УДК 004.04

РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ТЕКСТОВОЙ АНАЛИТИКИ ЮРИДИЧЕСКИХ ДОКУМЕНТОВ

[1]Д. С. Зуев, [2]М. Ф. Насрутдинов, [3]А. Ф. Хасьянов

Казанский (Приволжский) федеральный университет, г. Казань, ул. Кремлевская, д. 35, 420008

[1]dzuev11@gmail.com, [2]marat.nasrutdinov@kpfu.ru, [3]ak@it.kfu.ru

Аннотация

Обсуждено использование механизмов машинного обучения, анализа естественного языка и интеллектуального поиска в области юриспруденции. Основные ожидаемые результаты – методология применения алгоритмов текстовой аналитики и семантического анализа естественного языка (NLP) в задачах управления знаниями в судебном делопроизводстве, а также других видах юридической практики. Полученные результаты могут быть применены в области образования и управления знаниями в более широком контексте, поскольку исследование лежит на стыке юриспруденции, математической и компьютерной лингвистики.

Описан прототип многоагентной системы интеллектуального анализа текстов в юриспруденции, способной на имеющейся базе данных судебных документов выявлять общие зависимости, предоставлять для ознакомления юридические дела, близкие по тематике, рекомендовать наиболее вероятные исходы судебного рассмотрения или помечать важные места, на которые следует обращать внимание при процессуальных действиях с использованием инструментов текстовой аналитики.

***Ключевые слова**: аналитика и управление данными, интенсивное использование данных, электронные библиотеки, кластеризация, классификация судебных актов, рекомендательная система, микросервисная архитектура.*

ВВЕДЕНИЕ

Целью работы является создание интеллектуальной программной системы с набором инструментов текстовой аналитики и приложений, ориентированной

на решение прикладных задач в области юриспруденции. Создание такой системы полностью лежит в русле развития и использования современных информационно-коммуникационных технологий (ИКТ) при обработке больших массивов информации и предусматривает развитие и практическую реализацию системы управления юридическими знаниями на основе семантических технологий и онтологий. Планируемые исследования соответствуют идеологии Инициативы открытых архивов (Open Archive Initiative), основное назначение которой – повысить доступность и объединить в распределенной системе взаимосвязанных хранилищ корпуса различных текстов, включая как современные источники, так и источники, ставшие историческими.

Любой документ имеет собственные особенности, и юриспруденция не является исключением. Здесь фундаментальной задачей является изучение особенностей юридических документов с целью применения известных моделей и методов текстового анализа и машинного обучения к специфическому классу документов, каковыми являются исковые заявления, судебные дела и решения. Реализуемый проект развивает названное направление исследований и предполагает разработку и внедрение современной информационной системы, ориентированной на работу с юридическими документами как с отдельным классом электронных документов.

1. АКТУАЛЬНОСТЬ ИССЛЕДОВАНИЯ

Как известно, сегодня мы являемся участниками перехода от традиционного типа общества к информационному – обществу, в котором большинство работающих занято производством, хранением, переработкой и реализацией информации, особенно высшей её формы – знаний (см., например, https://ru.wikipedia.org/wiki/Информационное_общество). Информационное общество характеризуется высоким уровнем развития ИКТ и их интенсивным использованием всеми и всюду, а целью формирования такого общества является, в частности, развитие на основе ИКТ научно-исследовательской деятельности в любых областях человеческой жизнедеятельности.

В основе ИКТ лежит информация, а сами они во многом определяют содержание, масштабы и темпы развития других технологий. В частности, развитие облачных технологий позволило принципиально изменить подходы к созданию

сложных программных систем практически для всех предметных областей. Одной из важнейших признана задача интеграции разнородных электронных ресурсов в мировое информационное пространство. Определяющей составляющей такой интеграции является процесс семантического структурирования контента. Для этого в последние годы консорциум W3C (www.w3.org) активно разрабатывает технологии Семантического Веба, в частности, на платформе XML создан широкий спектр языков разметки текста, позволяющих не только учесть специфику предметных областей, но и повысить эффективность структурирования при автоматизированной обработке информации.

