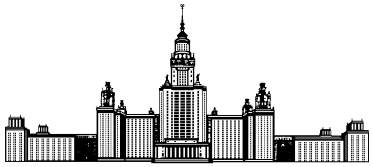
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических методов прогнозирования

КУРСОВАЯ РАБОТА

«Языковые модели для ранжирования фраз в полуавтоматической суммаризации научных статей»

Выполнила:

студентка 1 курса 117 группы

Дзюба Мария Эдуардовна

Научный руководитель:

д.ф-м.н., профессор РАН

Воронцов Константин Вячеславович

Москва, 2024

Содержание

1 Введение

С каждым годом объем научных исследований многократно возрастает, что приводит к значительному увеличению нагрузки на ученых во время проведения предварительного анализа и изучения предложенных ранее методов решения поставленных задач.

При написании научной работы важно структурировать информацию и размещать исследованную литературу в порядке, который будет наиболее понятен читателю. Для этого можно опираться на ранее написанные статьи по определенной тематике, чтобы выявить основные тенденции и приемы, используемые для ранжирования связанных материалов.

Системы суммаризации особенно полезны для ученых и экспертов, которым приходится тратить значительное количество времени на чтение научных публикаций. В настоящей работе рассматривается задача полуавтоматической суммаризации, цель которой состоит в оказании помощи пользователям при написании авторского обзора (реферата, дайджеста) по заданной тематической подборке.

Основной целью данного исследования является исследование методов ранжирования связанных статей при написании собственного исследования. При этом данные методы могут быть использованы на одном из первых этапов полуавтоматической суммаризации текстов.

Практическая значимость и актуальность данной работы заключается в том, что предложенная система может быть встроена в поисково-рекомендательную систему формирования и анализа тематических подборок англоязычных научных статей «Мастерская знаний».

В текущей версии описанной системы пользователям необходимо самостоятельно упорядочивать цитируемые документы. Интеграция одного или нескольких предложенных алгоритмов позволит исследователям выбирать наиболее подходящий для их логики повествования метод ранжирования и логично размещать уже изученные статьи в процессе написания собственного исследования.

2 Постановка задачи

Пусть D - коллекция научных публикаций, содержащих ссылки на публикации из множества C. Для каждого документа $d \in D$ выделена последовательность из k_d ссылок

$$X_d = (x_{d1}, ..., x_{dk})$$

$$x_{di} \in C$$

Постановка задачи заключается в построение модели ранжирования ссылок, которая принимает на вход произвольное неупорядоченное подмножество $X \subset C$ и выдает на выходе отношение линейного порядка на X, наиболее согласованное с наблюдаемыми ранее последовательностями $X_D = \{X_d : d \in D\}$.

Согласованность модели ранжирования $f(x)$ с последовательностью X_d предлагается измерять с помощью коэффициента корреляции Кендалла, равного доли правильно ранжированных пар:

$$\tau_f(X_d) = \frac{2}{k_d(k_d - 1)} \sum_{i<j} [f(x_{di}) < f(x_{dj})].$$

Согласованность модели ранжирования $f(x)$ со всей совокупностью наблюдаемых данных X_D определяется путем усреднения $\tau_f(X_d)$ по всем последовательностям X_d:

$$\tau_f(X_D) = \frac{1}{|D|} \sum_{d \in D} \tau_f(X_d).$$

Коэффициент ранговой корреляции Кендалла принимает значения в отрезке [0 , 1], чем выше значение, тем лучше.

Для построения параметрических моделей ранжирования $f(x, w)$ удобно минимизировать число неверно отранжированных пар (i, j). При этом, для улучшения пользовательского опыта, будем стремиться минимизировать также и количество пар с одинаковым выходом функции, тем самым повышая однозначность предложенной итоговой расстановки.

$$Q(x, w) = \sum_{d \in D} \sum_{i<j} [f(x_{dj}, w) - f(x_{di}, w) \leq 0] \to \min_w$$

Функция $M_{ij} = f(x_{dj}, w) - f(x_{di}, w)$ называется *парным отступом* и используется для определения аппроксимированных оптимизационных критериев в попарных

методах обучения ранжированию:

$$Q(x, w) \leq \tilde{Q}(x, w) = \sum_{d \in D} \sum_{i<j} \mathcal{L}(M_{ij}(w) \to \min_{w},$$

где $\mathcal{L}(M) \geq [M \leq 0]$ - гладкая верхняя оценка функции потерь.

2.1 Определения и обозначения

Граф цитирований - в информатике и библиометрии представляет собой ориентированный граф, который описывает цитаты в коллекции документов. Каждая вершина в графе представляет документ в коллекции, и каждое ребро направлено от одного документа к другому, который он цитирует.

Модель ранжирования - функция $f : C \to \mathbb{R}$, с помощью которой элементы заданного неупорядоченного подмножества ссылок $X = \{x_1, ..., x_k\} \subset C$ могут быть отранжированы по возрастанию $f(x^{(1)}) \leq f(x^{(2)}) \leq ... \leq f(x^{(k)})$.

