МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Институт№8 «Информационные технологии и прикладная математика» **Кафедра** 810Б

Направление подготовки 02.04.02 ФИИТ **Группа** М8О-203М-18

Квалификация (степень) магистр

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

На тему: Тематический поиск в коллекции юридических документов

Автор диссертации Герасименко Николай Александрович
(Фамилия, имя, отчество) подпись

Научный руководитель Абгарян Каринэ Карленовна
(Фамилия, имя, отчество) подпись

Научный руководитель Воронцов Константин Вячеславович
(Фамилия, имя, отчество) подпись

Рецензент Артёмова Екатерина Леонидовна
(Фамилия, имя, отчество) подпись

К з а щ и т е д о п у с т и т ь

Зав. кафедрой Абгарян К. К.
(Фамилия, инициалы) подпись

« 24 » мая 2020 г.

Москва 2020 г.

РЕФЕРАТ

Магистерская диссертация содержит 30 страниц, 3 таблицы, 5 рисунков. Список использованных источников содержит 7 позиций.

ИНФОРМАЦИОННЫЙ ПОИСК, РАЗВЕДОЧНЫЙ ПОИСК, ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, АДДИТИВНАЯ РЕГУЛЯРИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ, АНАЛИЗ ЮРИДИЧЕСКИХ ДОКУМЕНТОВ

Магистерская диссертация посвящена построению системы информационного поиска в коллекции актов арбитражных судов, с использованием тематического моделирования в качестве ключевой технологии. В качестве запроса поисковой системе может выступать произвольный документ коллекции. В ответ на поисковый запрос генерируется список документов коллекции, ранжированный по убыванию релевантности.

Для решения поставленной задачи построена тематическая модель коллекции актов арбитражных судов с помощью открытой библиотеки BigARTM. При построении модели учтена специфика предметной области путем добавления в модель модальностей, особых типов токенов, таких как ссылки на нормативно-правовые акты (НПА) и юридические термины. Юридические термины выделялись автоматически, с использованием алгоритма TopMine. Построенная модель показывает высокую интерпретируемость тем и точность поиска.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Специалисты в области юриспруденции часто сталкиваются в своей работе с необходимостью поиска документов в базах юридической практики. Исследование юридических баз данных – это рутинный процесс, требующий от практикующего юриста больших временных затрат. Использование традиционных систем поиска по четкому короткому запросу (known-item search) [1] упрощает задачу специалиста, однако в формате короткого запроса зачастую невозможно описать все характеристики дела, для которого идет поиск релевантной практики. В данной работе предлагается использовать иную парадигму информационного поиска – разведочный поиск, который позволяет в качестве запроса использовать целый документ или коллекцию документов [2]. При использовании тематического моделирования в качестве ключевой технологии для построения системы разведочного поиска, поиск называют тематическим.

Вероятностное тематическое моделирование – это современный метод машинного обучения, широко применяющийся для анализа текстовых коллекций. Тематическая модель позволяет определить, к каким темам относятся документы коллекции, и какие термины образуют каждую тему. В данной работе применяется метод аддитивной регуляризации тематических моделей (АРТМ) [3]. Данный подход позволяет учитывать при построении модели дополнительные лингвистические требования и экстралингвистические данные. Кроме того, АРТМ позволяет учесть специфику предметной области путем добавления в модель модальностей, особых типов токенов, таких как ссылки на нормативно-правовые акты и юридические термины.

Все необходимые инструменты для обучения тематической модели в рамках теории аддитивной регуляризации реализованы в библиотеке с открытым кодом BigARTM [4]. В данной работе использовался Python-

интерфейс данной библиотеки, ядро которой написано на C++ и является эффективной потоковой параллельной реализацией APTM.

Целью данной работы является построение системы тематического информационного поиска по коллекции юридических документов. В ходе выполнения данной работы необходимо решить следующие задачи:

1. Определить совместно с экспертами в предметной области модальности, которые могут улучшить качество поиска.
2. Организовать экспертную разметку поисковой выдачи для репрезентативного набора запросов для последующей оценки качества поиска.
3. Обучить тематическую модель коллекции юридических документов.
4. Оценить качество поиска с помощью тематической модели и сравнить его с результатами других подходов.

С помощью BigARTM была построена тематическая модель коллекции, состоящей из 124767 судебных актов, решений Арбитражного суда Московской области. Поиск был реализован с использованием косинусной меры близости, применяемой к полученным с помощью модели тематическим векторным представлениям документов, что позволяет оценить семантическое расстояние между документами. Точность тематического поиска сравнивалась с точностью нескольких базовых моделей (baselines): модель, основанная на классическом подходе TF-IDF, а также нейросетевая модель Doc2Vec. Поскольку векторные представления для документов, полученные с помощью TF-IDF имеют слишком большую длину, равную длине словаря, к полученной матрице коллекции применялся метод сингулярного разложения (SVD).

В разделе 1.1 описана концепция разведочного поиска и основные подходы, использующиеся для решения данного класса задач. В разделе 1.2 ставится задача тематического моделирования и описываются различные виды тематических моделей, в том числе подробно говорится об их аддитивной регуляризации. Также в данной главе описывается механизм

использования в тематических моделях модальностей и регуляризаторов. В разделе 2 описан процесс обучения тематической модели и проведена оценка качества тематического поиска. Во введении подведены итоги работы и сделаны основные выводы.

