправлении которых естественным образом будут заинтересованы как потребители данной информации, так и органы, генерирующие эту самую информацию.

С учетом вышеизложенного, становиться очевидным актуальность разработки систем автоматического анализа и поиска ошибок (неточностей) в электронных коллекциях судебных актов, что подразумевает, в том числе, разработку алгоритмов анализа судебных актов, представляющих собой структурированный текстовый документ.

2. Задача автоматического определения категории арбитражного спора

2.1 Постановка задачи

Основной целью данного исследования является разработка алгоритмов проверки "корректности" отнесения арбитражного судебного решения, размещенного на сервисе Высшего Арбитражного Суда Российской Федерации "Банк решений арбитражных судов" по адресу с сети Интернет http://ras.arbitr.ru, к определенной категории спора.

Зачастую потребители данного ресурса интересуются судебной практикой по той или иной категории юридического спора (например, аренда, ценные бумаги, товарные знаки и т.д.) за определенное время. Таким образом, ошибка при отнесении того или иного судебного решения к определенной категории спора может привести дезинформированию пользователя.

Ошибки при отнесении судебного акта к определенной категории спора могут возникнуть как вследствие технического сбоя при обработке или пересылке информации на ресурс, так и вследствие неправильного указания категории спора автором документа. Последнее зачастую может быть связано с тем обстоятельством, что правовые споры могут иметь такую сложность, что относить его только к одной категории будет не совсем корректно с юридической точки зрения. Поиск и выявление подобных ситуаций также может оказаться полезным для изучения и дальнейшего

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

принятия решения о модификации классификатора споров ответственными за его ведение лицами.

Для достижения сформулированной цели были поставлены следующие задачи:

- Разработать алгоритмы поиска судебных актов в сетевых хранилищах и их предварительной обработки.
- Разработать алгоритмы извлечения и анализа информации из структурированного текста судебного акта.
- Разработать алгоритмы интерпретации полученных результатов.

Результатом работы должен стать программный комплекс, который будет производить автоматическую выгрузку судебных актов из определенных сетевых хранилищ с последующей их классификацией по заранее определенным категориям спора. Одна из основных задач разрабатываемой системы обнаруживать и оповещать администратора о несоответствии установленной в хранилище категории спора судебного акта, с категорией определенной в ходе работы системы.

2.2 Объект исследования

В рамках данного исследования планируется обработка только судебных решений, т.е. судебных актов арбитражных судов первых инстанций, по которым заканчивается рассмотрение судебных дел по существу.

Судебное решение представляет собой структурированный текст, который условно можно разделить на четыре части:

- Вводная часть – в данной части судебного решения указываются: дата и место принятия решения суда; наименование суда, состав суда; секретарь судебного заседания; стороны и другие лица участвующие в деле, их представители; предмет спора или заявленное требование.
- Описательная часть – данная часть судебного решения содержит указание на требования (и их нормативно-правовые основания) истца, возражения ответчика и объяснения других лиц, участвующих в деле.
- Мотивировочная часть – в данной части судебного решения указываются установленные судом обстоятельства дела, доказательства, на основании которых судом делаются выводы, доводы, которыми суд отвергает или принимает те или иные доказательства. А также ссылки на законы, которыми руководствовался суд.
- Резолютивная часть – данная часть решения содержит выводы суда об удовлетворении иска либо об отказе в удовлетворении иска полностью или в части, указание на распределение судебных расходов, срок и порядок обжалования решения суда.

2.3 Задача формирования обучающей выборки

Источником данных для исследования является упомянутый выше сервис Высшего Арбитражного Суда Российской Федерации "Банк решений арбитражных судов", который является хранилищем всех судебных актов, принятыми всеми арбитражными судами Российской Федерации.

На данном этапе своего исследования, мы исходили из того, что к настоящему времени существующий сервис предоставляет только пользовательский интерфейс доступа к данным.

Проанализировав возможности использования данного интерфейса для организации автоматической выборки данных, были выявлены следующие функциональные ограничения:

- при изменении структуры пользовательского интерфейса, разрабатываемый прототип программного продукта не сможет гарантировать работоспособность и потребует переработки;
- при активном использовании разрабатываемой системой данного интерфейса повышается нагрузка на сервис, и как следствие, могут возникнуть подозрения на злоумышленные действия.

Перечисленные выше ограничения делают нецелесообразным использования данного интерфейса для организации автоматической выборки.