2.2 Обзор литературы

Существует множество методов для ранжирования ссылок, которые можно разделить на две категории: методы на основе графов и методы на основе машинного обучения.

2.2.1 Методы на основе графов

В поставленной задаче можем рассматривать цитирования как ссылки. Учитывая факт общедоступности графа цитирований, можем применить описанные ниже методы.

Методы на основе графов используют структуру ссылок между документами для определения их важности. Одним из таких методов является алгоритм PageRank[5], который был разработан для ранжирования веб-страниц. Он определяет важность страницы на основе количества ссылок, указывающих на нее, и значимость страниц, ссылающихся на нее. Аналогичные алгоритмы могут быть применены для ранжирования ссылок в задаче суммаризации.

Также были предложены различные подходы Weighted PageRank [6], [7], [8] - модификация алгоритма PageRank, которая учитывает не только количество ссылок,

указывающих на страницу, но и их вес. Вес ссылки может зависеть от различных факторов, таких как авторитетность сайта, на котором размещена ссылка, релевантность текста ссылки и т.д.

В уже существующих исследованиях были предложены следующие факторы, относительно которых выставляется вес для цитаты: Yu, Li и Liu (2004) [4] предложили добавлять вес каждой вершине графа в зависимости от года публикации работы.

2.2.2 Методы на основе машинного обучения

Одним из основных медотов ранжирования с помощью машинного обучения является использование градиентного бустинга над решающими деревьями. Наиболее популярными имплементациями являются XGBoost [2] и LightGBM [1]. Для дальнейшей работы был выбран алгоритм LightGBM из-за наличия возможности использования персонализированной оптимизационной функции.

3 Описание данных и используемого признакового пространства

3.1 Используемый корпус данных

В качестве данных предлагается использовать данные из коллекции S2ORC (The Semantic Scholar Open Research Corpus) [3]. В корпусе представлены данные о 81.1M статьях из разных научных областей, при этом полные тексты документов представленны для 8M статей. В этом наборе данных собраны статьи из сотен академических издательств и цифровых архивов. На сегодняшний день эта коллекция является самой крупной общедоступной коллекцией машиночитаемых академических текстов. Статьи данного корпуса объединены в граф цитирования, включающий 467 миллионов вершин.

Особенности данной коллекции документов:

1. В коллекции предоставляются метаданные о статьях, такие как авторы, год публикации, место публикации, ключевые слова и ссылки на другие исследования.

2. Для каждой статьи представлена аннотация.

3. Цитаты внутри абзацев связаны с элементами библиографии, приведенной в метаданных статьи, что позволяет однозначно сопоставить цитируемую статью с ее вершиной в графе цитирований.

4. При наличии текста статьи, он разделен на абзацы, внутри которых выделены цитирования, приложенные таблицы и рисунки.

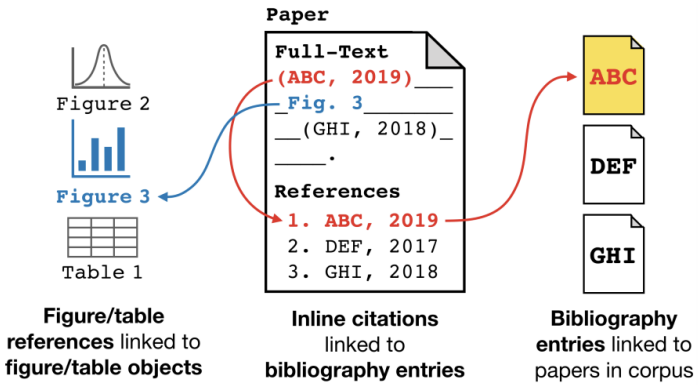


Рис. 1: Общий вид данных, представленных в выбранном корпусе

Общий объем данной коллекции составляет более 1Тб данных. В данной работе используется только его часть - ACL (Anthology of Computer Linguistics)[9]. Все статьи настоящего раздела посвящены компьютерным наукам и обработке естественного языка. Объем данной части S2ORC составляет 43к статей.

В качестве предварительной обработки и чистки данных были проведены следующие этапы:

1. удаления повторяющихся статей

2. удаления статей без полного текста работы

При этом, в данном исследование опустим требование на полноту метаданных в корпусе и попытаемся их восстановить. Для дальнейшего использования остаются доступны 31156 статей.

3.2 Анализ плотности распределения ссылок в зависимости от раздела

В данной части работы будет проведен анализ плотности цитирования в зависимости от раздела. Хочется проверить гипотезу о том, распределение плотности не совпадает в специализированных разделах и нет. Под специализированным разделом в этом случае будем рассматривать те части статей, в которых автор предполагал описание изученных работ других исследователей.