Создаваемая система должна позволять участникам юридического процесса правильно проводить подготовку соответствующих судебных дел, а также осуществлять планирование судебной деятельности. Эта система ориентирована на арбитражные суды, занимающиеся рассмотрением споров, связанных с предпринимательской деятельностью. В целом наш проект направлен на развитие российского правового государства, обеспечение доступности, открытости и прозрачности правосудия, формирование у граждан правосознания, основанного на верховенстве Права.

На сегодняшний день судопроизводство в России является областью, постоянно наращивающей присутствие информационно-коммуникационных технологий в каждодневной деятельности. Сегодня в судах используется ряд программных систем, позволяющих вести документооборот в электронной форме. Тем не менее, информационная и производственная нагрузки на судей по-прежнему остаются недопустимо большими. Так, например, только в рамках инициативы по борьбе с коррупцией на судей арбитражного суда Республики Татарстан приходится более 50 судебных дел в месяц. В таких условиях без использования специализированных автоматизированных информационных систем качественное повышение эффективности работы судов просто невозможно.

2. СУЩЕСТВУЮЩИЕ РАЗРАБОТКИ

Одним из направлений разработки специализированных автоматизированных информационных систем в судопроизводстве является создание интеллектуальных систем, способных на имеющейся базе данных судебных документов выявлять общие зависимости, предоставлять судьям для ознакомления близкие по

тематике дела, рекомендовать наиболее вероятные исходы или помечать важные места, на которые судебным работникам следует обращать внимание при процессуальных действиях.

В [1] описаны основные онтологии и подходы к работе с юридическими документами с точки зрения их семантики. Известны успешные реализации подобных информационных систем за рубежом, например, система «Case Cruncher Alpha» (www.case-crunch.com), разрабатываемая в Sidney Sussex College, Cambridge и ориентированная на прогнозирование решений юридических задач. Однако имеющиеся решения не учитывают особенности русского языка и кириллической транскрипции.

В судах также имеется программное обеспечение, которое позволяет автоматизировать часть рутинных операций. Эти программные комплексы направлены либо на автоматизацию документооборота в целом, либо представляют собой широчайшие базы данных тематических документов, найти в которых необходимую информацию в сжатые сроки не всегда представляется возможным, и не используют весь спектр семантических технологий и инструментария текстовой аналитики.

Проведенные исследования по семантическому структурированию информации в других предметных областях (см., например, [2, 3]), анализ инструментов текстовой аналитики (см., например, [4]) показывают наличие всех типовых инструментов, которые можно было бы применить и в данной предметной области для реализации поставленной задачи.

ЦЕЛЬ И ЗАДАЧИ

Разрабатываемая информационная система должна позволять участникам юридического процесса правильно проводить подготовку судебных дел, а также осуществлять планирование судебной деятельности. На данном этапе система ориентирована на арбитражные суды, занимающиеся рассмотрением споров в сфере предпринимательства.

Сегодня судебная система России серьезно загружена, количество судебных споров неуклонно растет. С одной стороны, это свидетельствует о развитии правового государства, когда все большее количество споров решается в правовом поле. С другой стороны, суды являются одной из самых забюрократизиро-

ванных сфер жизнедеятельности. При этом рост количества судебных дел повышает нагрузку как на судей, так и на вспомогательных технических работников.

Судебная система – это область, где объем работы с текстовыми документами весьма значителен, а процесс принятия решения всегда должен быть понятным и прозрачным. Поэтому, особенно в условиях роста нагрузки на сотрудников судов, требуются инструменты, позволяющие осуществлять интеллектуальный анализ поступающего информационного массива. Автоматизированный текстовый анализ позволяет выделить важные признаки документов (подсудность, характер спора, участвующие стороны и т. д.), осуществить поиск в судебной базе данных и представить похожие документы, по которым уже приняты решения, или даже спрогнозировать вероятное решение суда по рассматриваемому делу. Именно на этот аспект работы судебной системы нацелена наша система.

Задача создаваемого программного комплекса – помочь определить характер спора, осуществить поиск и проверку действия правовых норм, регулирующих спорные правоотношения, оказывать содействие в установлении компетентного суда (подсудность, подведомственность), статуса участников спора (действующее, ликвидированное, банкрот), определении круга обстоятельств, имеющих значение для рассмотрения спора, характера спорного правоотношения, нормы права, подлежащей применению (действует ли данная норма), а также проверять достаточность и комплектность представляемых документов.