ОСНОВНАЯ ЧАСТЬ

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1. Разведочный информационный поиск

1.1.1. Постановка задачи

Традиционные системы поиска по короткому запросу (known-item search) [1], такие как Google и Yandex, предназначены для поиска ответов на четко сформулированные вопросы. Использование таких систем предполагает хорошее знание необходимых терминов предметной области, а также понимание, что конкретно требуется найти. В случае отсутствия такого понимания пользователь поисковой системы двигается итерационно, просматривая результаты и уточняя на их основании свой запрос. Зачастую требуется немало итераций, прежде чем пользователь достигает нужного результата.

Пользователи традиционных систем поиска достаточно редко используют длинные специализированные запросы [5]. Пользователь старается не использовать длинных запросов, поскольку знает, что это ухудшит результаты поиска. Системы поиска по короткому запросу создавались и хорошо подходят для нахождения ответов на конкретные вопросы, например «Каков адрес ближайшей аптеки?» или «Какие существуют хорошие и недорогие средства от кашля?». Однако формат короткого запроса зачастую не позволяет подробно описать все аспекты интересующей пользователя темы. В силу чего поиск исчерпывающего ответа на вопрос может занять достаточно много времени.

Парадигма разведочного поиска позволяет пользователю более подробно описать свой запрос, получив таким образом и более полные, и более точные результаты. Процесс поиска также может иметь итерационную структуру, однако цель каждой итерации расширение познаний пользователя в выбранной области, а не только уточнение запроса. Разведочный поиск способствует более широкому и всестороннему изучению выбранной темы.

Формат длинного запроса также позволяет добавлять фрагменты найденных документов к изначальному запросу, таким образом расширяя спектр задействованных тем. Схематичное изображение итераций поиска по короткому запросу, и разведочного поиска представлены на (Рис. 1.1).

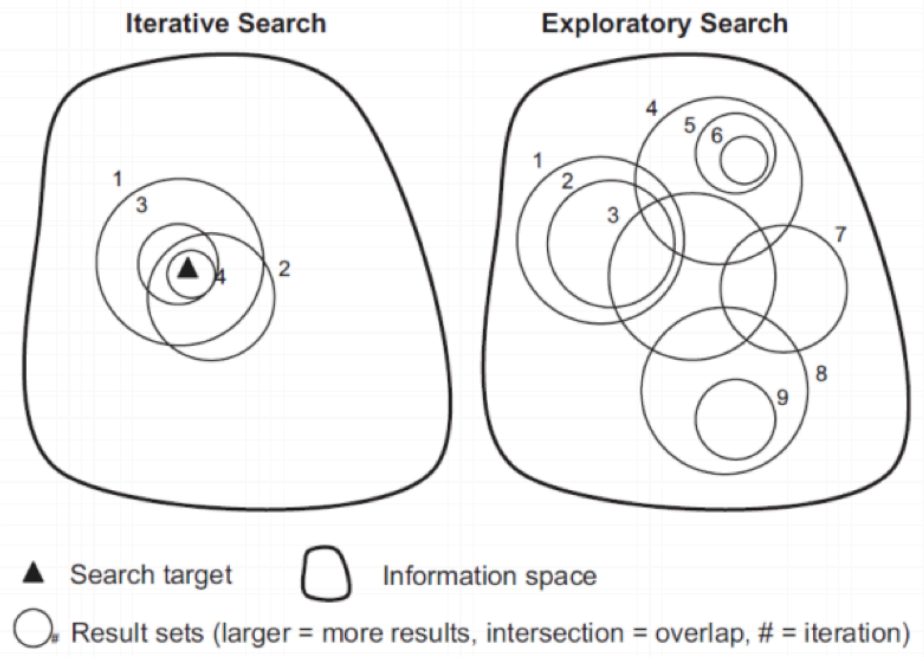

Рис. 1.1 Схематичное изображение итераций традиционного и разведочного поиска

Применительно к задаче поиска похожей судебной практики, в качестве запроса к системе разведочного поиска пользователь-юрист может использовать черновик судебного или законодательного акта, возможно составленный из фрагментов некоторого набора документов. В качестве результата специалист ожидает получить набор юридических документов, отсортированных по степени релевантности запросу. Полученный результат может быть использован для расширения и уточнения документа-запроса для последующего поиска.

Учитывая специфику задачи поиска судебной практики, как поиска по длинному запросу в большой коллекции документов, идея перехода от поиска

по короткому запросу к разведочному поиску представляется перспективной. В данной работе проверяется это предположение.

1.1.2. Тематический разведочный поиск

В случае использования тематического моделирования в качестве ключевой технологии при построении системы разведочного поиска, поиск называют тематическим. Обученная тематическая модель позволяет получить сжатые векторные представления как для документов коллекции, в которой идет поиск, так и для запроса. Такие векторные представления называются семантическими: расстояния между ними отражают смысловую разницу между документами, которым они соответствуют. Семантические векторные представления, полученные с помощью тематической модели, кроме того, обладают свойством интерпретируемости координат: каждая из них представляет собой степень принадлежности документа к определенной теме.