Для проведения данного анализа будем рассматривать всевозможные раззделы (вне зависимости от наименования) в которых есть хотя бы два цитирования. При этом из всего корпуса будем рассматривать топ-50 разделов, которые чаще всего встречаются.

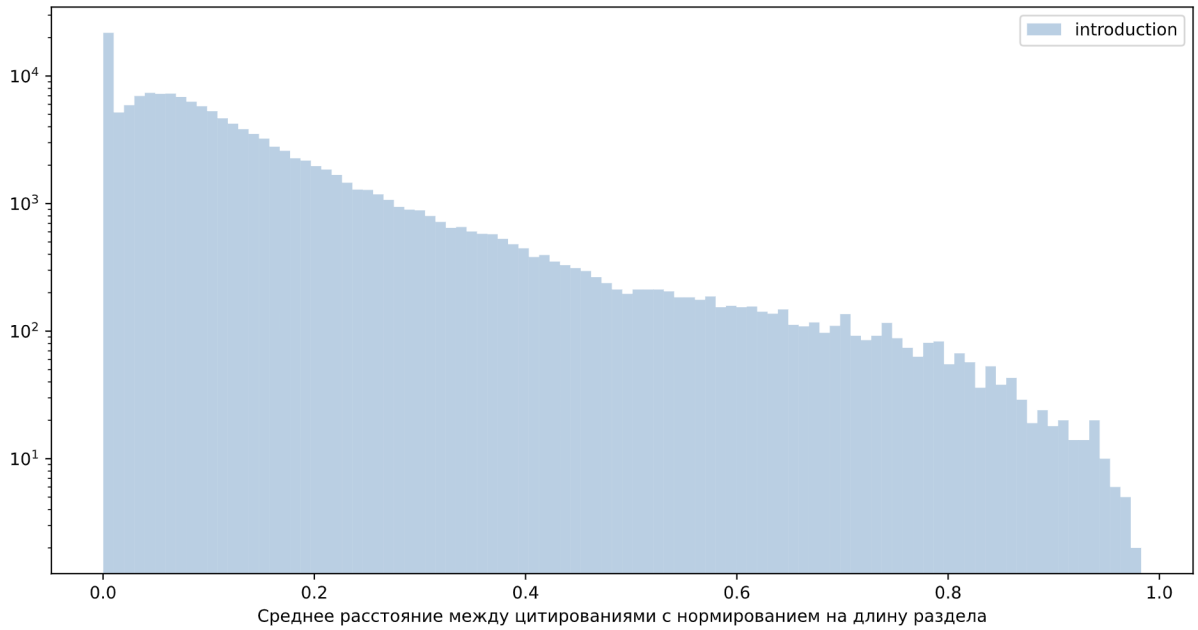
Примеры именований разделов приведены на изображение.

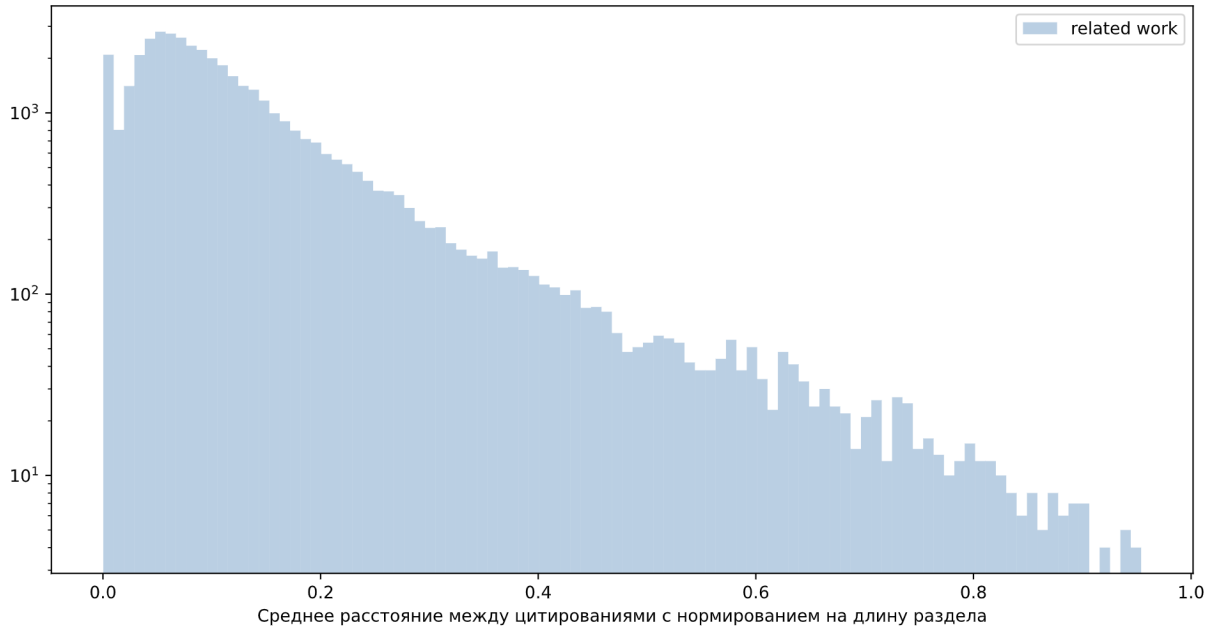
```
introduction
related work
conclusion
results
experiments
discussion
conclusions
evaluation
experimental setup
conclusion and future work
data
conclusions and future work
features
experimental results
results and discussion
background
model
method
methodology
error analysis
analysis
experiments and results
datasets
```

