**Министерство науки и высшего образования Российской Федерации**

Федеральное государственное автономное образовательное учреждение

высшего профессионального образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Наименование школы: ИШИТР

Наименование отделения: ОИТ

Наименование направления: Информационные технологии и

телекоммуникации

**Отчет по научно-исследовательской работе студента**

Выполнил: аспирант гр. А2-36 И.А. Рыбаченко

Проверил: к.т.н., доцент ОИТ ИШИТР А.О. Савельев

Томск – 2022

**Оглавление**

[Введение 3](#_Toc123004675)

[1. Научная информация 4](#_Toc123004676)

[1.1 Виды научных публикаций 4](#_Toc123004677)

[1.2 Структура статьи публикуемой в журнале 4](#_Toc123004678)

[1.3 Базы научного цитирования 5](#_Toc123004679)

[1.4 Ранжирование авторов научных публикаций 6](#_Toc123004680)

[1.5 Ранжирование научных журналов 7](#_Toc123004681)

[1.6 Инструменты систематизации изученных публикаций 7](#_Toc123004682)

[2. Поиск научной информации 8](#_Toc123004683)

[2.1 Методика поиска информации 8](#_Toc123004684)

[2.2 Поиск ключевых слов 8](#_Toc123004685)

[2.3 Определение релевантности статьи 9](#_Toc123004686)

[2.3 Извлечение структурированной информации 10](#_Toc123004687)

[2.4 Определение научного тренда 11](#_Toc123004688)

[3. Обзор предметной области поиска научной информации 12](#_Toc123004689)

[3.1 Методика проведения обзора 12](#_Toc123004690)

[3.2 Ключевые слова 12](#_Toc123004691)

[3.3 Анализ ключевых слов 13](#_Toc123004692)

[3.3.1 Ключевое слово citation database 13](#_Toc123004693)

[3.3.2 Ключевое слово information retrieval и literature review 15](#_Toc123004694)

[3.3.3 Ключевое слово summarization 15](#_Toc123004695)

[3.3.4 Ключевое слово text clustering 16](#_Toc123004696)

[3.4 Ключевые авторы 17](#_Toc123004697)

[Заключение 19](#_Toc123004698)

[Список литературы 20](#_Toc123004699)

Введение

С развитием технологий хранения данных и сети Интернет количество общедоступных научных публикаций значительно возросло и продолжает возрастать неубывающими темпами [1]. Разработка специальных алгоритмов работы с информацией может позволить ученым работать эффективнее, не тратя время на изучение некачественной или несоответствующей тематике информации и не упуская из виду значимые публикации.

Объект исследования. Процесс извлечения информации из слабоструктурированных источников и баз научных цитирований.

Предмет исследования. Модели, методы и алгоритмическое обеспечение процесса извлечения информации из слабоструктурированных источников и баз научных цитирований.

Цель работы. Изучение специфики предметной области поиска информации из слабоструктурированных источников и баз научных цитирований и определение тенденций её развития.

Задачи:

1. Составление карты ключевых слов, характеризующих предметную область;

2. Выделение наиболее значимых в предметной области публикаций;

3. Определение трендов развития предметной области.

Отчет состоит из трех глав. В первой главе описывается структура научных публикаций и баз цитирования. Во второй главе рассматриваются методики поиска научной информации. В третьей главе проводится системный обзор литературы с целью определения направлений развития поиска информации из слабоструктурированных источников и баз цитирования.

1. Научная информация

1.1 Виды научных публикаций

Определения актуальное состояние научной области можно при помощи анализа научных публикаций. К научным публикациям относятся следующие виды публикаций [2]:

1. Научная статья в журнале.

2. Научная статья в журнале, входящем в перечень ВАК.

3. Научная статья в сборнике научных статей.

4. Научная статья в сборнике трудов научной конференции.

5. Тезисы научного доклада или выступления.

6. Монография.

7. Раздел или глава в монографии.

8. Автореферат диссертации.

9. Научный отчет.

10. Рецензия в научном журнале на учебник, учебное пособие, учебно-методическое пособие, монографию.

11. Научные доклады, изданные отдельно или в научных сборниках.

12. Препринты.

Помимо научных публикаций для определения границ научной области также могут быть использованы патенты и заявки на регистрацию программы для электронной вычислительной машины. Данные документы отражают практическое применение научных знаний.

1.2 Структура статьи публикуемой в журнале

В Российской Федерации правила оформления статей, публикуемых в журнале или сборнике, регламентируются ГОСТ Р 7.0.7-2021 [3]. Нормативными ссылками ГОСТ определяет правила оформления заголовков и списка литературы. Ключевыми структурными элементами статьи являются заглавие, аннотация, ключевые слова, основная часть и библиографический список.

