Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.85

Малашков
Валентин Борисович

ИССЛЕДОВАНИЕ И ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТЕКСТОВ НАУЧНЫХ ПУБЛИКАЦИЙ

АВТОРЕФЕРАТ
на соискание академической степени
магистратехнических наук

по специальности 1-40 80 05 –Программная инженерия

Научный руководитель
Шульдова С.Г.
к.т.н., доцент

Минск 2021

КРАТКОЕ ВВЕДЕНИЕ

По примерным подсчетам, объем научных публикаций только в электронном виде составляет более ста миллионов монографий, статей, материалов конференций, тезисов докладов и т.д. Ежегодно объём опубликованных работ увеличивается на миллион единиц.

На данном этапе, исследователям необходимы инструменты, которые автоматизируют и ускоряют поиск как ранее опубликованных научных работ, так и новых публикаций, помогающие быстро и эффективно вникать в суть отдельно взятого исследования или целой области.

Несмотря на достаточно детальное оформление любых научных публикаций, чтобы исследователю вникнуть в новую для себя область, ему придется пройти через огромное количество тех самых публикаций. Есть различные способы освоить новую область с помощью публикаций, которые как раз и содержат обзор новых и старых подходов в решении конкретной взятой задачи или области, но в тоже время — это отдельные публикации, которые необходимо отыскать, а также они все равно могут не содержать актуальной информации.

В настоящей работе рассмотренысуществующие решения для анализа текстов научных публикаций с целью реферирования и выявления научных направлений и тематик, рассмотрены возможности современных поисковых систем научных публикаций, а также данные, которые они могут предоставить. Рассмотрены доступные наборы данных для разработки алгоритмов, а также различные сервера, где можно в свободном доступе получить как полноценную публикацию, так и метаданные, ассоциированные с ней.

Затем на основе собранных данных рассмотрены возможности современных алгоритмов для кластеризации и детектирования сообществ в графе цитирования. Рассмотрены возможности современных алгоритмов для векторизации текста с получением метрики для сравнения текстов публикаций друг с другом, возможности комбинирования данных алгоритмов для детектирования сгруппированных научных публикаций по схожим тематикам. А также подходы и существующий алгоритмы для обобщения полученных тематик научных исследований.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Несмотря на достаточно детальное оформление любых научных публикаций, чтобы исследователю вникнуть в новую для себя область, ему придется пройти через огромное количество тех самых публикаций.

Исследователям необходимы инструменты, которые автоматизируют и ускоряют поиск как ранее опубликованных научных работ, так и новых публикаций, помогающие быстро и эффективно вникать в суть отдельно взятого исследования или целой области.Для этого необходимо понимание современных и актуальных научных направлений, а значит необходимо инструменты, которые бы позволяли автоматически извлекать эту информацию.

Цель и задачи исследования

Цель диссертационной работы – анализ, применение и оценкаэффективности методов и алгоритмовобработки текстов научных публикаций и их метаданных для выявления актуальных научных направлений, а также извлечение ключевой информации для их описания.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Анализ современных решений и алгоритмов для извлечения ключевой информации из научных публикаций.
2. Анализ современных решений и алгоритмов для кластеризации научных публикаций на основе информации, полученной из неструктурированного текста, а также метаданных.
3. Разработка модели для анализа научных публикаций и извлечение информации о научных направлениях.
4. Разработка архитектуры программного обеспечения для анализа текстов научных публикаций.

*Объектом*исследования выступают текст научных публикаций, а также метаданные о самой публикации.

Предметом исследования является извлечение информации о научных направлениях из текстов научных публикаций.

Основной *гипотезой*, положенной в основу диссертационной работы, является повышение эффективности алгоритмов кластеризации для выявления научных направлений за счеттекста самих научных публикаций.

Новизна полученных результатов

В результате проведённого исследования были получены результаты сравнения нескольких обобщенных подходов к формированию кластеров научных публикацию для выявления научных направлений в сочетании с текстовой информацией самих публикаций. Рассмотрены и проанализированы возможности оценки эффективности, полученных обобщенных методов исходя из доступных данных.

Положения, выносимые на защиту

1. Обоснование выбора набора данных для выполнения исследования.
2. Результаты анализа алгоритмов извлечения информации из текстов научных публикаций.
3. Обобщенные подходы к формированию кластеров научных публикаций для выявления актуальных научных направлений.
4. Результаты анализа применения подходовдля кластерного анализа научных публикаций для выявления актуальных научных направлений.

Апробация результатов диссертации

Материалы, положенные в основу работы, докладывались на одиннадцатой международной научно-методической конференции «Дистанционное обу-

чение - образовательная среда XXI века» (Минск, Беларусь, 2019), а также на научном семинаре молодых ученых «Актуальные проблемы науки XXI века» Минского инновационного университета (Минск, Беларусь, 2021).

