Исследование результатов работы алгоритма LDA на отдельных классах патентов

А. Г. Кравец[1], Н. О. Шумейко[2], С. С. Васильев, В. В. Алейников
Волгоградский государственный технический университет
[1]agk@gde.ru, [2] nikitashumeyko92@gmail.com

***Аннотация*. Целью статьи является исследование схожести извлекаемых топиков из похожих классов патентов и возможности классификации данных документов по общей модели. Оптимальное количество топиков можно выбрать по интерпретации получающихся тем (например, экспертная оценка) на предмет связности слов в теме и отражения общего дискурса. В представленном наборе документов известна только общая тематика, предположить, какие подтемы могут выделиться не представляется возможным. В ходе исследования рассмотрена динамика изменения качества моделей при изменении параметров, по которой выбраны относительно оптимальные параметры. Однако вопрос оптимизации моделей требует более детального рассмотрения.**

Ключевые слова: машинное обучение; патентный поиск; тематическое моделирование; обработка естесвенного языка; латентное размещение Дирихле

I. Введение

Латентное размещение Дирихле [1] (LDA, от англ. Latent Dirichlet allocation) — применяемая в машинном обучении и информационном поиске порождающая модель, позволяющая объяснять результаты наблюдений с помощью неявных групп, благодаря чему возможно выявление причин сходства некоторых частей данных. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем и что появление каждого слова связано с одной из тем документа. В LDA каждый документ может рассматриваться как набор различных тематик. Подобный подход схож с вероятностным латентно-семантическим анализом (pLSA) с той разницей, что в LDA предполагается, что распределение тематик имеет в качестве априори распределения Дирихле. На практике в результате получается более корректный набор тематик.

Тематическая модель (topic model) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов. На выходе для каждого документа выдаётся числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем. Размерность этого вектора, равная числу тем, может либо задаваться на входе, либо определяться моделью автоматически. [2]

Перплексия [3] — критерий численной оценки качества вероятностной модели, равный экспоненте от минус усреднённого логарифма правдоподобия:

$$P = exp(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} ln\, p(w|d))$$

где n — длина коллекции в словах. Перплексия зависит от мощности словаря и распределения частот слов в коллекции.

$$p(w) = n_w/n$$

II. Автоматизация анализа патентной информации

Система автоматического позиционирования заявочных материалов на получение патента на изобретение в глобальном патентном пространстве на основе статистико-семантических подходов Cyber Examiner – это система для поддержки принятия решения экспертом при анализе заявки на получение патента. Пилотный проект системы Cyber Examiner был реализован по заказу Всемирной организации интеллектуальной собственности (Швейцария) [4].

Одним из наиболее важных этапов реализации системы является определение списка патентов, релевантных поданной заявке (рис. 1, 2) [5].

