Выделение библиографического блока в научных текстах

Александр Огальцов

Компания Антиплагиат, Высшая Школа Экономики
ММРО-2019

29 ноября 2019

Мотивация

Рассматриваемая задача — выделение библиографии из научного текста.
Связанные подзадачи:

- Выделение отдельных записей;
- Определение основных метаданных записи;
- Построение графа цитирований;
- Исключение библиографии при поиске неправомерных заимствований.

Задача нетривиальна из-за разброса стилей оформления библиографии.

Методы решения задачи

- Шаблоны, паттерны, правила — подходит для заданного формата документа или узкого набора форматов.
- Классификация текстовых сегментов.
- Использование форматирования — подходит для работы с pdf-документами.
- Алгоритмы HMM, CRF — учёт контекста и признаков форматирования.

Проблемы

- Разнообразие стилей документов выборки: существующие решения обучались и тестировались на узком наборе форматов.
- Разнообразие расширений документа: требуется, чтобы алгоритм работал с любым форматом.
- Отсутствие выборок на русском языке, удовлетворяющих критериям разнообразности стилей и форматов.
- Зашумленность текстового слоя документа.
- Высокие требования к скорости работы в условиях пиковых нагрузок.

Пример

СПИСОК ЛИТЕРАТУРЫ

1. Актуализированные ценности современного российского общества / отв. ред. И. А. Халий. М.: Институт социологии РАН, 2015. 273 с. URL: http://www.isras.ru/files/File/publ/Act_zennosti_sovr_obschestva.pdf (дата обращения: 12.10.2018).
2. В Североуральске для жителей открыты пункты выдачи бутилированной воды // Главное управление МЧС России по Свердловской области: оф. сайт. URL: http://66.mchs.gov.ru/operationalpage/operational/item/7199391/ (дата обращения: 21.10.2018).
3. Государственные доклады «О состоянии и об охране окружающей среды Свердловской области» 2006-2017 гг. // Министерство природных ресурсов и экологии Свердловской области: оф. сайт. URL: https://mprso.midural.ru/article/show/id/1126 (дата обращения: 21.10.2018).
4. Ермаков Д. С. Оценка прогресса и перспективы образования для устойчивого развития в России / Д. С. Ермаков // Вестник РУДН. 2010. № 2. С. 99-104.
5. Ермаков Д. С. Педагогическая концепция формирования экологической компетентности учащихся: автореф. дис. … докт. пед. наук / Д. С. Ермаков. М., 2009.

(a) Обычный биб. блок

с. 74–77].

Возвращаясь к главному вопросу, поставленному в статье, – сформированы ли в настоящее время единые основы методики расследования преступлений, посягающих на половую неприкосновенность малолетних и несовершеннолетних, – полагаем возможным ее дальнейшая детализация стимулируются рядом малоисследованных либо вновь выявленных элементов криминалистической характеристики данной видовой группы преступлений, кардинально влияющих на содержание процесса их расследования. Они составляют предмет отдельного рассмотрения.

1. Форма «1-П» (455), книга 11 «Сводный отчет по России "Сведения о преступлениях, по которым имеются потерпевшие"». 2017–2018 гг.

2. Корнакова С.В. Содержание и значение криминалистической характеристики преступления (на примере насильственных половых преступлений, совершенных несовершеннолетними в отношении несовершеннолетних) // Рос. следователь. 2019. № 5.

3. Брусницын Л.В. Новые правила допроса несовершеннолетних потерпевших и свидетелей на предварительном следствии и в суде // Уголовное право. 2015. № 3.

1. Form «1-P» (455), book 11 «Summary report on Russia "Information about the crimes for which there are victims"». 2017–2018.

2. Kornakova S.V. The content and value of the criminalistic characteristics of the crime (for example, violent sexual crimes committed by minors in relation to minors) // Russian investigator. 2019. № 5.