Для достижения поставленных целей поставлены следующие задачи и спроектирована архитектура системы [5]:

- создание портала для формирования шаблонов исковых заявлений с отслеживанием их жизненного цикла;
- разметка и анализ существующей базы судебных решений, исковых заявлений (классификация заявлений и решений, извлечение сущностей и фактов);
- подбор аналогичных дел и решений, рекомендательный сервис;
- сопоставление исковых заявлений и судебных решений;
- распределение судебных дел между судьями с учетом их специализации и текущей загрузки.

Фактически каждая из выделенных задач является автономным модулем разрабатываемой информационной системы, а сама система – практическая демонстрация совместного использования ряда семантических технологий и инструментов текстовой аналитики.

Важно отметить, что создание рекомендательной системы текстовой аналитики юридических документов с применением технологий искусственного интеллекта и использованием технологий Семантического Веба является новой задачей, аналоги подобных систем в России отсутствуют. Существующие мировые аналоги ориентированы на тексты на латинице и не работают с кириллическим текстами.

ПОДХОДЫ К РЕАЛИЗАЦИИ

Исследование свойств и возможностей специализированных языков разметки, построенных на базе расширенного языка разметки XML, является важным и актуальным, поскольку на этой базе постепенно строится концепция формального описания семантики различных понятий и явлений. То, что до сих пор не выработано универсального и всеобъемлющего подхода к описанию семантики, существенно ограничивает возможности обмена информацией, поскольку формализованным оказывается лишь синтаксис сообщений, а он лишен смысла: за обыкновенными и сходными по форме синтаксическими конструкциями может скрываться самая неожиданная семантика (смысл).

Предлагаемые нами подходы используют методы компьютерной и математической лингвистики. Анализ текста основан как на традиционных подходах, связанных с применением онтологий, так и на применении машинного обучения на базе обширного корпуса юридических документов, имеющихся в открытом доступе, включая артефакты правоприменительной практики судов различных инстанций.

Информация в системе должна храниться в формализованном и понятном компьютеру виде, сформированном на основе технологий Семантического Веба. Такой способ управления знаниями позволит создать инструменты для работы непосредственно с объектами знания (средства агрегации, семантического поиска и идентификации тождественных объектов) (см. [6–8]).

Как известно, большинство современных электронных архивов документов представляет собой наборы неструктурированных документов, на базе которых трудно организовать семантический поиск, извлечение метаинформации и различные информационные сервисы. В этом смысле архивы судов и наборы иных юридических документов не являются исключением. Кроме того, в настоящее время наблюдается значительное увеличение объема данных, включаемых в репозитории, что в свою очередь создает дополнительные трудности при обработке информации. Поэтому в условиях непрерывного роста объёмов, а также многообразия информации активно развиваются новые подходы, инструменты и методы обработки огромных объёмов данных, обозначаемых термином «большие данные» (Big Data). При управлении репозиториями электронных документов больших данных в полной мере остаются актуальными, а также появляются новые задачи, в их числе: семантическая разметка, организация поиска, выделение метаданных, формирование тематических кластеров документов, определение зависимостей, поиск близких документов и др. Насущными становятся проблемы анализа и управления данными в различных областях с интенсивным использованием данных. Часть описанных теоретических проблем применительно к юридической информации исследуется в рамках настоящего исследования.

Переход к представлению внутренней структуры знания создает новую парадигму представления, в которой основные акценты смещаются на выделение элементов (классов) и их взаимосвязей, что позволяет создавать различные сетевые концептуальные структуры (например, граф цитирования, граф концептов и др.). Выделение классов объектов и организация соответствующих репозиториев позволят создать новые вычислительные возможности по обработке данных, такие, как извлечение и обработка терминов, поиск близких результатов и т. п.

Семантический поиск позволит по введенному описанию объекта или после выделения термина в исковом заявлении получить дополнительную информацию (определение, свойства, связи с другими объектами, список документов, в которых встречается объект, указание источника, где он впервые введен). С помощью такого поиска, например, можно найти все судебные решения, в обосновании которых прямо или косвенно используются те или иные нормативные акты.

