Тематический поиск схожих дел в коллекции актов арбитражных судов*

Герасименко Н. А.[1]***, Артёмова Е. Л.***[2]***, Воронцов К. В.***[3]
nikgerasimenko@gmail.com
[1]МАИ; [2]НИУ ВШЭ; [3]МФТИ

В работе рассматривается задача информационного поиска по 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 и юридические термины. Юридические термины выделялись полуавтоматически, с использованием алгоритма TopMine [1]. С помощью данного алгоритма выделены коллокации, из множества которых экспертами удалены словосочетания, не относящиеся к предметной области или являющиеся шумовыми.

Ключевые слова: *LegalTech, BigARTM, тематическое моделирование, аддитивная регуляризация мультимодальных иерархических тематических моделей.*

1 Введение

Специалисты в области юриспруденции сталкиваются в своей работе с необходимостью поиска документов в базах судебной практики. Они ищут дела, схожие с теми, над которыми работают, не всегда при этом зная, по какому запросу или в какой именно тематике искать. Данная задача отнесится к классу задач информационного поиска.

Тематический разведочный поиск - это относительно новая парадигма информационного поиска, в рамках которой в качестве запроса, в отличие от привычных систем поиска по ключевым словам, выступает целый документ или коллекция документов.

Юридический язык или «язык права», по выражению Ю.В. Кудрявцева, «носит на себе отпечаток естественного языка», но при этом является «специфичным по построению» [5] Таким образом, юридические тексты обладают особой спецификой и понимание смысла юридического текста зачастую требует консультации с экспертом. В ходе совместной работы с экспертами в области юриспруденции авторами был выявлен ряд важных модальностей, без учета которых, по мнению экспертов, невозможно построение адекватной модели. Библиотека BigARTM [3], с помощью которой строится тематическая модель, обладает возможностями для учета этих модальностей. В модели учтена модальность ссылок на нормативно-правовые акты и модальность юридических терминов.

Данное исследование основано на подходе, описанном в работе [2], где он использован при построении тематической модели для разведочного поиска по статьям на порталах Habrahabr.ru и TechCrunch.com. Построенная авторами модель показала хорошие результаты на размеченных с помощью ассесоров данных.

2 Постановка задачи

Дана коллекция, состоящая из 7937 юридических документов, а именно актов арбитражных судов относящихся к делам о банкротстве. Для каждого документа известно его

*Задачу поставил: Воронцов К. В. Консультант: Артёмова Е. Л.

место в иерархии категорий, например, категория «Особенности банкротства отдельных категорий должников» или подкатегория «Внешнее урправление» категории «Процедры банкротства».

Задача состоит в построении тематической модели данной коллекции с помощью библиотеки BigARTM. Перед построением тематической модели необходимо провести предварительную обработку текстов, которая заключается в следующих шагах.

1. Проанализировав структуру документов, требуется выделить из них, при помощи контекстно-свободных грамматик и регулярных выражений, ссылки на нормативно-правовые акты (НПА).
2. Требуется выделить юридические темины с помощью алгоритма TopMine [1].

Ссылки на НПА и юридические термины будут учтены в модели в качестве модальностей.

Предварительная оценка модели будет проводиться по критериям перплексии, разреженности распределений тем в документах, а также разреженности распределений токенов в темах для модальностей ссылок на НПА и юридических терминов в соответствии со стандартной методологией оценки [4].

Внешняя оценка строится следующим образом. Полученные в результате моделирования тематические вектора для каждого документа кластеризуются при помощи алгоритма k-means. Затем, по критерию Rand Index делается оценка, в какой степени картина кластеризации согласованна с картиной принадлежности документов их категориям: оказались ли в одном кластере документы из одной категории, и оказались ли документы из разных категорий в разных кластерах.