Основной текст статьи может быть структурирован и состоять из введения, текста статьи и заключения. В текст статьи могут быть выделены разделы (например, «Материалы и методы», «Результаты», «Обсуждение»).

Стоит отметить, что применение государственных стандартов является добровольным. Как правило журналы публикуют свои собственные памятки с требованиями к оформлению статей. Подтверждением тому могут послужить первые 5 журналов, взятые из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук (по состоянию на 01.02.2022 года) [4].

1.3 Базы научного цитирования

Информация о научных статьях агрегируется в базах цитирования. Существует большое количество баз цитирования. Одними из наиболее популярных являются:

1) Scopus

2) Web Of Science

3) PubMed

4) Google Scholar

5) Microsoft Academics

6) Elibrary

7) ACM

Базы цитирования различаются между собой количеством индексируемых журналов. Так, Google Scholar индексирует большое количество журналов, тогда как Web Of Science и Scopus ограничивают включаемые журналы по рейтингу [5]. Базы цитирования могут различаться тематикой. Например, база PubMed посвящена медицинской теме, а ACM компьютерной.

Scopus и Web Of Science обновляют свои базы мета-информации 1-2 раза в неделю [6]. Для подготовки мета-информации могут применяться автоматизированные системы, такие как Open Journal System. Иногда для извлечения мета-информации применяются программные скрипты [7]. Собранная мета-информация может быть неточная и научные базы могут упускать 5-10% цитирований [8].

Полнотекстовые статьи могут быть найдены на таких ресурсах как ScienceDirect, Arxiv, CyberLeninka или ResearchGate. ResearchGate помимо предоставления полнотекстовых статей предоставляют ученым возможность подписываться друг на друга и выстраивать коммуникацию. Особое место среди источников полнотекстовых публикаций занимает SciHub, не требующий для работы наличия платной подписки.

Информацию о связях между публикациями может быть представлена в графическом виде, что позволяет быстрее находить значимые статьи в интересующей области. Например, графическую информацию представляет ресурс ConnectedPapers.

1.4 Ранжирование авторов научных публикаций

Для оценки научного вклада автора может быть использована метрика Author Impact Factor (AIF) [9]. AIF вычисляется как отношение количества цитирований работ, опубликованных в год T к количеству работ, опубликованных за период dT, предшествующий году T.

H-индекс (индекс Хирша) вычисляется следующим образом. Учёный имеет индекс *h*, если *h* из его *Np* статей цитируются как минимум *h* раз каждая, в то время как оставшиеся (*Np — h*) статей цитируются не более чем *h* раз каждая.

G-индекс вычисляется следующим образом. Автор имеет рейтинг g, если g его наиболее цитируемых публикаций имеют не менее g2 цитирований.

1.5 Ранжирование научных журналов

Jourlnal Citation Report (JCR) вычисляется на основе данных Web Of Science и отражает частоту, с которой статьи журнала цитируются в научной литературе. SCImago Journal Rank (SJR) вычисляется также как и JCR, но на основе данных Scopus. В области компьютерной науки индексы JCR и SJR коррелируют между собой и могут использоваться взаимозаменяемо [10]. SJR обычно принимает большие значения, чем JCR.

Существуют также и другие метрики оценки журналов — Russian Science Citation Index (RSCI), Source Normalized Impact Per Paper (SNIP).

1.6 Инструменты систематизации изученных публикаций

Для хранения и структурирования изученных статей может быть использованы такие программы как Mendeley, 4author. Данные инструменты также позволяют создавать библиографические ссылки в нужном формате.

Для хранения заметок может быть использована программа Obsidian. Программа позволяет хранить заметки в формате markdown и устанавливать связи между заметками.

При создании заметок может быть использован метод zettelkasten. Данный метод подразумевает первоначальное накопление материала в неструктурированном виде и регулярные операции по структуризации накопленного материала.

2. Поиск научной информации

2.1 Методика поиска информации

Для поиска научной литературы может применяться следующий алгоритм [11]:

1) Сформулировать проблему

2) Выбрать базу данных для поиска информации

3) Сформулировать ключевые слова

3) Найти работы в базе данных по ключевым словам

4) Оценка результата и переход на шаг 1 при необходимости

В результате изучения литературы может быть составлен литературный обзор. Литературные обзоры по широте изучаемого вопроса могут быть разделены на системные. полусистемные и интеграционные [12]. Обзоры литературы могут быть найдены в базах цитирования по ключевому слову literature review. Как правило, в начале каждого обзора предоставляется методика сбора и анализа данных.

Смежной к поиску научной информации задачей является задача поиска информации в корпоративной базе знаний [13]. Спецификой корпоративного сегмента являются требование информационной безопасности о разграничении доступов к документам.