Важно отметить, что в различных документах объект может обозначаться различными терминами.

Базовые научные разработки, положенные в основу системы, разработаны его участниками в рамках таких проектов, как «Thinking and understanding» и др. (см. [9]).

Система построена с использованием сервис-ориентированной архитектуры (Service-oriented architecture, SOA), точнее, с использованием микросервисов. Напомним, что SOA – это построение среды, в которой отдается предпочтение слабым связям, абстрагированию низкоуровневой логики, гибкости, а также возможности многократного использования и обнаружения компонентов [10, 11]. Дальнейшим развитием парадигмы сервис-ориентированной архитектуры можно считать появление архитектуры микросервисов [12]. Термин «Microservice Architecture» получил распространение в последние несколько лет для описания способа проектирования приложений в виде набора независимо развертываемых сервисов.

Архитектурный стиль микросервисов – это подход, при котором единое приложение строится как набор небольших сервисов, каждый из которых работает в рамках собственного процесса и взаимодействует с остальными. Сервисы построены вокруг бизнес-потребностей и развертываются независимо с использованием полностью автоматизированной среды. Централизованное управление минимизировано, а сами сервисы могут быть написаны на разных языках программирования и использовать разные технологии хранения данных. Более того, внутри каждого микросервиса вполне может быть задействована собственная база данных (см. [12]).

С учетом достаточно большого количества модулей системы необходимо выбрать подход к организации всего приложения и минимизировать зависимости, связанные с изменениями внутри отдельных модулей. При этом очевидно, что модули текстовой аналитики со временем будут изменяться, возможна реализация различных алгоритмов классификации и аналитики в зависимости от массива обрабатываемых документов. К тому же итеративный процесс разработки системы требует простых механизмов независимого обновления различных модулей. Это обуславливает применение архитектуры микросервисов для создания системы.

Одной из важнейших задач формируемой информационной системы являются поиск и предоставление аналогичных решений по схожим судебным искам. Таким образом, необходим сервис поиска аналогичных документов, или рекомендательный сервис.

Существуют два основных типа рекомендательных систем: контент-ориентированные и социальные (см., например, [13]). Первые основаны на представлении предпочтений пользователей путем анализа содержимого рекомендательных элементов. Системы второго типа моделируют предпочтения, оценивая близость профилей пользователей. Ниже под рекомендательным сервисом будем понимать информационную систему, которая:

1) формирует модель предметной области на основе массива документов (включая подготовительные операции – приведение к векторному виду, кластеризацию и т. п.);

2) получает на вход документ и выдает список документов, близких к входному.

В разрабатываемой рекомендательной системе реализованы следующие основные этапы:

- извлечение ключевых слов из документов коллекции на основе онтологий;
- представление каждой публикации в виде вектора, компоненты которого соответствуют концептам онтологии;
- определение значения компоненты – это вес соответствующего понятия в данной статье (вычисляется с использованием количества его упоминаний в тексте статьи и количества упоминаний связанных понятий);
- использование в качестве меры близости между публикациями косинусной меры близости между их векторами.

Важным аспектом выполнения проекта является апробирование результатов на принципиально различных коллекциях электронных документов, таких, как математические, гуманитарные и юридические документы, что позволит построить универсальную модель системы, адаптировать разрабатываемые сервисы к различным документам. Полная реализация проекта позволит фактически построить программную систему управления электронными документами,

способную консолидировать управление разнородными электронными научными коллекциями, и в конечном итоге представить набор универсальных семантических сервисов анализа научных данных.

Подход, разрабатываемый в рамках настоящего проекта, полностью соответствует мировой практике перехода на электронные средства представления и обмена информацией.

ТЕКУЩИЕ РЕЗУЛЬТАТЫ

Описываемая система находится на начальном этапе разработки, на текущий момент предложена верхнеуровневая архитектура системы [5], проведен обзор функциональности имеющихся систем, работающих в данной предметной области, определен набор востребованных дополнительных сервисов для системы. Разработаны концептуальная и инфологическая модели системы. Определен жизненный цикл искового заявления как электронного документа, построены основные процессы системы – загрузка, редактирование, поиск и просмотр [5].