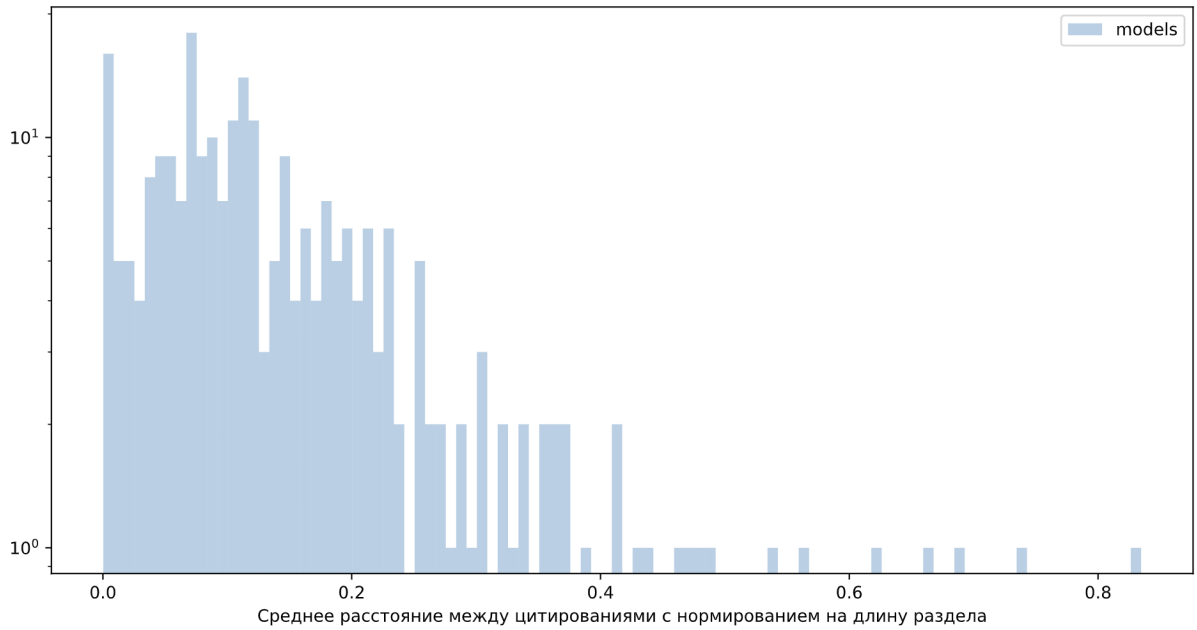
Рис. 2: Наиболее популярные названия разделов в корпусе данных

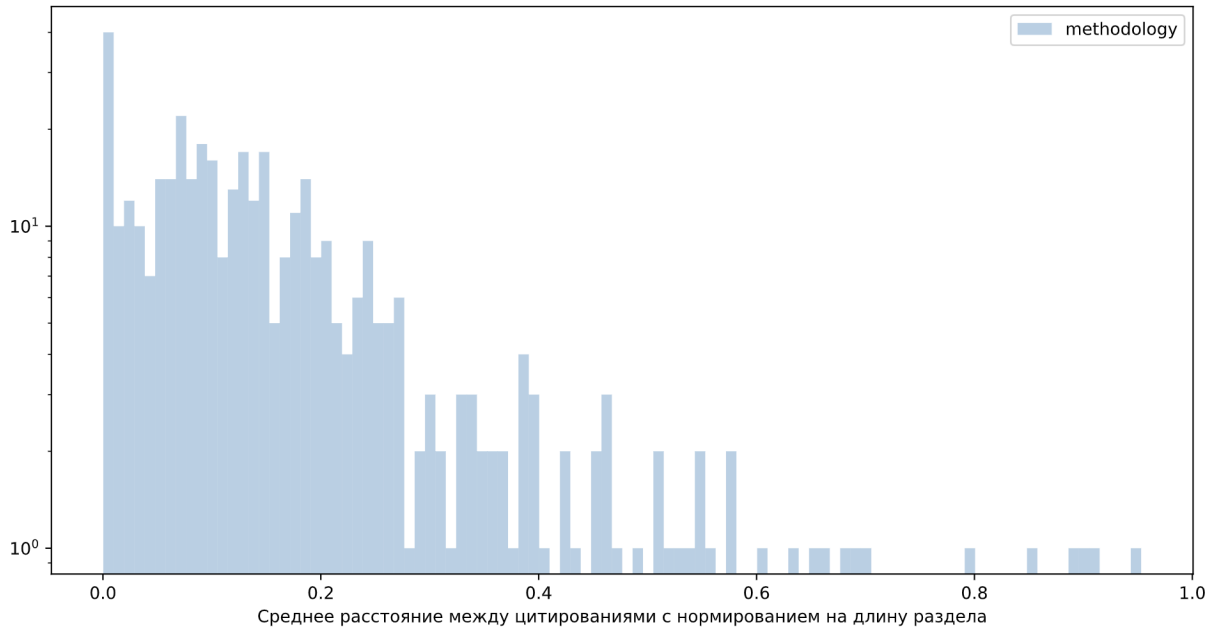
Для универсального сбора данных будет приводить все названия разделов к нижнему регистру.