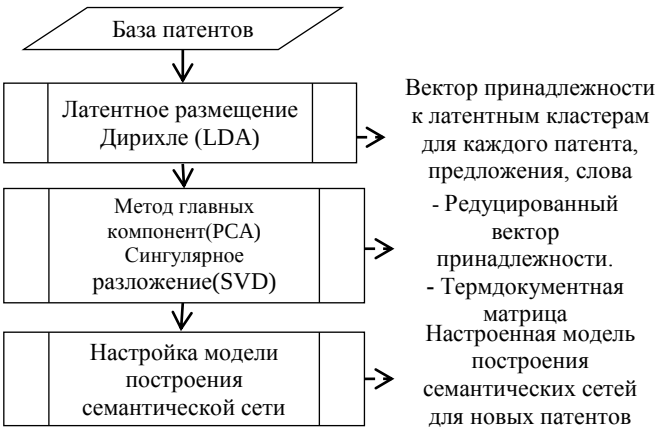

Рис. 1. Алгоритм обработки существующей базы патентов

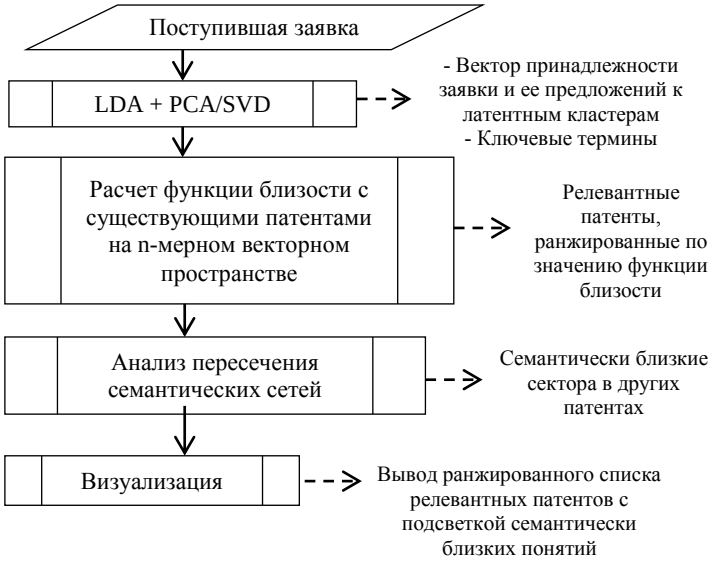

Рис. 2. Алгоритм обработки поступившей заявки

Текст заявки поступает в систему через веб-интерфейс [6]. Самая главная информация хранится в формуле изобретения. Именно уникальность этой информации необходимо проверить эксперту.

III. Обучение моделей и эксперименты

A. *Исходные данные*

В качестве исходных данных использовались тексты заявок пяти классов патентов:

A (HUMAN NECESSITIES),

B (PERFORMING OPERATIONS; TRANSPORTING),

G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING),

H (PHYSICS),

F (ELECTRICITY).

Исходные файлы представлены в формате XML, из которых для обучения извлекался раздел «Формула изобретения» (Claims). Был разработан скрипт парсинга XML-файлов.

Извлеченные пункты формулы изобретения собирались в единую строку. При этом для увеличения статистической значимости зависимые пункты формул уточнялись по первой ссылке (вида «по п.1») на другие пункты.

Таким образом, документ патента представлял собой строку из набора пунктов формулы изобретения, раскрытых при необходимости до первой ссылки на иные пункты.

Порядок текстовой обработки включал следующие этапы:

- токенезацию (встроенные средства python);
- приведение к нижнему регистру (встроенные средства python);
- отбрасывание токенов, длинной меньше двух символов (т.к. наблюдалось выраженное содержание элементов формул) (встроенные средства python);
- удаление пунктуации и стоп-слов (пакет nltk);
- лемматизация слов (пакет pymorphy2).

Для каждого класса были сформированы обучающая (4 000 патентов) и тестовая (1 000 патентов) выборки.

Для обучения модели использовалась библиотека gensim, полученные модели визуализировались средствами библиотеки pyLDAvis.

Цель экспериментов - исследование зависимости достигаемого качества модели и времени обучения от значений параметров.

Условия проведения экспериментов. Серия экспериментов проводится с реализацией LDA в библиотеке gensim (версия функции с распараллеливанием обучения). Настраиваются параметры:

- количество проходов обучения по коллекции (P);
- гиперпараметры модели (значение параметра α, параметр β дублировался);
- количество извлекаемых топиков (K).

B. *Количество проходов*

Из пяти обучающих выборок a-train.sample, b-train.sample, f-train.sample, g-train.sample и h-train.sample была сформирована общая, на которой и обучалась модель со следующими параметрами:

- количество скрытых тематик: 2;
- количество проходов по коллекции документов: 1, 5, 10, 15, 20, 25, 30, 50;
- остальные параметры по умолчанию.