Алгоритм тематического поиска имеет следующую структуру.

Дано:

- коллекция документов D.
- множество запросов Q.

Алгоритм:

1. Обучить тематическую модель коллекции D.
2. Получить с помощью обученной тематической модели векторные представления документов коллекции D.
3. Получить с помощью обученной тематической модели векторные представления документов-запросов Q.
4. Пользуясь косинусной мерой близости, найти k ближайших документов коллекции D для каждого запроса из Q.

Критерии качества:

- Precision@k - доля релевантных документов среди первых k найденных.
- Recall@k - доля k первых найденных релевантных документов среди всех релевантных.

Первым шагом алгоритма является обучение тематической модели рассматриваемой коллекции документов. В результате мы получаем модель, с помощью которой можем вычислить векторные представления для документов коллекции $d \in D$ и запросов $q \in Q$. Далее для каждого запроса q мы можем найти k ближайших документов коллекции D, которые и будут поисковой выдачей системы. Близость между векторами оценивается с помощью косинусной меры, которая вычисляется по формуле:

$$cossim = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \cdot \|\vec{d}\|} = \frac{\sum_{i=1}^{n} \overrightarrow{q_i} \cdot \overrightarrow{d_i}}{\sqrt{\sum_{i=1}^{n} q_i} \sqrt{\sum_{i=1}^{n} d_i}} \quad (1.1)$$

где $\vec{q}, \vec{d}$ – векторные представления документа и запроса.

Поисковая выдача системы оценивается с помощью критериев Precision@k, являющейся мерой точности поиска, и Recall@k, являющейся мерой полноты. Данные метрики имеют следующий вид:

$$\text{Precision@}k = \frac{TP}{k} \quad (1.2)$$

$$\text{Recall@}k = \frac{TP}{P} \quad (1.3)$$

Где TP $(True\ Positive)$ – количество релевантных документов в поисковой выдаче, P $(Positive)$ – количество релевантных документов во всей коллекции.

Также может быть использована агрегированная оценка качества поиска, представляющая собой среднее гармоническое между $\text{Precision@}k$ и $\text{Recall@}k$:

$$F_1@k = \frac{\text{Precision@}k + \text{Recall@}k}{2 \cdot \text{Precision@}k \cdot \text{Recall@}k} \quad (1.4)$$

1.1.3. Оценивание качества разведочного поиска

В случае, если в коллекции может находится большое количество релевантных документов для каждого запроса, оценивание качества может быть проведено на подмножествах коллекции, содержащей релевантные и нерелевантные документы. Для каждого запроса с помощью одной из базовых моделей формируется поисковая выдача заведомо большей длины, чем значения k, которые будут использоваться при оценке качества поиска. Полученное подмножество коллекции размечается асессорами, обладающими достаточной экспертизой в предметной области. В данной работе, поскольку поиск проводился в коллекции юридических документов, в качестве асессоров выступали практикующие юристы.

Для оценки релевантности документов им была предложена трехбалльная система, которая формулировалась в понятных юристу терминах:

- документ точно окажется полезен при работе над делом
- документ может оказать полезен в некоторых частных случаях
- документ бесполезен при работе над делом

Для асессоров была сформулирована понятная задача, с которой они уже могли сталкиваться. Более опытные коллеги зачастую дают стажерам-

юристам задачу найти похожую судебную практику, набросав общее описание дела, в том числе из фрагментов других дел. Задача разметки поисковой выдачи представляет собой проверку работы стажера, которого, в данном случае, заменяет поисковая система.

Использование базовой модели позволяет не только быстрее найти релевантные документы, но и неочевидные нерелевантные. Разметка только релевантных и взятие случайных документов в качестве нерелевантных делает задачу поиска слишком простой, поэтому от этого способа было решено отказаться.

После того, как поисковые выдачи базовой модели для набора запросов размечены, можно провести оценку каждой из них по критериям $\mathrm{Precision}@k$, $\mathrm{Recall}@k$ и $F_1@k$. Результирующее качество поиска может быть получено из подходящей статистики для набора результатов метрик для всех запросов, например с помощью среднего или медианы. Имеет смысл также изучить более детально результаты для разных запросов и оценить их сложность для системы тематического поиска.

1.2. Вероятностное тематическое моделирование

1.2.1. Постановка задачи

Пусть D – множество документов (коллекция), W – множество всех употребляемых в коллекции токенов (словарь коллекции). Каждый документ коллекции $d \in D$ описывается последовательностью токенов $(w_1, \dots, w_{n_d}) \in W$, причем каждому токену w ставится в соответствие число n_{dw} его вхождений в документ d.

Таким образом, коллекция может быть представлена в виде матрицы частотных оценок вероятности встретить токен w_i в документе d_j:

$$F = (f_{wd})_{W \times D} \tag{1.5}$$

$$f_{wd} = \frac{n_{dw}}{n_d} \tag{1.6}$$

Тематическая модель строится в предположении, что существует множество тем T, затронутых в документах коллекции. Тогда коллекция представляет собой множество троек $(w_i, d_i, t_i), i = 1 \dots n$, порожденных дискретным распределением $p(w, d, t)$, определенном на конечном вероятностном пространстве $W \times D \times T$. Причем токены и документы коллекции являются наблюдаемыми переменными, в то время как темы – скрытыми.