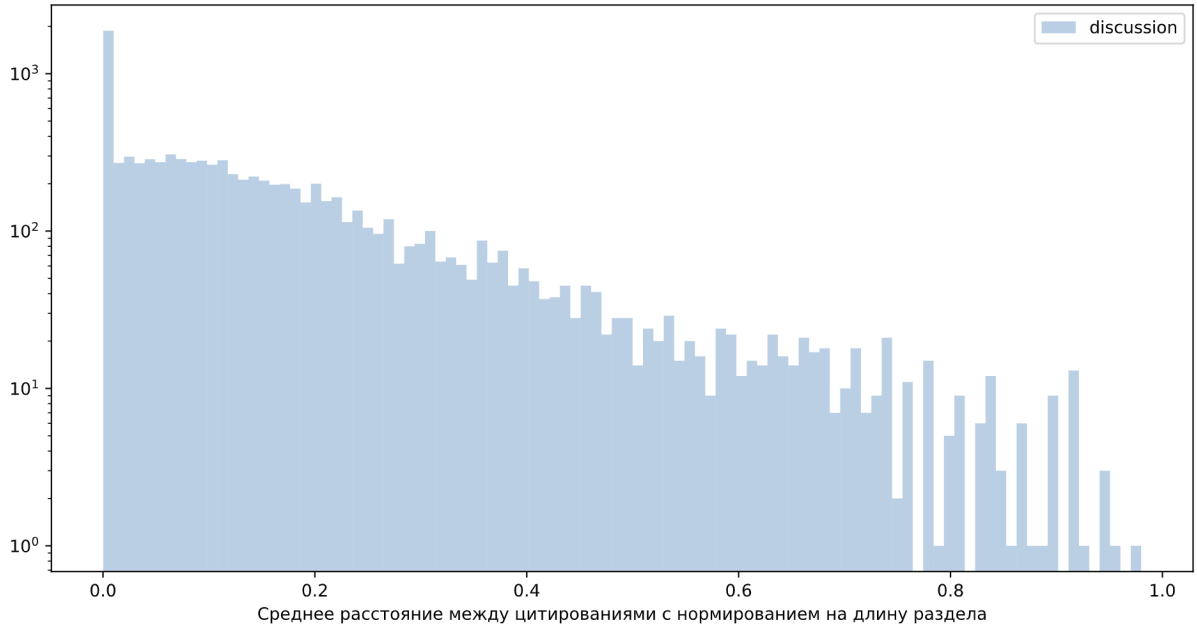
Приведем примеры распределений для нескольких «специализированных» разделов, и для не «специализированных».


introduction
10^4
10^3
10^2
10^1
0.0
0.2
0.4
0.6
0.8
1.0
Среднее расстояние между цитированиями с нормированием на длину раздела


related work
10^3
10^2
10^1
0.0
0.2
0.4
0.6
0.8
1.0
Среднее расстояние между цитированиями с нормированием на длину раздела


models
10^1
10^0
0.0
0.2
0.4
0.6
0.8
Среднее расстояние между цитированиями с нормированием на длину раздела


methodology
10^1
10^0
0.0
0.2
0.4
0.6
0.8
1.0
Среднее расстояние между цитированиями с нормированием на длину раздела



В данных примерах под «Специализированными» разделами понимаются разделы "Introduction"и "Related work".

Как видно из приведенных графиков, можно выделить тенденцию в близком расположении цитирований в «неспециализированных» разделах. При этом, в отведенных под это разделах (introduction/related work) упоминания уже проведенных исследований расположены как близко, так и далеко друг от друга.

3.3 Метод сбора упорядоченной последовательности ссылок

В качестве объекта выборки рассматривается пара <Исходная статья, Связанная статья>, а также характеристики каждой из статей.

Алгоритм сбора данных в необходимом для моделей машинного обучения формате состоит из следующей последовательности шагов - среди всех представленных абзацей статьи выделяем все, связанные с введением, обзором литературы и предыдущими исследованиями и объединяем их в одноименные части; внутри каждой части выделяем цитирования и однозначно сопоставляем их с элементами библиографии таким образом формируя набор пар <Исходная статья, Связанная статья>.