В рамках решения задачи классификации проведены предварительный анализ судебных документов, отбор значимых признаков для определенных категорий судебного спора, проведен латентно-семантический анализ для выявления общей структуры типовых документов. На тестовой выборке проверены алгоритмы байесовской классификации, k-ближайшего соседа и деревьев решений, разрабатывается модель на основе искусственной нейронной сети. На следующем этапе планируется увеличить выборку исковых арбитражных заявлений и рассмотреть большее число типов возможных судебных споров, а также разработать программные модули, выполняющие задачи отбора информативных признаков и классификации.

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 2.8712.2017/8.9.

СПИСОК ЛИТЕРАТУРЫ

1. *Peroni S.* Semantic Web Technologies and Legal Scholarly Publishing Law, Springer, Governance and Technology Series, 2014. V. 15. doi 10.1007/978-3-319-04777-5

2. *Елизаров А. М., Жижченко А. Б., Жильцов Н. Г., Кириллович А. В., Липачёв Е. К.* Онтологии математического знания и рекомендательная система для коллекций физико-математических документов //Доклады Академии наук. 2016. Т. 467, № 4. С. 392–395. doi: 10.1134/S1064562416020174

3. *Елизаров А. М., Липачёв Е. К., Невзорова О. А., Соловьев В. Д.* Методы и средства семантического структурирования электронных математических документов //Доклады Академии наук. 2014. Т. 457, № 6. С. 642–645. doi 10.7868/S0869565214240049

4. *Грант С. Ингерсолл, Томас С. Мортон, Эндрю Л. Фэррис*. Обработка неструктурированных текстов. Поиск, организация и манипулирование/ Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 414 с.: ил.

5. *Зуев Д. С., Марченко А. А., Хасьянов А. Ф.* Применение инструментов интеллектуального анализа текстов в юриспруденции // CEUR Workshop Proceedings. 2017. V. 2022. P. 214–218. http://ceur-ws.org/Vol-2022/paper35.pdf

6. Digital Mathematics Library: a vision for the future. International Mathematical Union, 2006. http://www.mathunion.org/fileadmin/IMU/Report/ dml_vision.pdf.

7. *Olver P. J.* What's happening with the World Digital Mathematics Library? http://www.math.umn. edu/~olver/t_/wdmlb.pdf

8. Developing a 21st century global library for mathematics research. Washington, D.C.: The National Academies Press, 2014. 131 p. arxiv.org/pdf/1404.1905; http://www.nap.edu/catalog/18619/developing-a-21st-century-global-library-for-mathematics-research.

9. *Toschev A., Talanov M.* Thinking Lifecycle as an Implementation of Machine Understanding in Software Maintenance Automation Domain// Jezic G., Howlett R., Jain L. (eds) Agent and Multi-Agent Systems: Technologies and Applications. Smart Innovation, Systems and Technologies. 2015. Vol 38. Springer, Cham. doi: 10.1007/978-3-319-19728-9_25

10. *Gold N. et al.* Understanding Service Oriented Software. IEEE Software. 2004. V. 21, No. 2. P. 71–77.

11. *Jones S.* Toward an Acceptable Definition of Service. IEEE Software. 2005. V. 22, No. 3. P. 87–93.

12. *Fowler M.* Microservices a definition of this new architectural term. https://martinfowler.com/articles/microservices.html

13. *Ricci F., Rokach L., Shapira B., Kantor P.B.* Recommender Systems Handbook. N.Y.: Springer, 2011. 842 p.

RECOMMENDER SYSTEM OF TEXT ANALYTICS OF LEGAL DOCUMENTS

[1]D. S. Zuev, [2]M. F. Nasrutdinov, [3]A. F. Khassianov

Kazan (Volga region) Federal University

[1]dzuev11@gmail.com, [2]marat.nasrutdinov@kpfu.ru, [3]ak@it.kfu.ru

Abstract

The paper discusses the use of machine learning mechanisms, natural language analysis and intellectual search in the field of jurisprudence. The main expected results are the methodology for applying text-based analytics and semantic natural language processing (NLP) algorithms in knowledge management cases in different types of legal practice. The obtained results can be applied in the field of education and knowledge management in a wider context, since the study lies at the union of jurisprudence, mathematical and computer linguistics.