Формальная постановка задачи. Дана коллекция документов D, $|D| = n$. Каждый документ $d \in D$ принадлежит одной из категорий $T_i \in T$. С другой стороны, каждый документ d принадлежит одному из кластеров $Y_j \in Y$, полученных в результате применения алгоритма k-means к множеству тематических векторов документов коллекции. Требуется построить тематическую модель таким образом, что

$$\frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \to \max, \tag{1}$$

где
$a = |S^a|$, где $S^a = \{(d_i, d_j) \mid d_i, d_j \in Y_k, d_i, d_j \in T_l\}$,
$b = |S^b|$, где $S^b = \{(d_i, d_j) \mid d_i \in Y_{k_1}, d_j \in Y_{k_2}, d_i \in T_{l_1}, d_j \in T_{l_2}\}$,
$c = |S^c|$, где $S^c = \{(d_i, d_j) \mid d_i, d_j \in Y_k, d_i \in T_{l_1}, d_j \in T_{l_2}\}$,
$d = |S^d|$, где $S^d = \{(d_i, d_j) \mid d_i \in Y_{k_1}, d_j \in Y_{k_2}, d_i, d_j \in T_l\}$,
$1 \leqslant i, j \leqslant n, i \neq j, 1 \leqslant k, k_1, k_2 \leqslant r, k_1 \neq k_2, 1 \leqslant l, l_1, l_2 \leqslant s, l_1 \neq l_2$.

3 Вероятностное тематическое моделирование на основе аддитивной регуляризации

Пусть M - множество модальностей, каждой из которой соответствует набор токенов W_m, называемый словарем модальности. Пусть W - множество токенов из всех словарей, соответствующих модальностям из M. Каждый документ коллекции D с длинной n_d представляет собой набор токенов $w_1, ..., w_{n_d}$ из множества W.

В соответствии с теорией аддитивной регуляризации тематических моделей [4] для каждой модальности вводится критерий логарифма правдоподобия и с помощью EM-алгоритма максимизируется их взвешенная сумма. Также к сумме добавляются регуляризаторы - дополнительные критерии, необходимые поскольку в общем случае задача имеет бесконечно много решений.

Регуляризаторы сглаживания и разреживания имеют одинаковый вид и отличаются только знаками коэффициентов α и β, для регуляризатора разреживания они отрицательны.

$$R(\Phi,\Theta) = \beta \sum_{m\in M}\sum_{t\in T}\sum_{w\in W^m} \beta_w ln\varphi_{wt} + \alpha \sum_{d\in D}\sum_{t\in T} \alpha_t ln\theta_{td} \to \max. \quad (2)$$

Регулятор сглаживания вводит в модель требование схожести распределений φ_{wt} с распределением β_w и θ_{td} с распределением α_t. Регуляризатор разреживания, в свою очередь, способствует появлению нулевых элементов в распределенях φ_{wt} и θ_{td}, что позволяет находить более компактные представления документов.

Регуляризатор декоррелирования вводит в модель требование различности тем путем минимизации ковариации между столбцами матрицы Ф.

$$R(\Phi) = -\tau \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \varphi_{wt}\varphi_{ws} \to \max. \quad (3)$$

Также побочным эффектом работы регуляризатора декоррелирования является разреживание матрицы Ф, поэтому в случае его применения, можно не применять регуляризатор разреживания для нее.

4 Вычислительный эксперимент

Эксперимент проводился на коллекции, состоящей из 7937 юридических документов, а именно актов арбитражных судов относящихся к делам о банкротстве. О каждом документе была известна его категория, эта информация в последствии использовалась для внешней оценки качества модели. При построении модели использовались, помимо модальности текста, модальность ссылок на НПА и модальность юридических терминов.

В рамках предобработки документов была проведена лемматизация при помощи морфологического анализатора pymorphy2, были исключены 5% наиболее высокочастотных слов, а также слова общей лексики из списка stop-words библиотеки nltk.

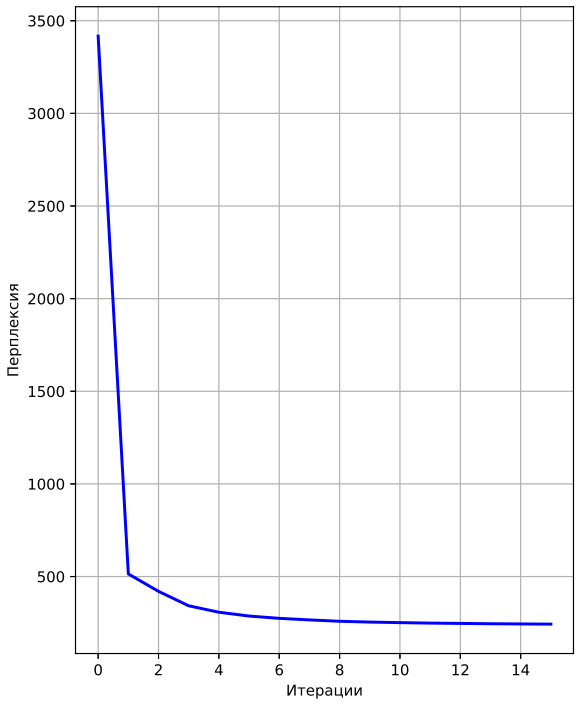
При построении тематической модели использовались регуляризаторы декоррелирования для матрицы Ф терминов в темах и регуляризатор разреживания для матрицы Θ тем в документах. Выбор параметров модели производился путем перебора значений по сетке с использованием набора критериев качества: перплексия, разреженность распределений токенов в темах, разреженность распределений тем в документах, а также внешнего критерия качества. Подбор весов модальностей также производился с помощью перебора по сетке.