Для обучения и валидации поделим исходный датасет по году публикации. В качестве обучения будем рассматривать статьи, которые были опубликованы раньше

2018 года. В валидацию будем добавлять статьи, опубликованные после 2018 года, соответсвенно. Таким образом в обучение попали 24681 статей, в валидацию 6475 статей.

3.4 Признаковое пространство

В данном исследовании хотелось максимально ослабить необходимые к данным требования, так как для дальнейшей эксплуатации модели предполагается стремиться к минимальному ограничению пользователя системы при подборе статей.

В рамках условий описанных выше будем опираться на следующие общедоступные метаданные:

1. Год публикации
2. Авторы публикации
3. Заголовок статьи

Так же, при наличии будут использованы данные, которые могут быть получены путем анализа графа цитирований. Однако отсутствие этих данных не является останавливающим фактором при обработке последовательности ссылок.

В случае отсутсвия одного или нескольких факторов в обучающей и/или валидационной выборках предлагается заменять отсутствующие значения на заранее вычисленные статистики по всему датасету. При отсутствие значения категориального признака предлагается заменять его значение на специально зарезервированную для таких случаев категорию -1.

Вычисляемые на основе данных признаки:

1. Количество пересекающихся авторов
2. Разница в годе публикации
3. Принадлежность к определенному разделу (закодирована с помощью one hot encoding)
4. Год публикации исходной статьи

5. Год публикации цитируемой статьи

6. min/mean/max разницы в годах публикации рассматриваемых в группе статей

7. min/mean/max разницы в годах публикации внутри каждого из разделов в группе ранжируемых статей

4 Модели для решения поставленной задачи

4.1 Подходы к обучению и валидации моделей

В данной задаче можно выделить три подхода к обучению моделей.

В первом подходе все ссылки рассматриваются вне зависимости от раздела. Это позволяет создать наиболее полное и интегрированное представление о цитируемой литературе. Это может быть полезно для исследований, охватывающих широкий спектр тем и дисциплин. Однако, одним из недостатков этого подхода является то, что он может приводить к замешательству и путанице, так как разные разделы статьи могут иметь разные цели и контексты, что затрудняет точное понимание взаимосвязей между цитатами, особенно при повторе цитирующихся документов в различных разделах. При этом в данном подходе при наличие цитирование одной ссылки более одного раза будем учитывать только ее первое вхождение.

Второй подход является модификацией первого подхода. В нем так же будут рассматриваться ссылки вне зависимости от раздела, но в данном случае при наличии повторяющихся ссылок, они будут оставаться в датасете. При этом, будет добавлен признак принадлежности к конкретному разделу.

Третий подход предполагает отдельное рассмотрение последовательностей ссылок внутри каждого из выделенных разделов. Данный подход предоставляет более структурированное ранжирование. Этот метод помогает исследователям фокусироваться на конкретных аспектах их работы и контексте, в котором были сделаны цитаты. Он также обеспечивает лучшую организацию и более ясную картину, что может быть важным для детального анализа и интерпретации данных.

Далее в работе будут проведены сравнения целевых метрик каждого из подходов на двух вариантах валидации.

4.2 Набор базовых моделей

4.2.1 Модель на основе года публикации

Одной из наиболее базовых моделей является модель, основанная на сортиировке статей по году их публикации, начиная от самых ранних до наиболее поздних. Этот подход позволяет исследователям получить хронологическую последовательность развития исследования по интересующей теме.

В данной модели вероятно появления большого числа объектов с одинаковыми значениями алгоритма, что сильно снижает однозначность ранжирование системой.

4.2.2 Модели на основе статистики совместного распределения

Рассмотрим последовательность ссылок $(x_1, ..., x_k)$. Назовем дистанцией между эллементами x_i и x_j в этой последовательности разность $\delta_{x_i x_j} = j - i$. Дистанция δ_{uv} положительна, если u располагается левее v и отрицательна, если правее.

Обозначим через $R_{uv}(D)$ мультимножество значений дистанций δ_{uv}, наблюдаемых во всех последовательностях X_d, $d \in D$ для данной пары $(u, v) \in C^2$.

Ссылки в последовательности могут повторяться. По умолчанию будем полагать, что в случае повторных ссылок в $R_{uv}(D)$ заносятся только первые вхождения ссылок в последовательность.