3. Brusnitsyn L.V. New rules for the interrogation of juvenile victims and witnesses during the preliminary investigation and in court // Criminal law. 2015.

4. Kalnitsky V.V., Ovchinnikova M.M. Video recording by the investigator of the testimony of a

(b) «Интересный» биб. блок

Предлагаемое решение: постановка задачи

Пусть документ последовательность строк

$$D = \{l_1, \ldots, l_N\},$$

обозначим классы текст/библиография как 0, 1 соответственно. Тогда требуется построить отображение

$$f: \ \{l_1, \ldots, l_{|D|}\} \longrightarrow \{0,1\}^{|D|}.$$

$$\hat{f} = \arg\max_{f \in \mathcal{F}}(F_1(f(D)),$$

где $\mathcal{F}$ — множество рассматриваемых моделей.
Основные предположения:

- Текст разбит на строки;
- Строка принадлежит одному из классов — текст/библиография.

Предлагаемое решение: критерии качества

Задача поставлена как бинарная классификация с несбалансированными классами (примерно 1/15), поэтому в качестве критериев качества будем использовать $Precision$, $Recall$ и их среднее гармоническое F-критерий:

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN},$$

$$F = \frac{2 \cdot P \cdot R}{P + R},$$

- TP — библиографические строки, определенные верно;
- TN — строки текста, определенные верно;
- FP — строки текста, определенные как библиографические;
- FN — библиографические строки, определенные как строки текста.

Данные

- Выборка из 1000 документов (800/100/100) была подготовлена и размечена вручную.
- Документы отбирались по принципу максимальной вариативности по стилю и формату документа.
- В выборку помимо прочих типов вошли сборники трудов конференций, которые имеют множество библиографических блоков, которые не располагаются в конце документа.

Алгоритм

Алгоритм состоит из двух основных моделей:

- Модель, основанная на правилах.
  - Использует ключевые слова и нумерацию.
  - Имеет высокую точность и низкую полноту.
- Признаковая модель.
  - Предсказывает вероятность принадлежности к классу.
  - Применяется операция склейки близких строк

Модель на правилах

Псевдокод:

```
Result: Rule-based method
for line in Text do
    if line in keywords-set then
        if numeration in next k lines then
            while numeration lasts do
                add line to references
            end
        end
    end
end
```

Примеры начала списка литературы:

- Bibliograficheskij spisok,
- Материалы судебной или иной юридической практики,
- Научная литература и материалы периодической печати

Признаковая модель

Случайный лес был обучен на следующем наборе признаков:

- Дату с 18xx по 20xx;
- Ключевое слово для начала библиографии;
- Инициалы в разном формате;
- Номера страниц;
- Номера выпусков;
- URL;
- Специальные термины.
- Строка начинается с числа, за которым следует пробел и буква;
- Доля слов, длиннее 3 букв;
- Длина строки.

Признаковая модель: склейка

Находятся только наиболее вероятные строки, затем делаются операции склейки-удаления:

- Объединить библиографиеские строки, отстоящие друг от друга менее чем на K строк;
- Удалить блоки с менее чем L идущих подряд библиографических строк.
- Объединить библиографиеские строки, отстоящие друг от друга менее чем на M строк;
- Удалить блоки с менее чем N идущих подряд библиографических строк.

Алгоритм склейки: пример

Список литературы

1. Polezhaev V.A.: Automated citation graph building from a corpora of scientific documents. In: Computer Research and Modeling, 2012, vol. 4, no. 4, pp. 707–719
2. Vasiliev A., Kozlov D., Samusev S., Shamina O.: Automatic document metadata extraction from Russian scientific articles. In: Proceedings of RCDL2007, 2007, vol. 1, pp. 175–181
3. Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic Document Metadata Extraction Using Support Vector Machines. In: JCDL, 2003
4. Beeferman, D., Berger, A., Lafferty, J.: Statistical Models for Text Segmentation. In: Machine Learning, 1999, no. 34. pp. 177–210
5. Ogaltsov, A.V., Bakhteev, O.Y.: Automatic metadata extraction from scientific PDF documents. In: Inform. Primen., 2018, no. 12:2 , pp 75–82