Для каждого значения параметра регуляризации производилось по 8 итераций EM-алгоритма, после чего выбиралось то значение, при котором модель улучшилась по одному или нескольким критериям качества, существенно не ухудшившись ни по одному из них.

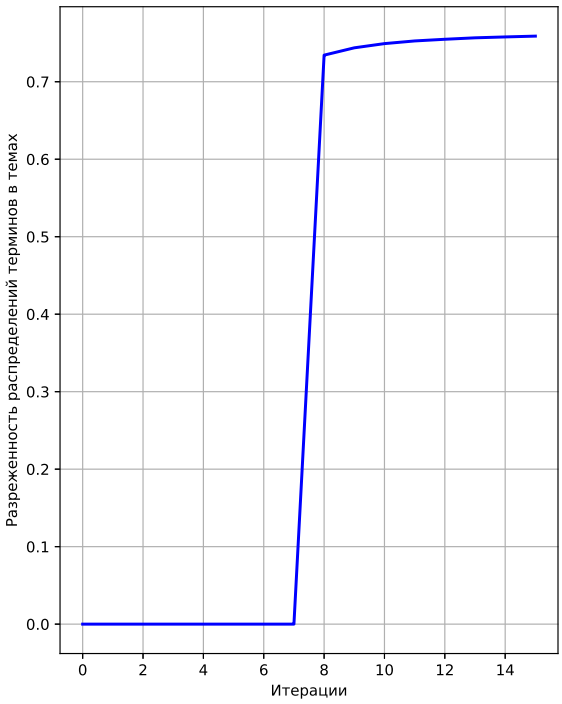
Регуляризатор декоррелирования матрицы Ф добавлялся в модель изначально, а спустя 8 итераций EM-алгоритма добавлялся регуляризатор разреживания для матрицы Θ.

На рисунке 1 изображены графики зависимости перплексии и разреженности распределений терминов в темах модели при найденных оптимальных значениях параметров регуляризации: $\tau = 3100$, $\alpha = -1.5$, $\beta = 0.25$.

В ходе сравнения показателей внешнего критерия качества - согласованности кластеризации тематических векторов и принадлежности документов категориям - модели APTM и модели tf-idf ссылок на НПА получены результаты, отраженные в таблице 1.



(*а*) Перплексия (термины)



(*б*) Разреженность Θ (термины)

Рис. 1 Зависимости перплексии и разреженности распределений терминов в темах.

Таблица 1 Сравнение показателей внешнего критерия качества.

Модель	ARI	AMI
TF-IDF по словам	12%	12.5%
TF-IDF по ссылкам на НПА	17%	22%
APTM	37.5%	42%

5 Заключение

Целью разведочного поиска является облегчение для человека задачи поиска необходимой информации, и поиск для специальстов в области юриспруденции является конкретной реализацией этой идеи. В работе рассмотрена задача информационного поиска по коллекции юридических документов, а именно актов арбитражных судов.

При помощи библиотеки BigARTM построена тематическая модель, строющая для документов коллекции сжатые векторные представления и, таким образом, позволяющая реализовать поиск, используя косинусную меру близости.

Был реализован метод внешней оценки качества тематической модели, заключающийся в вычислении согласованности между картиной кластеризации тематических векторов документов коллекции и их принадлежности различным категориям дел.

Результаты показывают, что тематическая модель способна корректно строить векторные представления юридических документов, что, в перспективе, дает возможность построения системы разведочного поиска, с использованием тематического моделирования в качестве ключевой технологии.

Литература

[1] 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.

[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] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

[4] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323, 2015.

[5] Kudriavtsev Yu.V. *Нормы права как социальная информация.* Юрид. лит-ра, 1981.