

Тематический поиск схожих дел в коллекции актов арбитражных судов*

Герасименко Н. А.¹, Артёмова Е. Л.², Воронцов К. В.³

nikgerasimenko@gmail.com

¹МАИ; ²НИУ ВШЭ; ³МФТИ

В работе рассматривается задача информационного поиска по коллекции актов арбитражных судов. В качестве запроса поисковой системе может выступать любой 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: добавлены модальность ссылок на нормативно-правовые акты, а также модальность юридических терминов, выделенных полуавтоматически, с использованием алгоритма TopMine.

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

1 Введение

Специалисты в области юриспруденции регулярно сталкиваются в своей работе с необходимостью поиска документов в базах судебной практики. Зачастую 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 может быть отнесена к классу задач информационного поиска.

Существует несколько парадигм информационного поиска, одной из которых является тематический разведочный поиск, в рамках которого в качестве запроса может выступать целый документ или даже коллекция документов.

Юридические тексты обладают особой спецификой и, не смотря на то, что могут иногда казаться обычными текстами на естественном языке, не могут в полной мере рассматриваться таким образом. Понимание истинного смысла юридического текста зачастую требует консультации с экспертом.

В ходе совместной работы с экспертами в области юриспруденции был выявлен ряд важных модальностей, без учета которых невозможно построение адекватной модели. Библиотека BigARTM [3], с помощью которой строится тематическая модель, обладает возможностями для учета этих модальностей. В данной работе 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 Nabrаhаbr.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 место в иерархии категорий, например, категория «Особенности банкротства отдельных

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

категорий должников» или подкатегория «Внешнее управление» категории «Процедуры 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 термины 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].

Окончательная оценка будет строиться 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.

Таким образом, формальная постановка задачи может быть сформулирована в следующем виде. Пусть D - коллекция документов, $|D| = n$. 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

$$\frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}} \rightarrow \max, \quad (1)$$

где:

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

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 длиной n_d представляет собой набор токенов w_1, \dots, w_{n_d} из множества W .

В соответствии с теорией аддитивной регуляризации тематических моделей для каждой модальности вводится критерий логарифма правдоподобия и с помощью ЕМ-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.

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

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 распределениях 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 Φ .

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

Также побочным эффектом работы регуляризатора декоррелирования является разреживание матрицы Φ , 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 rymorphu2, были исключены 5% наиболее высокочастотных слов, а также слова общей лексики из списка stop-words библиотеки nltk.

При построении тематической модели использовались регуляризаторы декоррелирования и сглаживания для матрицы Φ 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. Подбор весов модальностей также производился с помощью перебора по сетке.

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.