Список литературы

1. Polezhaev V.A.: Automated citation graph building from a corpora of scientific documents. In: Computer Research and Modeling, 2012, vol. 4, no. 4, pp. 707–719
2. Vasiliev A., Kozlov D., Samusev S., Shamina O.: Automatic document metadata extraction from Russian scientific articles. In: Proceedings of RCDL2007, 2007, vol. 1, pp. 175–181
3. Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic Document Metadata Extraction Using Support Vector Machines. In: JCDL, 2003
4. Beeferman, D., Berger, A., Lafferty, J.: Statistical Models for Text Segmentation. In: Machine Learning, 1999, no. 34. pp. 177–210
5. Ogaltsov, A.V., Bakhteev, O.Y.: Automatic metadata extraction from scientific PDF documents. In: Inform. Primen., 2018, no. 12:2 , pp 75–82

Интегральный алгоритм: качество

Итоговый результат — объединение строк, отобранных, как библиографические от обеих моделей.
Метрики качества на тестовой выборке:

	Precision	Recall
Модель на правилах	0,91	0,68
Признаковая модель	0,88	0,8
Интегральный алгоритм	0,94	0,9

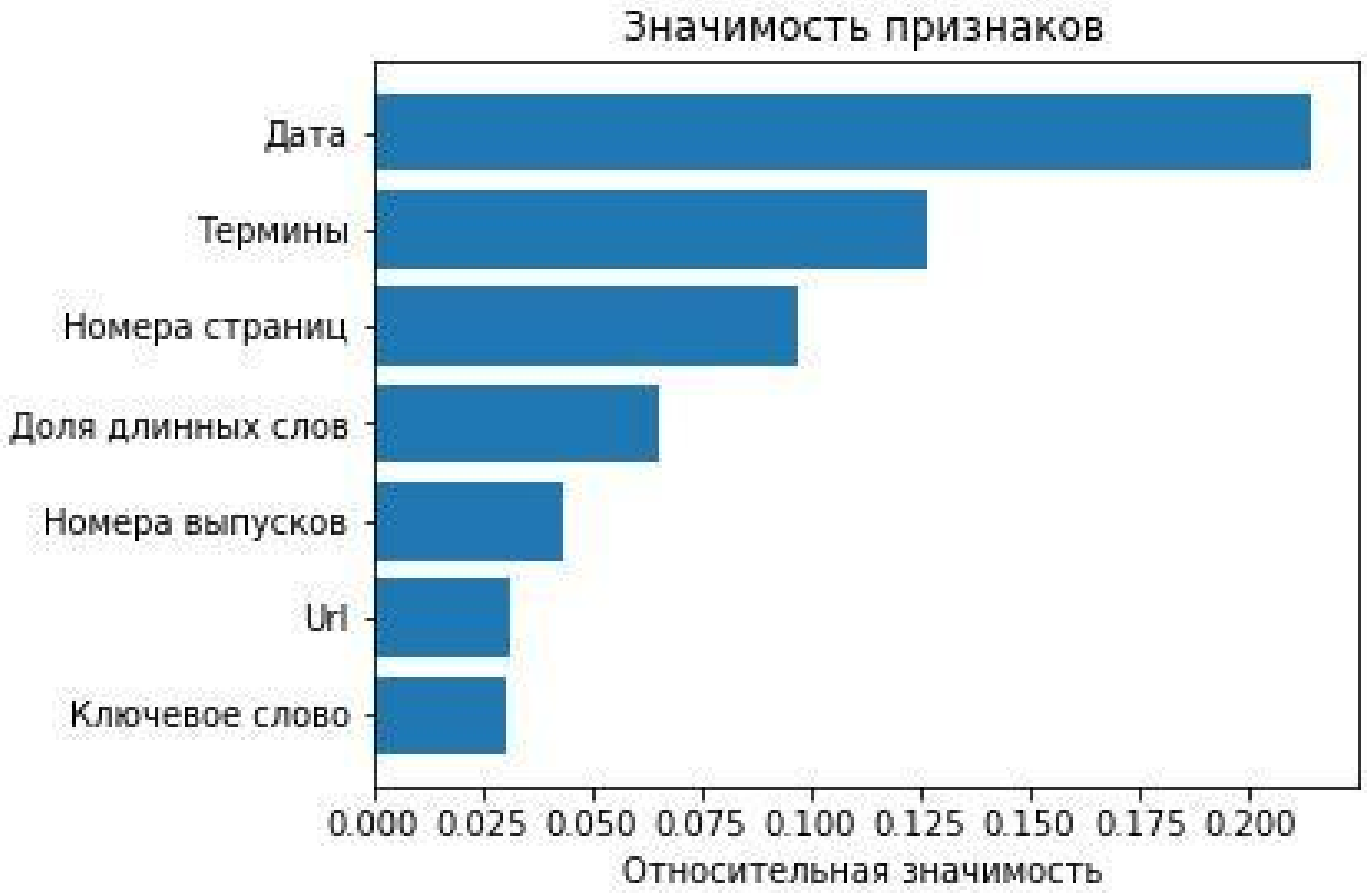
Сравнение с существующим решением

- Стресс-тест 100 pdf-документов из разных журналов и областей науки.
- В основном это двухколоночные статьи со смесью русского и английского языка.
- Сравнение с open-source инструментом Cermine, использующим информацию о форматировании и SVM.

	Precision	Recall	F-score
Cermine	0.74	0.47	0.49
Предложенный метод	**0.91**	**0.73**	**0.77**