Результаты серии экспериментов представлены на рис. 3, из которых видно, что увеличение проходов увеличивает время обучения и при количестве проходов более 8 временные затраты несопоставимо увеличиваются по сравнению с точностью. В дальнейших экспериментах будем использовать значение параметра равное 10 проходам по коллекции.

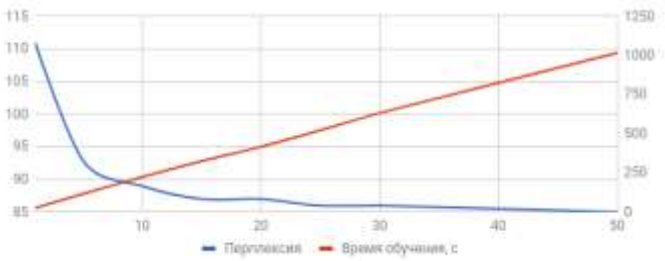

Рис. 3. График перплексии и времени обучения

C. Оценка влияния гиперпараметров на качество модели

Подбор геперпараметров модели предполагает поиск значений путем перебора некоторых значений в интервале (например [0,2]) с небольшим шагом, что довольно трудоемко. В ряде источников [7, 8] говорится об эмпирическом подборе данных параметров. В ходе экспериментов применялись эмпирические значения гиперпараметров и исследовалась тенденция изменения перплексии модели.

Статичные параметры.

Обучающая выборка: коллекция патентов (20 тыс. документов)

Количество топиков K: 2

Количество проходов P: 20

Изменяемые параметры.

Гиперпараметры модели α: 0,01; 0,1; 0,3; 0,5; 1,1; 1,25; auto (библиотека подбирает наилучшее значение сама); default (по умолчанию режим simmetric)

Сопоставление изменяемых параметров визуализировано на рис. 4.

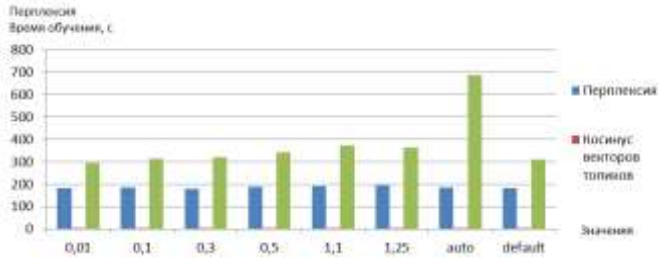

Рис. 4. Поведение модели при изменении гиперпараметров

В результате лучшим значение параметра α из представленного набора является коэффициент 1.1. Значение автоподбора параметров библиотекой не выделяется, зато время обучения существенно увеличилось. Т.к. в среднем значения перплексии не сильно изменяется при разных значениях гиперпараметров (и возможно будет зависеть от набора данных и прочих параметов) в следующих экспериментах оставим значение по умолчанию.

D. Поиск количества скрытых тематик

Цель экспериментов - исследование схожести извлекаемых топиков из похожих классов и возможности классификации по общей модели.

Оптимальное количество топиков можно выбрать по интерпретации получающихся тем (например, экспертная оценка) на предмет связности слов в теме и отражения общего дискурса. В представленном наборе документов известна только общая тематика, предположить, какие подтемы могут выделиться, не представляется возможным.

Сделаем предположение, что чем разнороднее темы (при определенном K), тем удачнее они выделены. Сравнение похожести векторов тем проводится с помощью косинусной меры.

Для каждой модели, независимо от изменяемых параметров, сохранен следующий набор характеристик:

- Файл обучающих данных;
- Количество извлекаемых топиков;
- Длина документа / словаря;
- Время обучения модели;
- Значение перплексии для модели;
- Топики с наборами 30 наиболее популярных слов по каждой из них;
- Косинусная мера между всеми топиками модели;
- Визуализация представления топиков модели (библиотека pyLDAvis).

Параметры модели (рис. 5–10):

- Количество скрытых тематик: 2, 3, 4, 5, 6, 7;
- Количество проходов по коллекции документов: 10
- Остальные параметры по умолчанию.