Поэтому было решено, что на данном этапе документы должны загружается в разрабатываемое приложение непосредственно пользователем в ручном режиме с использованием следующих параметров отбора информации: временной период, список судов, категория спора, тип судебного документа и некоторые другие параметры, используя которые можно выделить интересующую нас группу судебных решений.

Для получения обучающей и контрольной выборок были отобраны с помощью экспертов за последний год по 100 судебных решений для 10 категорий споров, существенно различающихся между собой с юридической точки зрения (разная нормативная база, использование специфической терминологии и т.п.).

2.4 Извлечение информации и построение вектора признаков

Следующий этап состоит в анализе того, какая информация, содержащаяся в каждом судебном решении, оказывает существенное влияние на принадлежность к той или иной категории спора.

Нами было выдвинуто предположение, что такой информацией является перечень ссылок на статьи нормативно-правовых актов (законов, кодексов и т.п.) в мотивировочной части судебного решения. Выделение мотивировочной части каждого судебного решения осуществлялось с

помощью шаблонов, т.е. как часть текста, заключенную между предложениями, содержащими следующие фразы "Исследовав материалы дела, суд приходит к выводу, что …" и "Суд решил". После извлечения фрагмента текста между вышеуказанными фразами или близкими им по смыслу, происходит поиск и приведение к единой унифицированной форме всех упоминаний и ссылок на нормативно-правовые акты, реализованный при помощи регулярных выражений, которые учитывают разные варианты написания. Например, считать одинаковыми по смыслу следующие написания: "ст.5", "стат. 5" и "статьёй 5" или "НК РФ" и "Налоговый кодекс Российской Федерации" и т.п.

Обработав все отобранные для исследования судебные решения, были полученные следующие параметры: Для каждой категории и для каждой статьи нормативно-правовых актов были получены: частота встречаемости в категории, возможное число употреблений в отдельных документах соответствующей категории, а также группы статей одновременно в документах категории.

Другой комплект параметров основывается на ограниченном наборе специализированных или ключевых слов (словосочетаний), уникально характеризующих ту или иную категорию спора, например, "вексель", "ипотека", "транспортная экспедиция" и т.п. Они были получены следующим образом: все слова в выделенном фрагменте были приведены к нормальной форме. Из всех слов был сделан словарь и каждому слову приписывался набор пар, содержащих номер категории и частоту встречаемости в этой категории. Представляли интерес только те слова, которые встречались не более чем в трех категориях.

2.5 Методы классификации

Задачу автоматического определения категории споров можно сформулировать в общем виде. Пусть задано некоторое множество документов D = {d_1 … d_s} и некоторое множество категории споров C = {c_1… c_m}. Каждый документ задается вектором признаков $X = (X_1, X_2, \ldots, X_n)$, который может быть получен для каждого предложения текста, где $X_i = \{ x_{i1}, x_{i2}, \ldots, x_{iri} \}, i = 1,2 \ldots n$ — множество возможных значений признака X_i. Разыскивается неизвестная функция F, которая определяет принадлежность документа к определенной категории F: DxC -> {0, 1}. Построение классификатора F’ проводиться на подмножестве документов D’ с заранее определенной категорией c_i, где i принимает значения от 1 до m. Такая выборка документов называется обучающей, и задача относится к классу задач обучения с учителем.

До решения задачи классификации была решена задача поиска информативных признаков и уменьшение размерности признакового пространства.

Для нахождения эффективного классификатора судебных решений по категориям споров, можно использовать различные метрические алгоритмы классификации (например, метод потенциальных функций, ближайших соседей и т.д.) с последующим их анализом, вариацией параметров [1,2,3,4]. Для анализа эффективности работы алгоритмов классификации потребуется привлечение экспертов в данной области. Для построения классификатора авторы использовали **Метод, основанный на индуктивном построении систем правил**, вида "ЕСЛИ ... ТО …" с весами по обучающей выборке. Данный метод предложил Чистяков С.П. для определения авторства текстов [5]. Правила используются при создании классификатора, позволяющего провести классификацию текстов или групп текстов. Классификаторы, основанные на правилах, имеют хорошую интерпретируемость решений. При классификации текст относится к той категории, к которой было отнесено большинство из составляющих его предложений.