При построении тематической модели принимается гипотеза «мешка слов», утверждающая, что порядок слов в документе не важен. Также принимается гипотеза «мешка документов», утверждающая, что не важен порядок документов в коллекции. Для того, чтобы не терять информацию о взаимном расположении слов, могут быть использованы, наряду с обычными словами, устойчивые словосочетания и коллокации, выделенные из коллекции с помощью различных алгоритмов.

В тематическом моделировании темы представляются дискретными распределениями на множестве токенов, а документы – дискретными распределениями на множестве тем. Пусть $p(w|t)$ – вероятность, с которой токен w встречается в теме t, а $p(t|d)$ – вероятность, с которой тема t встречается в документе d. Вычисление таких вероятностей для всех $t \in T, w \in W, d \in D$ равносильно вычислению матриц:

$$\Phi = p(w|t)_{W \times T} \tag{1.7}$$

$$\Theta = p(t|d)_{T \times D} \tag{1.8}$$

Такие матрицы называются матрицами терминов-тем и тем-документов соответственно. Таким образом, задача тематического моделирования

сводится к оценке параметров $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$ по коллекции D. Это задача стохастического матричного разложения матрицы F терминов-документов на произведение матриц Φ терминов-тем и матрицы Θ тем-документов (Рис. 1.2):

$$F_{W\times D} \approx \Phi_{W\times T} \times \Theta_{T\times D} \tag{1.9}$$

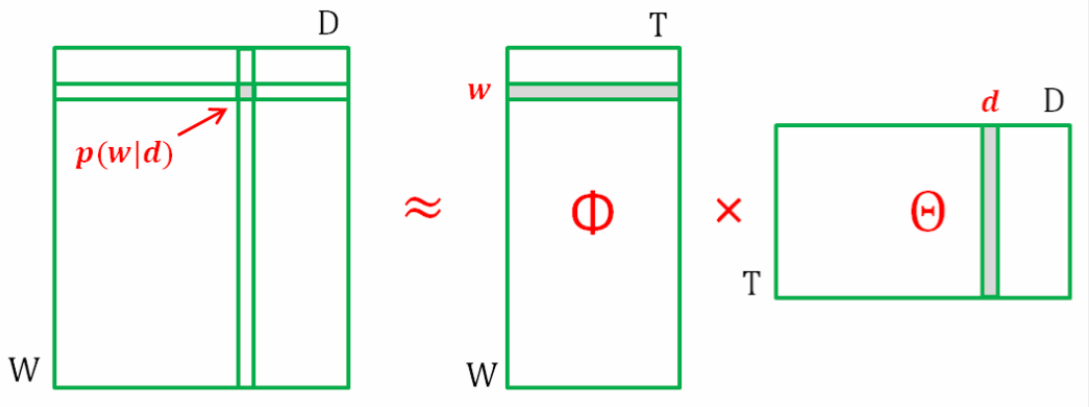

Рис. 1.2 Стохастическое матричное разложение матрицы терминов-документов

1.2.2. Аддитивная регуляризация тематических моделей

Одним из традиционных подходов в тематическом моделировании является вероятностный латентный семантический анализ (PLSA) [6]. В рамках этого подхода задача оценки параметров $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$ решается путем максимизации логарифма правдоподобия, с условием нормировки столбцов матрицы Φ и строк матрицы Θ и неотрицательности всех элементов этих матриц:

$$\mathrm{L}(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\mathrm{L}(\Phi,\Theta) = \ln \prod_{d\in D} \prod_{w\in d} p(w|d)^{n_{dw}} = \sum_{d\in D} \sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \tag{1.10}$$

$$\sum_{w\in W} \phi_{wt} = 1, \qquad \phi_{wt} \geq 0, \qquad \sum_{t\in T} \theta_{td} = 1, \qquad \theta_{td} \geq 0.$$

Методом решения данной задачи является итерационный ЕМ-алгоритм, который представляет собой метод простых итераций для решения системы нелинейных уравнений, возникающих из условий Каруша-Куна-Таккера для оптимизационной задачи.

Проблема данного подхода заключается в том, что в таком общем виде, задача тематического моделирования имеет бесконечно много решений, поскольку, если $F = \Phi\Theta$ является решением задачи, то для любых невырожденных матриц S таких, что $\Phi' = \Phi S$ и $\Theta' = \Theta S$ – стохастические матрицы, $F = (\Phi S)(S^{-1}\Theta)$ будет являться решением задачи. Задачи такого типа называются некорректно поставленными по Адамару.

Общепринятым способом работы с такими задачами является дополнение условий задачи дополнительными требованиями. Такой подход называется регуляризацией, а требования – регуляризаторами. В теории аддитивной регуляризации (АРТМ) [3] регуляризаторы R_i добавляются в качестве слагаемых к логарифму правдоподобия (2.6):

$$\mathrm{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta} \qquad (1.11)$$

$$\sum_{w\in W} \phi_{wt} = 1, \qquad \phi_{wt} \geq 0, \qquad \sum_{t\in T} \theta_{td} = 1, \qquad \theta_{td} \geq 0,$$

где $\tau_i \geq 0$ – коэффициент регуляризации.