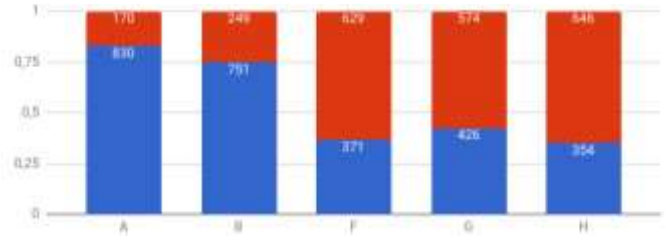

Рис. 5. Распределение классов патентов по 2 темам

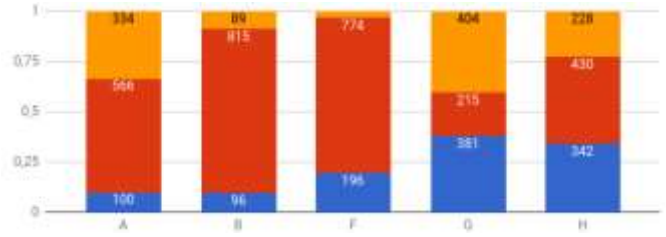

Рис. 6. Распределение классов патентов по 3 темам

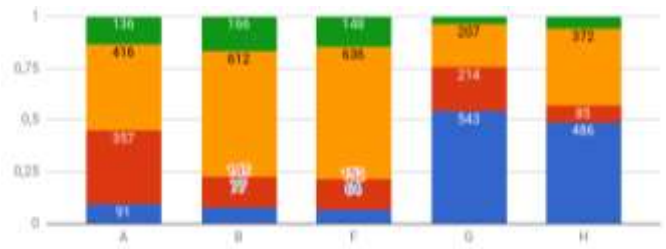

Рис. 7. Распределение классов патентов по 4 темам

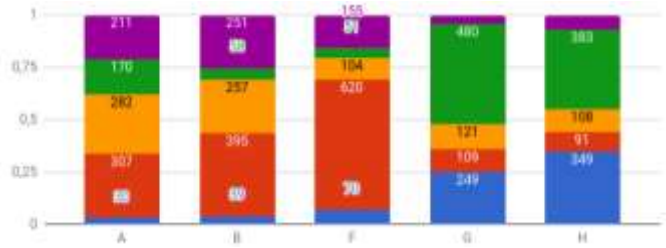

Рис. 8. Распределение классов патентов по 5 темам

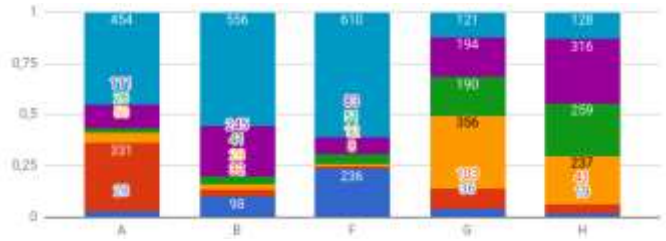

Рис. 9. Распределение классов патентов по 6 темам

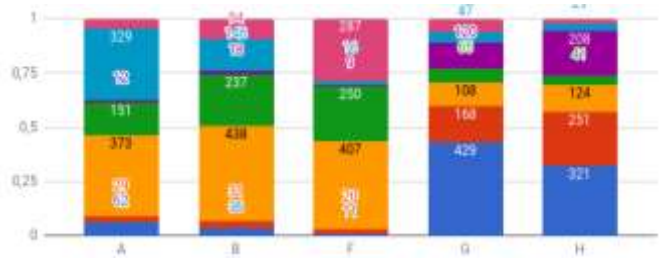

Рис. 10.Распределение классов патентов по 7 темам

IV. Выводы и обсуждение результатов

Исходя из полученных результатов, можно сделать следующие выводы. При распределении представленной коллекции документов по двум темам, можно выделить ярко выраженное сходство двух классов патентов A (HUMAN NECESSITIES) и B (PERFORMING OPERATIONS; TRANSPORTING), и менее выраженное, но присутствующее сходство классов G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING), H (PHYSICS) и F (ELECTRICITY).