Обозначим Y — категорийный (классовый) признак с множеством возможных значений $L = \{ 0, 1, \ldots, m-1 \}, \ m \geq 2.$ Предполагается существование неизвестного совместного распределения $P(x, y)$ признаков $X_1, X_2, \ldots, X_n, Y.$ Из распределения $P(x, y)$ имеется обучающая выборка $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}.$

Строим правила вида "ЕСЛИ <предпосылка> ТО <заключение> <с весом w>", где предпосылка имеет C вид:

$$C = \{X_{\alpha_1} = x_{\alpha_1\beta_1}\} \& \{X_{\alpha_2} = x_{\alpha_2\beta_2}\} \& \ldots \& \{X_{\alpha_r} = x_{\alpha_r\beta_r}\},$$

где $\alpha_i = 1,2,\ldots,n, \ \beta_i = 1,2,\ldots,r_i, \ i = 1,2,\ldots,r,$ то есть рассматриваем цепочки конъюнкций длины $r,$ состоящие из упорядоченных пар "признак-значение", перебирая все возможные комбинации и исключая рассмотрение в одной цепочке одного признака с разными значениями.

Заключение имеет вид: $C_i^* = \{Y = i\}, i \in L,$ то есть классовый признак в случае выполнения "предпосылки" принимает определенное значение классового признака. Вес $w \in (0,1)$ является мерой влияния предпосылки правила на заключение.

Получаются правила вида $C \Rightarrow C_i^* < w >.$ Если существует два правила с одним и тем же заключением, то для вычисления общего веса правил пользуемся формулой для комбинации весов:

$$w_1 \oplus w_2 = \frac{w_1 w_2}{w_1 w_2 + (1 - w_1)(1 - w_2)}.$$

Если $\Re$ — множество правил, то функция комбинации весов применяется к весам всех правил, входящих в $\Re$, для которых имеется одинаковое заключение. Получаем *композиционный вес* для

множества правил $W(C_i^*|C,\Re)=\oplus_\alpha w_\alpha$. Тогда множество правил $\Re$ индуцирует некоторый классификатор $f_\Re: X \to D$, который относит предложение, $x=(x_1,x_2,\ldots,x_n)\in X$, определяемое набором признаков, к тому классу, для которого композиционный вес максимальный:

$$f_\Re(x)=\arg\max_i W\left(C_i^*|C(x),\Re\right).$$

В множестве правил $\Re$ нас интересуют только те, которые по обучающей выборке показали статистически значимые отличия распределения классового признака.

При построении статистического критерия на основе классификатора использовалось модельное предположение, что набор признаков имеет биномиальное распределение. Рассматривался случай, когда классификатор использовался для построения статистического критерия проверки нулевой гипотезы H_0 о том, что некоторый текст принадлежит категории "А" против альтернативной гипотезы H_1, что текст принадлежит категории "Б". В качестве статистики критерия использовалось количество предложений, отнесенных классификатором к категории "Б".

2.6 Итоги работы

Предварительно проведенные исследования показывают более высокую точность поиска, чем результаты статьи [6].

3. Заключение

Рассмотренная в работе задача автоматического определения категории арбитражного спора не единственная, которую можно сформулировать, анализируя электронную коллекцию судебных актов. Однако, подходы к их решению, принципы извлечения текстовой информации, выбор параметров, как представляется, будут схожими с описанными в данной работе.

Литература

[1] Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. Москва, Наука, 1970.

[2] Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. Математические вопросы кибернетики, Т. 13, стр. 5–36, 2004.

[3] Воронцов К.В. Профили компактности и выделение опорных объектов в метрических алгоритмах классификации. Искусственный интеллект, № 2, стр. 30-33, 2006.

[4] Загоруйко Н.Г. Методы распознавания и их применение. Москва, Советское радио, 1972.

[5] Суровцова Т.Г., Чистяков С.П. О построении статистических критериев для атрибуции авторства литературных текстов // Вестник Санкт-Петербургского университета, серия 10, прикладная математика, информатика и процессы управления. – Вып. 3. – 2009. – С. 138-143.

[6] Губин М.В., Меркулов А.И. Автоматическое выделение гипертекстовых переходов в текстах документов. Труды Международной конференции «Диалог 2004», стр. 155-158, 2004.

Discrepancy Search in Digital Collections of Judicial Acts

© Alexander Rogov, Yury Sidorov, Ilya Burlak

In this paper preconditions of errors occurrence, inconsistencies and inaccuracies in the electronic storage of court decisions and their impact on consumers of this information are considered. The specific problem of automatic determination of a category of arbitration dispute is studied.