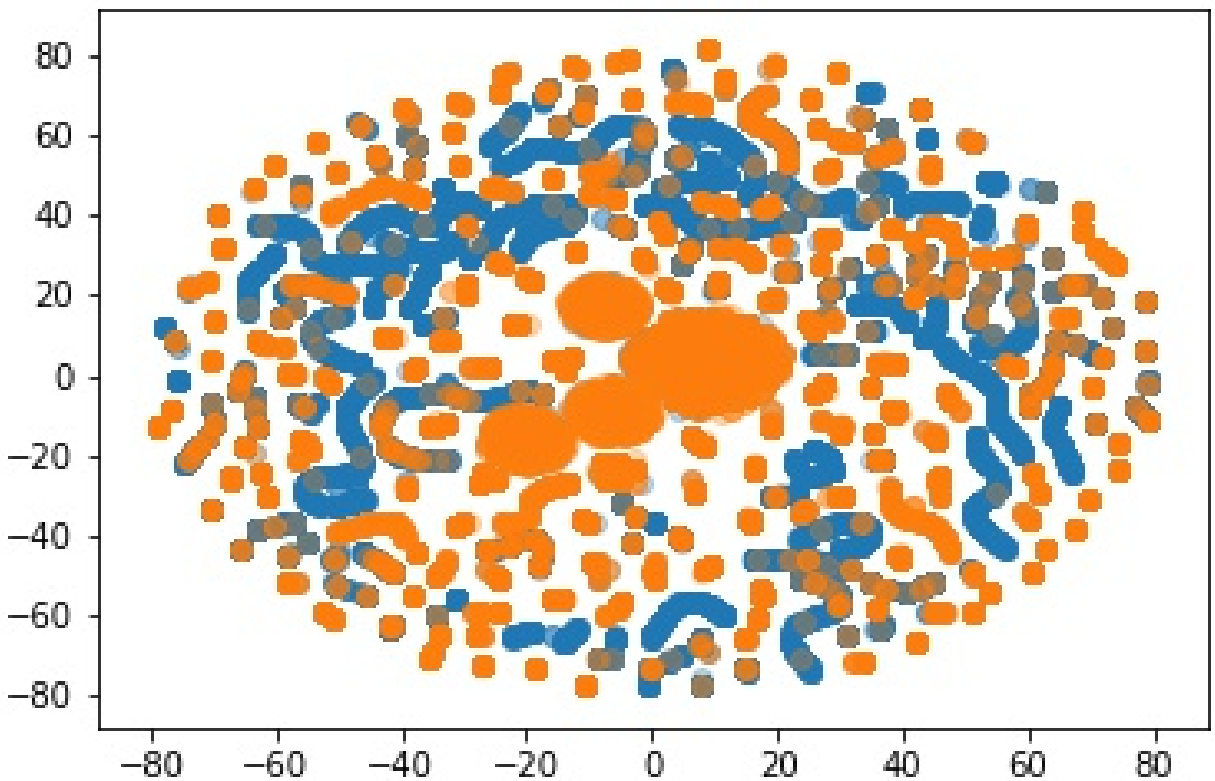
Анализ важности признаков

Был проведен качественный анализ полученной модели. На рисунке представлены относительные значимости признаков при классификации тестовых объектов.



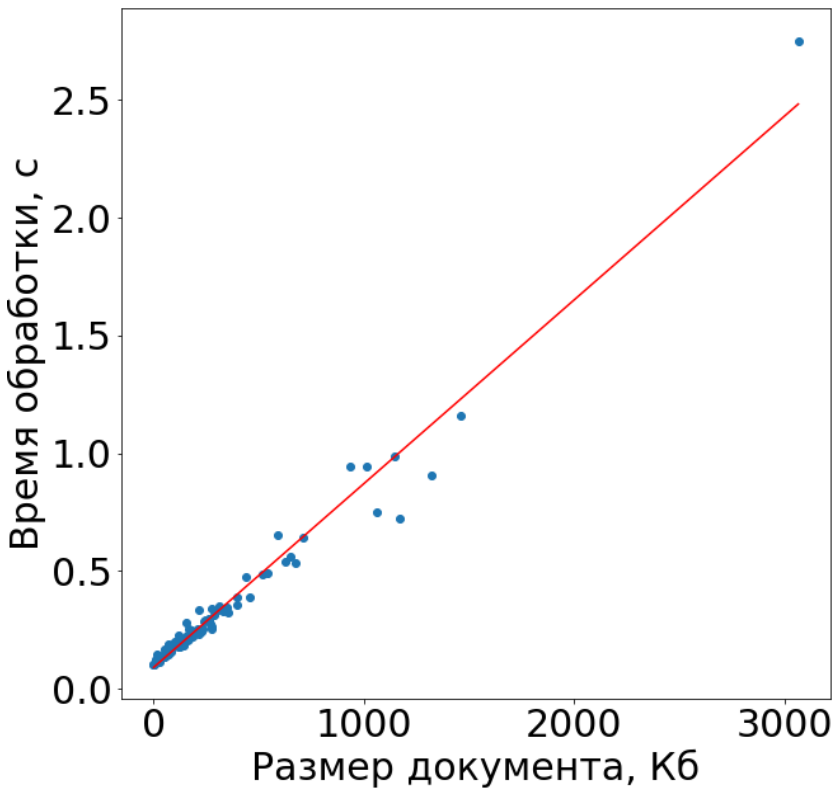
Проекция признакового описания объектов

Синим цветом обозначены библиографические строки.



Производительность

Время обработки документа в зависимости от размера текстового слоя.



Заключение

Предложен алгоритм извлечения библиографического блока со следующими полезными свойствами:

- Универсальность относительно форматирования документа
- Универсальность относительно расширения документа
- Устойчивость к зашумлённости текстового слоя
- Производительность 300 документов в минуту

Планы:

- Поддержка других языков
- Выделение отдельных биб. записей и их парсинг