Полученная задача также решается с помощью ЕМ-алгоритма, с добавлением соответствующего слагаемого на М-шаге. Первое уравнение системы представляет собой Е-шаг (expectation), на котором вычисляются условные вероятности для тем для каждой пары термин-документ, на основании значений, которые были вычислены на предыдущем М-шаге.

Второе и третье уравнения – это М-шаг (maximization), на котором вычисляются новые значения параметров ϕ_{wt} и θ_{td}.

$$\begin{cases} p_{tdw} = p(t|d,w) = \underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\dfrac{\partial R(\Phi,\Theta)}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{w\in W}{\text{norm}}\left(\sum_{w\in W} n_{dw}p_{tdw} + \theta_{td}\dfrac{\partial R(\Phi,\Theta)}{\partial \theta_{td}}\right), \end{cases} \quad (1.12)$$

где $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_t,x_s\}}$ – операция нормирования вектора.

В качестве начального приближения ϕ_{wt} и θ_{td} могут быть взяты равномерные ортонормированные вектора или случайные вектора, удовлетворяющие условиям, накладываемым на ϕ_{wt} и θ_{td}.

1.2.2.1. Регуляризатор разреживания

Регуляризатор разреживания матрицы тем-документов Θ формализует так называемую гипотезу разреженности, состоящую в том, что каждый документ относится к малому количеству тем. В практических задачах разумно использовать сильно разреженные матрицы Φ и Θ, в которых около 90% значений являются нулями.

Разреженность распределения обратно пропорционально его энтропии, а равномерное распределение имеет максимальную энтропию. Поэтому требование разреженности эквивалентно максимизации KL-дивергенции между распределениями θ_{td} и равномерным распределением α_t. Регуляризатор, таким образом, представляет из себя суммарную KL-дивергенцию по всем темам $t \in T$ и документам $d \in D$.

$$\sum_{d\in D} \mathrm{KL}(\alpha_t||\theta_{td}) \to \max_{\Theta}$$

$$\mathrm{R}(\Theta) = -\alpha \sum_{d\in D}\sum_{t\in T} \alpha_t \ln\theta_{td} \to \max_{\Theta}, \quad (1.13)$$

где α – коэффициент регуляризации.

1.2.2.2. Регуляризатор декоррелирования

Регуляризатор декоррелирования формализует предположение о различности тем, как распределений на множестве токенов, максимизируя ковариации между темами – столбцами матрицы ϕ:

$$\mathrm{R}(\Phi) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W}\phi_{wt}\phi_{ws} \to \max_{\Phi}, \tag{1.14}$$

где τ – коэффициент регуляризации.

1.2.3. Мультимодальное тематическое моделирование

Мультимодальные тематические модели позволяют использовать, помимо обычного теста, также дополнительную информацию о документе (метаинформацию), такую как: авторы, теги, даты, ссылки и т.д. Базовой модальностью являются просто слова текста, другие модальности учитываются аналогично, у каждой модальности m имеется свой словарь токенов W_m.

Токены разных модальностей учитываются в модели аналогично унимодальному случаю. Каждый токен $w \in W_m$ появляется в теме $t \in T$ с вероятностью $p(w|t)$. Для каждой модальности формируется матрица терминов-тем:

$$\Phi_m = (p(w|t))_{W_m\times T} \quad \forall m \in M \tag{1.15}$$

Общая матрица Ф формируется объединением матриц Φ_m для всех модальностей, записанных в виде столбца:

$$\Phi = p(w|t)_{W\times T} \tag{1.16}$$

Матрица тем-документов Θ остается неизменной по сравнению с унимодальным случаем.

Логарифм правдоподобия обобщается на мультимодальный случай следующим образом:

$$\mathrm{L}(\Phi,\Theta)=\sum_{m\in M}\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+\sum_{i}\tau_i R_i(\Phi,\Theta)\to\max_{\Phi,\Theta} \quad (1.17)$$

Причем для каждой модальности регуляризаторы, связанные с Ф, определяются независимо, со своими коэффициентами регуляризации.

1.2.3.1. Модальность юридических терминов

В ходе обсуждения задачи со специалистами в предметной области были выявлены важные модальности, использование которых может существенно улучшить качество модели. Одна из этих модальностей – модальность юридических терминов.

Юридические термины как и вообще термины предметной области могут быть извлечены из текста двумя способами: с помощью словаря этих терминов и с помощью автоматического извлечения. В данной работе был выбран второй вариант.

Термины извлекались как коллокации, n-граммы, встречающиеся в тексте гораздо чаще, чем можно было бы ожидать при случайном соединении. Для извлечения коллокаций использовался метод TopMine [7].

1.2.3.2. Модальность ссылок на нормативно-правовые акты

Другой важной модальностью юридического текста являются ссылки на нормативно-правовые акты (НПА). Информация о том, какие законодательные акты упоминаются в тексте документа, достаточно однозначно может указывать на затронутые в тексте темы. Например, если речь идет об административном правонарушении, в тексте будут указаны статьи Кодекса об административных правонарушениях (КоАП РФ), а если речь идет об уголовном деле – Уголовный Кодекс (УК РФ).