По мультимножеству $R_{uv}(D)$ определяются эмперические оценки функции распределения, математического ожидания и дисперсии дистанций δ_{uv}:

$$F_{uv}(z) = \frac{1}{|R_{uv}|} \sum_{\delta \in R_{uv}} [\delta \leq z] - \text{выборочная функция распределения;}$$

$$\mu_{uv} = \frac{1}{|R_{uv}|} \sum_{\delta \in R_{uv}} \delta - \text{оценка средней дистанции;}$$

$$\sigma^2_{uv} = \frac{1}{|R_{uv}|} \sum_{\delta \in R_{uv}} (\delta - \mu_{uv})^2 - \text{оценка среднеквадратичного отклонения.}$$

Базовая эвристическая модель ранжирования $f_0(x)$ основана на том, чтобы посчитать по всей выборке, насколько чаще элемент x находится правее всех остальных элементов заданного (неупорядоченного) множества X, чем левее:

$$f_0(x) = \sum_{u \in X \setminus x} \frac{1}{|R_{uv}|} \sum_{\delta \in R_{uv}} [\delta \geq 0].$$

Второй вариант - учесть дистанцию, насколько x правее других элементов последовательности:

$$f_\delta(x) = \sum_{u \in X \setminus x} \frac{1}{|R_{uv}|} \sum_{\delta \in R_{uv}} \delta[\delta \geq 0].$$

Третий вариант - учесть все дистанции между x и другими объектами вне зависимости от их взаимного расположения. При этом, если x стоит правее объекта, то дистанция между ними добавляется с положительным знаком, если левее - с отрицательным:

$$f_{\delta_{all}}(x) = \sum_{u \in X \setminus x} \frac{1}{|R_{uv}|} \sum_{\delta \in R_{uv}} \delta.$$

4.2.3 PageRank

Алгоритм PageRank может быть представлен в виде следующего уравнения:

$$PR(x_k) = \frac{(1 - coeff)}{D} + coeff \cdot \sum_{i=1}^{n} \frac{PR(x_i)}{C_i},$$

где n - количество ссылающихся на документ x статей, $coeff$ - коэффицент затухания (положим его равным 0.85), D - общее число документов, C_i - количество ссылок, ссылающихся на i-й документ.

Для оценки PageRank для представленных статей будем использовать граф цитирований. Граф цитирований в данном случае будем строить на основе статей в обучающей части данных. При этом, при вычислении PageRank для тестовых данных, при отсутствии интересующей статью в графе цитирований - будем использовать среднее значение PageRank по обучающей выборке в качестве замены.

4.2.4 Модели на основе семантической близости статей

В качестве ответов ранжирующей модели будет рассматривать расстояние между заголовками исходной и связанной статей.

В качестве функции расстояния рассмотрим коссинусную близость двух векторных представлений заголовков рассматриваемых публикаций. В качестве модели определяющей векторные представления заголовка рассмотрим две предобученные языковые модели all-mpnet-base-v2 и all-MiniLM-L6-v2. Данные модели представляют собой языковые модели на основе трансформеров. Основной задачей, под которую

данные модели обучались, является задача семантической близости двух предложений.

4.3 Градиентный бустинг

Рассмотрим последовательность групп связанных ссылок $X_D = (X_1, ..., X_k)$ и их целевое ранжирование $(y_1, ..., y_k)$.

Строим алгоритм в виде:

$$f_n(x) = \sum_{i=1}^{n} b_i(x).$$

Пусть построен $f_t(x)$, тогда будем обучать алгоритм b_t на выборке $(x, -\mathcal{L}'(y, f_t(x)))$.

Тогда

$$f_{t+1}(x) = f_t(x) + b_t(x).$$

В задаче ранжирования антиградиент функции потерь $\mathcal{L}(y, f_t(x))$ вычисляется отдельно для каждой группы связанных ссылок X_d уже после сортировки документов по оценкам.

Минимизация $\mathcal{L}$ напрямую затруднительна из-за ее негладкости. Вместо этого будем использовать гладкие верхние оценки — функции, которые сверху оценивают $\mathcal{L}$ и являются непрерывно дифференцируемыми.

Логистическая функция потерь 1 :

$$\mathcal{L}_{\log(1+\exp(\text{-M}))} = \sum_{v<u} \log(1 + \exp(-(M)))$$

Логистическая функция потерь 2 :

$$\mathcal{L}_{log(1+M^2)} = \sum_{v<u} \log(1 + M^2)$$

Экспоненциальная функция потерь:

$$\mathcal{L}_{\exp(\text{M})} = \sum_{v<u} \exp(-(M))$$

Франк-Вульф:

$$\mathcal{L}_{\text{frank-wolfe}} = \sum_{v<u} \frac{1}{1 - M}$$

где $M(f(x), u, v) = f(x_u) - f(x_v)$.

4.4 Факторы для градиентного бустинга

В качестве факторов градиентного бустинга будем использовать признаки, описанные в разделе 3.3, а так же выходы моделей, описанных в 4.2. Помимо выходов модели посчитаем так же min/mean/max значений моделей внутри каждой группы документов, а так же внутри каждого из разделов внутри документа.

5 Анализ и сравнение работы моделей

5.1 Первый подход к обучению

В таблице ниже приведены замеры работы моделей, обученных на полный текст с удалением повторяющихся ссылок. Замер без нейросетевых фичей.