При распределении классов патентов по 3 темам – очевидным является наличие общего у следующих классов – A (HUMAN NECESSITIES), B (PERFORMING OPERATIONS; TRANSPORTING), G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING). При распределении классов патентов по 4 темам, можно наблюдать схожее распределение, что и при распределении по трем темам.

Весьма близкими будут результаты при разбиении на 5, 6 и 7 тематик, с той лишь разницей, что при распределении классов по 5 и 7 темам можно выделить схожесть по одной из тематик у классов H (PHYSICS) и F (ELECTRICITY), а при распределении по 5 темам лишь у классов A (HUMAN NECESSITIES), B (PERFORMING OPERATIONS; TRANSPORTING), G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING), собственно как и в экспериментах 2, 3 и 4.

Таким образом, можно заключить, что с помощью сформированной из пяти обучающих выборок общей модели удалось, в результате поиска по различному количеству общих тематик, выявить следующие наиболее близкие классы из рассматриваемых в данной работе: A (HUMAN NECESSITIES), B (PERFORMING OPERATIONS; TRANSPORTING), G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING). Также, некоторые прогоны показали, что скрытой схожестью обладают классы (PHYSICS) и F (ELECTRICITY). Также, можно заключить, что распределение по меньшему количеству тем дает более ярко выраженный результат. Так, в первом эксперименте очевидной схожестью обладали классы A и B, при дальнейшем увеличении числа общих тематик, это сходство хоть и не пропало, но стало менее заметным.

В результате проделанной работы были исследованы результаты работы алгоритма LDA на пяти классах русскоязычных патентов.

Рассмотрена динамика изменения качества моделей при изменении параметров, по которой выбраны относительно оптимальные параметры. Однако вопрос оптимизации моделей требует дальнейшего более детального исследования.

Проведены сравнения выделенных тем на основе косинусной меры, по результатам которых можно грубо провести оценку качества кластеризации. Т.к. при большом количестве топиков (рис. 7–9) возрастает количество схожих векторов. В целом проблема выбора количества кластеров относится к вопросу интерпретации содержимого и предполагает более глубокую проработку.

Список литературы

[1] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022.

[2] MachineLearning – статья тематическое моделирование http://www.machinelearning.ru/wiki/.

[3] Brown, Peter F.; et al. (March 1992). "An Estimate of an Upper Bound for the Entropy of English". Computational Linguistics. 18, 2007.

[4] Korobkin, D., Fomenkov, S., Kravets, A., Kolesnikov, S., Dykov, M. Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting (2015) Communications in Computer and Information Science, 535, pp. 124-136.

[5] Korobkin, D., Fomenkov, S., Kravets, A., Kolesnikov, S. Methods of statistical and semantic patent analysis (2017) Communications in Computer and Information Science, 754, pp. 48-61.

[6] Kravets, A., Shumeiko, N., Lempert, B., Salnikova, N., Shcherbakova, N. “Smart Queue” Approach for new technical solutions discovery in patent applications (2017) Communications in Computer and Information Science, 754, pp. 37-47.

[7] Kravets, A.G., Mironenko, A.G., Nazarov, S.S., Kravets, A.D. Patent application text pre-processing for patent examination procedure (2015) Communications in Computer and Information Science, 535, pp. 105-114.

[8] Kravets, A. G., Kravets, A. D., Rogachev, V.A., Medintseva, I. P.Cross-thematic modeling of the world prior-art state: rejected patent applications analysis (2016) Journal of Fundamental and Applied Sciences, Vol. 8, SI 3, pp. 2542-2552