Была написана программа на Python для извлечения ссылок на НПА из документов. В программе используются регулярные выражения и рекурсивные функции для извлечения сложных ссылок.

1.2.4. Оценивание качества тематических моделей

При использовании тематической модели для построения системы разведочного поиска оценка качества производится по поисковой выдаче. Такие критерии качества называются внешними, поскольку не учитывают архитектуру тематической модели, и могут быть использованы одинаково для оценки любой другой модели, использующейся для векторизации документов коллекции.

Помимо внешних критериев качества существуют внутренние критерии, которые используются в процессе обучения тематической модели для подбора коэффициентов регуляризации.

В задачах классификации и регрессии для оценки качества модели используется понятие «ошибки» или «потери». Для тематических моделей эти понятия не могут быть четко определены. Одной из принятых метрик качества является перплексия, которая используется в компьютерной лингвистике для оценки языковых моделей. Эта метрика тесно связана с правдоподобием и

представляет собой степень несоответствия модели токенам w, наблюдаемым в документах коллекции:

$$\mathcal{P}(D,p) = \exp\left(-\frac{1}{n}\mathrm{L}(\Phi,\Theta)\right) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right) \qquad (1.18)$$

Модель тем лучше, чем меньше ее перплексия. Перплексия может быть интерпретирована следующим образом. Если термины w порождены равномерным распределением $p(w) = 1/V$ на словаре длины V, то перплексия модели $p(w)$ сходится к V с ростом его длины. Чем сильнее отличается от равномерного распределения распределение $p(w)$, тем меньше будет перплексия модели.

Два других важных внутренних критерия качества тематической модели – это разреженность матриц терминов-тем Φ и тем-документов Θ. Эти метрики представляют собой доли нулей в соответствующих матрицах и позволяют оценить, насколько выполняется гипотеза разреженности распределений тем в документах, а также терминов в темах, предполагающую, что в одной теме с высокой вероятностью появляется небольшое количество терминов.

С помощью внутренних критериев качества может быть реализован жадный алгоритм подбора коэффициентов регуляризации, заключающийся в последовательном включении в модель различных регуляризаторов и выбора, таким образом, наилучшего коэффициента.

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1. Предварительная обработка данных

Эксперименты проводились на коллекции, состоящей из 124767 судебных актов, решений Арбитражного суда Московской области. Из каждого судебного решения была извлечена установочная часть, поскольку остальные части документов содержат только техническую судебную информацию, либо дублируют информацию из установочной части. Документы, длина установочной части которых оказалась меньше 4000 символов, были удалены из коллекции.

Из установочных частей документов были извлечены токены двух дополнительных модальностей: ссылки на нормативно-правовые акты и юридические термины. Юридические термины и обычные слова были приведены к начальной форме с помощью библиотеки Pymorphy2, затем были удалены общие шумовые слова русского языка, а также шумовые слова, специфичные для задачи, такие как «установить», «решить», «суд» и т.д. Из созданных таким образом «мешков слов» трех модальностей для каждого текста был составлен файл в формате Vowpal Wabbit, требующемся для работы с библиотекой BigARTM.

При инициализации словаря тематической модели с помощью BigARTM были удалены токены, встречающиеся менее, чем в 5 документах коллекции, а также токены, встретившиеся более, чем в 85% документов.

2.2. Обучение тематической модели

При обучении модели использовалось два регуляризатора: регуляризатор декоррелирования распределений терминов в темах и регуляризатор разреживания распределений тем в документах. Подбор коэффициентов

регуляризации осуществлялся по алгоритму, аналогичному использованному в работе [2].

На первом этапе производился подбор коэффициента для регуляризатора декоррелирования. Для каждого из тестируемого набора значений коэффициента τ проводилось по 8 итераций EM-алгоритма, после чего выбиралось наилучшее значение по критериям перплексии, разреженности матрицы Φ и разреженности матрицы Θ. Затем в выбранную таким образом наилучшую модель добавлялся регуляризатор разреживания Θ и проводилось еще 8 итераций EM-алгоритма для каждого из тестируемого набора значений коэффициента разреживания α. Для модели с полученной таким образом комбинацией коэффициентов τ и α проводилось еще 3 итерации EM-алгоритма. Зависимости внутренних критериев качества тематических моделей от количества итераций EM-алгоритма для итоговой модели на (Рис. 2.1) и (Рис. 2.2).