Модель	Результат
Year model	0.258
$L_{log(1+exp(M))}$	0.248
$L_{(log(1+M^2)}$	0.247
$L_{exp(M)}$	0.270
$L_{frank-wolfe}$	0.270
L_{M^2}	0.271

Таблица 1: Результаты

Замеры работы моделей на полном тексте с удалением повторов и использованием нейросетевых факторов

Модель	Результат
Year model	0.258
$L_{log(1+exp(M))}$	0.258
$L_{(log(1+M^2)}$	0.257
$L_{exp(M)}$	0.265
$L_{frank-wolfe}$	0.270
L_{M^2}	0.263

Таблица 2: Результаты

Так же были проведены замеры работы моделей на полном тексте с удалением повторов и нейросетевыми фичами, выходы которых были просуммированы с годом публикации статьи и отранжированы по этому значению.

Модель	Результат
Year model	0.258
$L_{log(1+exp(M))}$	0.294
$L_{(log(1+M^2)}$	0.3
$L_{exp(M)}$	0.298
$L_{frank-wolfe}$	0.310
L_{M^2}	0.3

Таблица 3: Результаты

Из приведенных таблиц, можем сделать вывод, что добавление подобранных нейросетевых факторов не дает значимого прироста к качеству модели, при этом усложняет инференс моделей, а так же их использование в реальном времени. Поэтому дальнейшие анализы подходов к обучению будут проведены без использования нейросетевых факторов.

5.2 Второй подход к обучению

Замеры работы моделей на полном тексте без удаления повторов, но с сохранением информации, из какого раздела пришла статья без нейросетевых фичей

Модель	Результат
Year model	0.258
$L_{log(1+exp(M))}$	0.236
$L_{(log(1+M^2)}$	0.281
$L_{exp(M)}$	0.2
$L_{frank-wolfe}$	0.237
L_{M^2}	0.288

Таблица 4: Результаты

5.3 Третий подход к обучению

Замеры работы модели при обучение по разделам, замеряем качество по разделам.

Модель	Результат
Year model	0.262
$L_{log(1+exp(M))}$	0.261
$L_{(log(1+M^2)}$	0.260
$L_{exp(M)}$	0.263
$L_{frank-wolfe}$	0.272
L_{M^2}	0.26

Таблица 5: Результаты

5.4 Сравнение трех подходов

Как видно из приведенных таблиц, самым оптимальным вариантом обучения является второй подход. При этом, если рассматривать композицию года и выхода модели, можно добиться результатов близкох к 0.31.

Если сравнивать честное качество моделей без дополнительных эвристик, лучшее качество получается при использование квадратичной функции как оптимизационной.

При этом так же хочется отметить, что добавление нейросетевых факторов не дало значимого прироста при текущих замерах.

6 Заключение

В данной работе были исследованы различные подходы к решению задачи построения модели ранжирования ссылок в системе полуавтоматического реферирования.

Были рассмотрены алгоритмы, основанные на графе цитирования, статистических характеристиках совместного распределения статей в корпусе документов, а также на моделях машинного обучения.

Анализ качества моделей, проведенный на выбранной коллекции, показал, что наилучшие результаты демонстрирует алгоритм, использующий градиентный бустинг над решающими деревьями для оптимизации специально подобранной функции потерь.

Список литературы

[1] Ke G. et al. Lightgbm: A highly efficient gradient boosting decision tree //Advances in neural information processing systems. – 2017. – Т. 30.

[2] Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – С. 785-794.

[3] Lo K. et al. S2ORC: The semantic scholar open research corpus //arXiv preprint arXiv:1911.02782. – 2019.

[4] Yu P. S., Li X., Liu B. On the temporal dimension of search //Proceedings of the 13th international World Wide Web conference on Alternate track papers posters. – 2004. – С. 448-449.

[5] Chen P. et al. Finding scientific gems with Google's PageRank algorithm //Journal of informetrics. – 2007. – Т. 1. – №. 1. – С. 8-15.

[6] Ding Y. Applying weighted PageRank to author citation networks //Journal of the American Society for Information Science and Technology. – 2011. – Т. 62. – №. 2. – С. 236-245.

[7] Xing W., Ghorbani A. Weighted pagerank algorithm //Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. – IEEE, 2004. – С. 305-314.

[8] Yan E., Ding Y. Discovering author impact: A PageRank perspective //Information processing management. – 2011. – Т. 47. – №. 1. – С. 125-134.

[9] Bird S. et al. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics //LREC. – 2008.

[10] Li H. A short introduction to learning to rank //IEICE TRANSACTIONS on Information and Systems. – 2011. – Т. 94. – №. 10. – С. 1854-1862.

[11] Burges C. J. C. From ranknet to lambdarank to lambdamart: An overview //Learning. – 2010. – Т. 11. – №. 23-581. – С. 81.