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 печатных работы в сборниках материалов международных научных конференций. В сборнике материалов одиннадцатой международной научно-методической конференции «Дистанционное обучение - образовательная среда XXI века» и статья в рецензируемом сборнике научных статей молодых ученых перечня ВАК «Актуальные проблемы науки XXI века» Минского инновационного университета.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области: источники данных и их поставщики, примеры современных методов и готовых решений для анализа и реферирования научных публикаций и научных направлений. Вторая глава посвящена обзору и категоризации существующих алгоритмов в контексте анализа текстов научных публикаций и их метаданных, описание обобщенных подходов к решению поставленной задачи. В третьей главе представлен предварительный анализ данных, на которых апробирован разработанный алгоритм, описана разработка конечного программного обеспечения, которая реализует полностью весь процесс анализа данных, от его извлечения до отображения конечного результата, оценка эффективности алгоритма.Общий объем работы составляет 76 страниц, из которых основного текста – 58 страницы, включая 20 рисунков и 20 таблиц, список использованных источников из 43 наименований на 4 страницах и 3 приложений на 12 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе были подробно рассмотрены существующие решения анализа текстов научных публикаций для решения задачи автоматического обобщения как научных публикаций, так и научных направлений. Рассмотрены доступные источники данных, которые можно использовать в разработке конечного алгоритма, программного обеспечения, а также при оценке его эффективности.

В рамках второй главы были рассмотрены более подробно основные методы для векторного представления текстов научных публикаций. Начиная от простых моделей на основе статических методов, заканчивая современными моделями на основе нейронных сетей. Рассмотрены меры схожести применимые к сравнению векторов представления текстов научных публикаций, и их применимость в той или иной задаче. Рассмотрены методы для обнаружения сообществ в сетевых структурах. В конце предоставлены два обобщенных под-

хода к решению задачи анализа текстов научных публикаций с точки зрения выявления научных направлений.

В третьей главе проведен предварительный анализ данных, выбранный для исследования. Рассмотрены метрики для оценки кластерного анализа. Произведена оценка работы алгоритмов, сравнительный анализ и отобран выигрышный подходи алгоритмы.

ЗАКЛЮЧЕНИЕ

Были проанализированы существующие подходы к организации обработки и анализа текстов научных публикаций, а также смежных тематик. Были проанализированы различные источники данных, как готовые исследовательские наборы данных, так и открытые API-сервисы. Произведен сравнительный анализ и принято решение по выбору набора данных для разработки, отладки и оценки методов и алгоритмов для обработки и анализа текстов научных публикаций, а именно набор данных «DBLP» от исследовательской группы AMiner.

Далее были подробно описаны и проанализированы различные алгоритмы для векторного представления текстов естественного языка, структурные алгоритмы для обнаружения сообщества в сетевых графах, а именно лувенский метод, алгоритмы для векторного представления вершин графа, а именно node2vec, а также алгоритм k-среднихдля кластерного анализа. В целях демонстрации результатов работы кластерного анализа, был рассмотрен алгоритм «PositionRank»для отбора ключевых фраз из общего текста в рамках полученных кластеров научных публикаций.

Произведен предварительный анализ выбранного набора данных, а также его сэмплирование на основании графа цитирования. Произведен анализ и оценка работы двух основных подходов, а именно:

1. комбинирование лувенского метода и оценки схожести текстов через векторное представление;
2. комбинирование результатов векторного представления текстов научных публикаций и вершин графа цитирования, а затем кластерный анализ с помощью k-средних.

С помощью метрик ARIи AMI, был произведен сравнительный анализ двух подходов к решению поставленной задачи. Наилучший результат показал второй подход, а именно сочетание результатов node2veci word2vecс последующим кластерным анализом с использованием алгоритма k-средних.

Из наблюдений во время экспериментов с настройками алгоритмов и их сравнения, можно сделать вывод, что огромную часть контекстной информации о научной публикации несет в себе как раз ссылки на источники, а не сам текст, при этом комбинирование текстовой информации и информации о ссылках на источники, позволяет получить более качественные результаты.

Дальнейшие исследования можно развивать в сторону включения остальных признаков, таких как: соавторство, время публикации,наукометрические показатели, пересечение множества ссылок на источники и так далее.

Также можно развивать исследование в сторону применения современных архитектур глубоких нейронных сетей на базе архитектуры Transformerдля генерации осознанного реферата из группированных текстов научных публикаций в рамках одной тематики, для краткого изложения научного направления.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1.Малашков, В. Б. Основные подходы к оценке эффективности публикационной деятельности профессорско-преподавательского состава кафедр / Малашков В. Б., Шульдова С. Г., Лапицкая Н. В. // Актуальные проблемы науки XXI : сборник научных статей молодых ученых. – 2020. – №1(9). – С. 25–30.

2. Малашков, В. Б. Анализ текстовых отзывов пользователей дистанционных курсов на основе алгоритмов машинного обучения / Малашков В. Б., Шульдова С. Г. // Дистанционное обучение – образовательная среда XXI века : материалы XI Международной научно-методической конференции, Минск, 12-13 декабря 2019 г. / редкол. : В. А. Прытков [и др.]. – Минск : БГУИР, 2019. – С. 191.