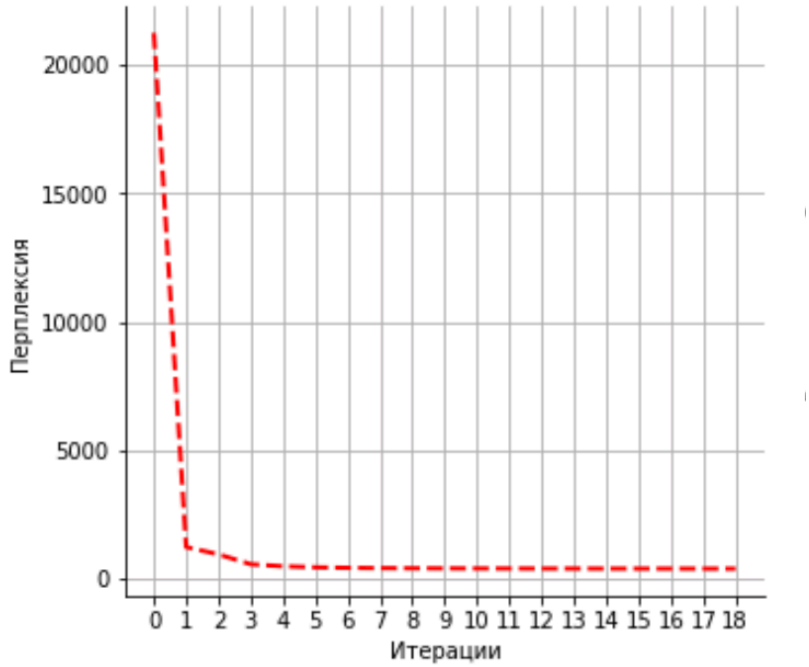

Рис. 2.1 Зависимость перплексии от количества итераций

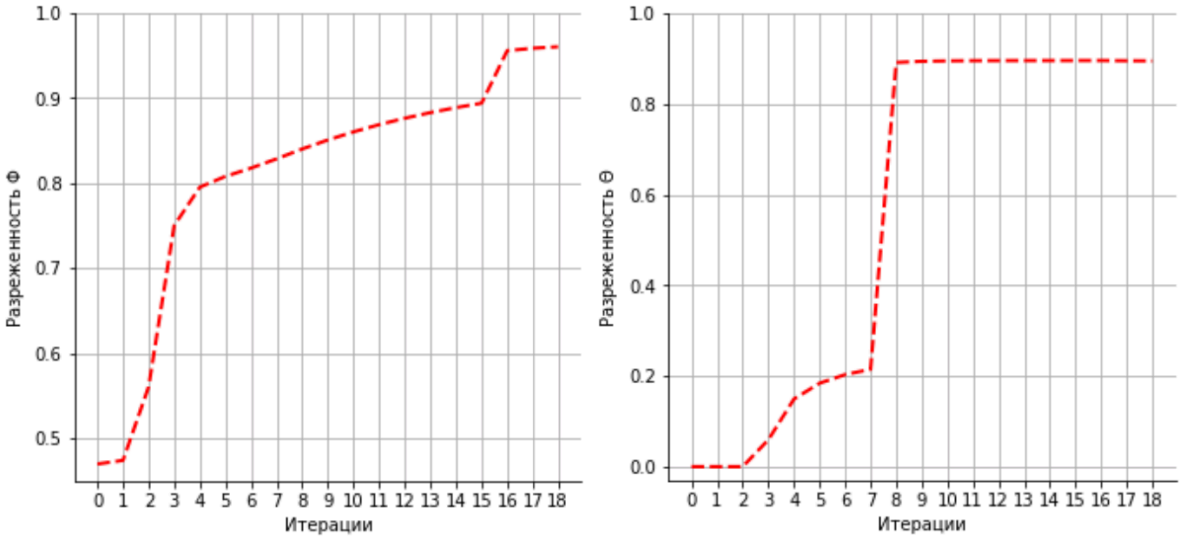

Рис. 2.2 Зависимость разреженности матриц Φ и Θ от количества итераций

2.3. Оценка качества разведочного поиска

Для оценки качества поиска использовались размеченные поисковые выдачи для набора запросов. Всего было размечено 72 поисковых выдачи, сделанных с помощью базовой модели. В результате отбора по необходимому количеству релевантных и нерелевантных документов валидационная выборка была сформирована из 30 поисковых выдач.

Для сравнения на всей коллекции были обучены базовые модели (baselines): традиционная модель TF-IDF и нейросетевая модель Doc2Vec.

Поскольку размерность векторов, построенных с помощью TF-IDF, которая равна длине словаря коллекции, слишком велика для адекватной работы поисковой системы, использован метод сингулярного разложения матриц (SVD) для понижения размерности до 100. Использовалась реализация TF-IDF библиотеки Sklearn.

Модель Doc2Vec была обучена для двух вариантов: размерности 100 и 200. Использовалась реализация Doc2Vec библиотеки Gensim. Результаты

оценки точности и полноты поиска по метрикам Precision@k и Recall@k для базовых моделей представлены в (Таблице 2.1).

Таблица 2.1. Качество поиска для различных моделей

	APTM	PLSA	TF-IDF	Doc2Vec	
Метрика	100	100	100	100	200
Precision@5	0.79	0.71	0.76	0.75	0.75
Precision@10	0.87	0.8	0.83	0.84	0.83
Precision@15	0.93	0.87	0.89	0.91	0.91
Precision@20	0.94	0.91	0.9	0.93	0.91
Recall@5	0.11	0.09	0.08	0.1	0.09
Recall@10	0.16	0.14	0.16	0.16	0.16
Recall@15	0.18	0.16	0.17	0.16	0.17
Recall@20	0.23	0.2	0.21	0.21	0.22

В результате использования жадного алгоритма подбора гиперпараметров тематической модели была выбрана модель со следующими значениями коэффициентов регуляризации:

- для декоррелирования распределений терминов в темах $\tau = 10^6$
- для разреживания распределений тем в документах $\alpha = -0.5$.