We describe a prototype of a multi-agent system of intellectual analysis of legal texts that is capable of identifying general dependencies on the existing database of legal documents, providing legal cases with similar topics, recommending the most likely outcomes of judicial review.

Keywords*: data analytics and data mining, data intensive domains, digital libraries, clustering, classification of judicial acts, recommender system, micro-service architecture.*

REFERENCES

1. Peroni S. Semantic Web Technologies and Legal Scholarly Publishing Law, Springer, Governance and Technology Series, 2014. V. 15. doi 10.1007/978-3-319-04777-5

2. *Elizarov A.M., Zhizhchenko A.B., Zhil'tsov N.G. Kirillovich A. V., Lipachev E.K.* Mathematical knowledge ontologies and recommender systems for collections of documents in physics and mathematics// Doklady Mathematics. 2016. V. 93, No 2. P. 231–233. https://doi.org/10.1134/S1064562416020174

3. Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solov'ev V.D. Methods and means for semantic structuring of electronic mathematical documents // Doklady Mathematics. 2014. V. 90. No 1. P. 521–524. doi 10.7868/S0869565214240049

4. *Grant S. Ingersoll, Thomas S. Morton, and Andrew L. Farris* Taming Text: How to Find, Organize, and Manipulate It / Manning Publications, 2012. – 320 p. ISBN: 9781933988382

5. *Zuev D.S., Marchenko A.A., Khassianov A.F.* Text Mining Tools in Legal Documents // CEUR Workshop Proceedings. 2017. V. 2022. P. 214–218. http://ceur-ws.org/Vol-2022/paper35.pdf

6. Digital Mathematics Library: a vision for the future. International Mathematical Union, 2006. http://www.mathunion.org/fileadmin/IMU/Report/ dml_vision.pdf.

7. *Olver P.J.* What's happening with the World Digital Mathematics Library? http://www.math.umn. edu/~olver/t_/wdmlb.pdf

8. Developing a 21st century global library for mathematics research. Washington, D.C.: The National Academies Press, 2014. 131 p. arxiv.org/pdf/1404.1905; http://www.nap.edu/catalog/18619/developing-a-21st-century-global-library-for-mathematics-research.

9. *Toschev A., Talanov M.* Thinking Lifecycle as an Implementation of Machine Understanding in Software Maintenance Automation Domain// Jezic G., Howlett R., Jain L. (eds) Agent and Multi-Agent Systems: Technologies and Applications. Smart Innovation, Systems and Technologies. 2015. Vol 38. Springer, Cham. doi: 10.1007/978-3-319-19728-9_25

10. *Gold N. et al.* Understanding Service Oriented Software. IEEE Software. 2004. V. 21, No 2. P. 71–77.

11. *Jones S.* Toward an Acceptable Definition of Service. IEEE Software. 2005. V. 22. No 3. P. 87–93.

12. *Fowler M.* Microservices a definition of this new architectural term. https://martinfowler.com/articles/microservices.html

13. *Ricci F., Rokach L., Shapira B., Kantor P.B.* Recommender Systems Handbook. N.Y.: Springer, 2011. 842 p.

СВЕДЕНИЯ ОБ АВТОРАХ

ЗУЕВ Денис Сергеевич – кандидат технических наук, заместитель директора по научной деятельности Высшей школы информационных технологий и интеллектуальных систем Казанского федерального университета.

Denis Sergeevich ZUEV – PhD, Deputy director for research, Higher Institute of Information Technology an Intelligent Systems

email: dzuev11@gmail.com

НАСРУТДИНОВ Марат Фаритович – кандидат физико-математических наук, заместитель директора по образовательной деятельности Высшей школы информационных технологий и интеллектуальных систем Казанского федерального университета.

Marat Faritovich NASRUTDINOV – Deputy Director for Education at Higher Institute for Information Technology and Information Systems of Kazan Federal University

marat.nasrutdinov@kpfu.ru

ХАСЬЯНОВ Айрат Фаридович – PhD, директор Высшей школы информационных технологий и интеллектуальных систем Казанского федерального университета.

Airat Faridovich *KHASSIANOV* – Phd, Director at Higher Institute for Information Technology and Information Systems of Kazan Federal University

ak@it.kfu.ru

Материал поступил в редакцию 31 июля 2018 года