Количество тем было выбрано равным 100, веса модальностей 1, 2, 10 для слов, коллокаций (юридических терминов) и ссылок на нормативно-правовые акты соответственно.

Результаты для моделей с разным количеством тем представлены в (Таблице 2.2). Для каждой модели был проведен подбор гиперпараметров в соответствии с алгоритмом, описанным в разделе 2.2, поскольку коэффициенты регуляризация не инвариантны относительно количества тем.

Эксперименты по подбору количества тем модели показали, что рост качества поиска останавливается при количестве тем, равном 100. Дальнейшее увеличение количества тем не приводит к увеличению точности и полноты,

приводя при этом к увеличению времени обучения моделей и векторизации документов с их помощью.

Таблица 2.2 Качество поиска для моделей с разным количеством тем

	Количество тем модели				
Метрика	25	50	100	200	300
Precision@5	0.7	0.71	0.79	0.75	0.75
Precision@10	0.79	0.8	0.87	0.84	0.83
Precision@15	0.85	0.87	0.93	0.91	0.91
Precision@20	0.9	0.91	0.94	0.93	0.91
Recall@5	0.09	0.09	0.11	0.1	0.09
Recall@10	0.13	0.14	0.16	0.16	0.16
Recall@15	0.16	0.16	0.18	0.16	0.17
Recall@20	0.2	0.2	0.23	0.21	0.22

В (Таблице 2.3) представлены результаты для тематических моделей с разными комбинациями модальностей (Слова, Ссылки на НПА, Юридические Термины).

Таблица 3.3 Качество поиска для моделей с разными комбинациями модальностей

	Комбинация модальностей			
Метрика	C	CH	CT	TH
Precision@5	0.73	0.78	0.74	0.78
Precision@10	0.84	0.86	0.84	0.85
Precision@15	0.9	0.89	0.9	0.91
Precision@20	0.9	0.91	0.91	0.92
Recall@5	0.09	0.1	0.09	0.1
Recall@10	0.13	0.13	0.12	0.14
Recall@15	0.15	0.16	0.15	0.17
Recall@20	0.2	0.21	0.2	0.2

По результатам видно, что модальность юридических терминов и модальность слов не дают существенного прироста при совместном использовании, а модальность ссылок на нормативно-правовые акты дает прирост в точности поиска 4-5%.

ЗАКЛЮЧЕНИЕ

Системы разведочного поиска позволяют использовать при поиске длинные запросы, позволяющие подробно описать обстоятельства дела, для которого практикующий юрист ищет похожую практику. Такой подход упрощает для специалиста процесс поиска интересующих его документов.

В данной работе реализована система разведочного поиска в коллекции актов арбитражных судов с использованием тематического моделирования с аддитивной регуляризацией в качестве ключевой технологии.

С помощью экспертов в предметной области были выявлены важные модальности, способные улучшить качество поиска: нормативно-правовые акты и юридические термины.

С использованием «жадного» алгоритма подбора параметров регуляризации была обучена мультимодальная тематическая модель с высоким уровнем разреженности матриц Φ и Θ.

Поиск был реализован с применением косинусной меры близости к тематическим векторам документов коллекции. Для оценки качества поиска была организована разметка экспертами в области юриспруденции поисковой выдачи для набора документов-запросов.

Система показала высокие результаты точности поиска, заметно превышающие результаты baseline-моделей, а темы модели были оценены практикующими юристами как хорошо интерпретируемые.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] White, Ryen W. Exploratory Search: Beyond the Query-Response Paradigm / Ryen W. White, Resa A. Roth. Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan and Claypool Publishers, 2009.

[2] Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. Multi-objective topic modeling for exploratory search in tech news. In Communications in Computer and Information Science, pages 181–193. Springer International Publishing, nov 2017.

[3] K.Vorontsov, O.Frei, M.Apishev A.Yanina P.Romov M.Dudarenko. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections / M.Apishev A.Yanina P.Romov M.Dudarenko K.Vorontsov, O.Frei. — 2014.

[4] Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // AIST'2015, Analysis of Images, Social Networks and Texts. Springer International Publishing Switzerland, 2015. Communications in Computer and Information Science (CCIS), pp. 370–384.

[5] Shah, C., C. Hendahewa, and R. Gonzalez-Ibanez. 2016. Rain or shine? forecasting search process performance in exploratory search tasks. Journal of the Association for Information Science and Technology 67(7):1607–1623.

[6] T.Hoffman. Probabilistic Latent Semantic Analysis / T.Hoffman // Uncertainity in Artificial Intelligence. — 1999. http://cs.brown.edu/~th/papers/ Hofmann-UAI99.pdf.

[7] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment, 8(3):305–316, 2014.

[8] Герасименко Н.А., Нагибина Д.А., Воронцов К.В. Тематический поиск в коллекции юридических документов. Сборник трудов IV Международной научно-технологической конференции студентов и молодых ученых «Молодежь. Инновации. Технологии» Новосибирск, 28 – 30 апреля 2020 года