2.2 Поиск ключевых слов

Wei Lu в своей работе [14] исследует как авторы подбирают ключевые слова при написании статей. Выбор ключевых слов определяется тремя факторами - содержанием работы, знанием научной области и опытом изучения статей других авторов. Ключевые слова определенные содержанием работы часто встречаются в заголовке или аннотации. Ключевые слова, полученные от знания научной области, берутся из списка самых популярных в предметной области ключевых слов. Ключевые слова обусловленные опытом изучения статей других авторов часто встречаются в заголовке или аннотации цитируемой статьи. Более чем в половине случаев ключевые слова встречаются в заголовке или в аннотации. По появлению новых ключевых слов можно отследить эволюцию предметной области.

Для подбора ключевых слов к заданной тематике могут использоваться специальные словари ключевых слов. Отмечается, что использование таких словарей не очень эффективно, поскольку малознакомому с предметной областью ученому сложно ориентироваться в словаре [15].

Для поиска ключевых слов по заданному документу может быть использован метод TF-IDF [16]. Слова с наибольшим значением метрики TF-IDF потенциально могут являться ключевыми. Для оценки релевантности ключевых слов может быть использован метод Монте-Карло. Ключевые слова случайным образом группируются в поисковые запросы. Результаты запросов сравниваются между собой на пересечение.

Ключевые слова могут быть использованы для составления поискового запроса в базу цитирования. Об особенностях составления поисковых запросов написано в статье [17].

2.3 Определение релевантности статьи

Для ранжирования статей часто используется количество цитирований, может применяться алгоритм PageRank и Hyperlink-induced Topical Search (HITS) [18]. Фильтрация работ на основе связей показывает лучшие результаты чем на основе содержимого статьи. В качестве критериев качества используется требование, что список литературы должен включать популярные работы, обзоры литературы, новые работы и работы из смежных областей.

Свежие статьи не имеют цитирований и поэтому при фильтрации статей недостаточно опираться только на количество цитирований. Можно помимо количества цитирований смотреть на рейтинг журнала или автора. Также рейтинг статей может оцениваться на основе графа знаний, содержащего информацию как из самих статей, так и из их взаимосвязей [19].

Для отбора релевантной литературы может использоваться [20]:

1) ручной анализ и классификация источников

2) автоматический поиск по ключевым словам

3) автоматическое определение тематики статьи на основе пропорции слов, относящихся к заданной тематике

Ручной анализ релевантности источников подразумевает их прочитывание. Чтобы не читать всю статью целиком, можно сгенерировать краткое содержание статьи. Краткое содержание статьи может быть сгенерировано автоматически [21].

Для автоматической классификации статей на релевантные и нерелевантные может использоваться машинное обучение [22]. Алгоритмы машинного обучения могут сократить количество прочитываемых статей на 30% без существенной потери релевантных статей.

2.3 Извлечение структурированной информации

Существует задача извлечения структурированной информации по конкретной предметной области. Собранная информация представляется в виде группы ключ-значение или в виде графа связанных сущностей.

Для извлечения информации из новостных текстов определенной тематики может быть составлена онтологическую модель предметной области. Данный метод может быть применен в самых разных областях — энергетике [23], криминалистике [24][25], медицине [26]. Для выделения из текста онтологических терминов могут быть использованы шаблоны LSPL [27], семантико-синтаксический разбор текста [28] или классификаторы [29].

Информация, собранная из разных источников, может иметь разную структуру. Вопрос приведения информацию к единому виду рассматривается в статье [30].

Для визуализации связей между извлеченными объектами может использоваться программа PULLENTI [25].

2.4 Определение научного тренда

Для определения тренда может быть проанализирована динамика изменения ключевых слов [31]. Для анализа строится граф из ключевых слов, популярных в разрезе по годам.

Для отслеживания изменения тенденций в научной области может быть использовано извлечение сущностей на основе онтологии [32].

Для определения границ научной области может быть использована кластеризация публикаций. Для кластеризации может использоваться не только сам текст статьи, но и информация об авторах [33]. Также может быть использован метод анализа совместных цитирований [34].

3. Обзор предметной области поиска научной информации

3.1 Методика проведения обзора

Целью обзора является определение направлений развития области поиска информации в слабоструктурированных источниках и базах цитирования.

Для обзора было собрано 40 источников на заданную тематику и исходя из них определены ключевые слова. Список источников собирался итеративно через изучение цитирований прочитанных статей, оценку их релевантности и пополнение списка ключевых слов.

3.2 Ключевые слова

При анализе первичного набора источников наиболее значимыми показались следующие ключевые слова:

1) citation database. Находит статьи про то, как искать информацию в базах цитирования, про рейтинги статей и журналов, про сравнение баз цитирования.

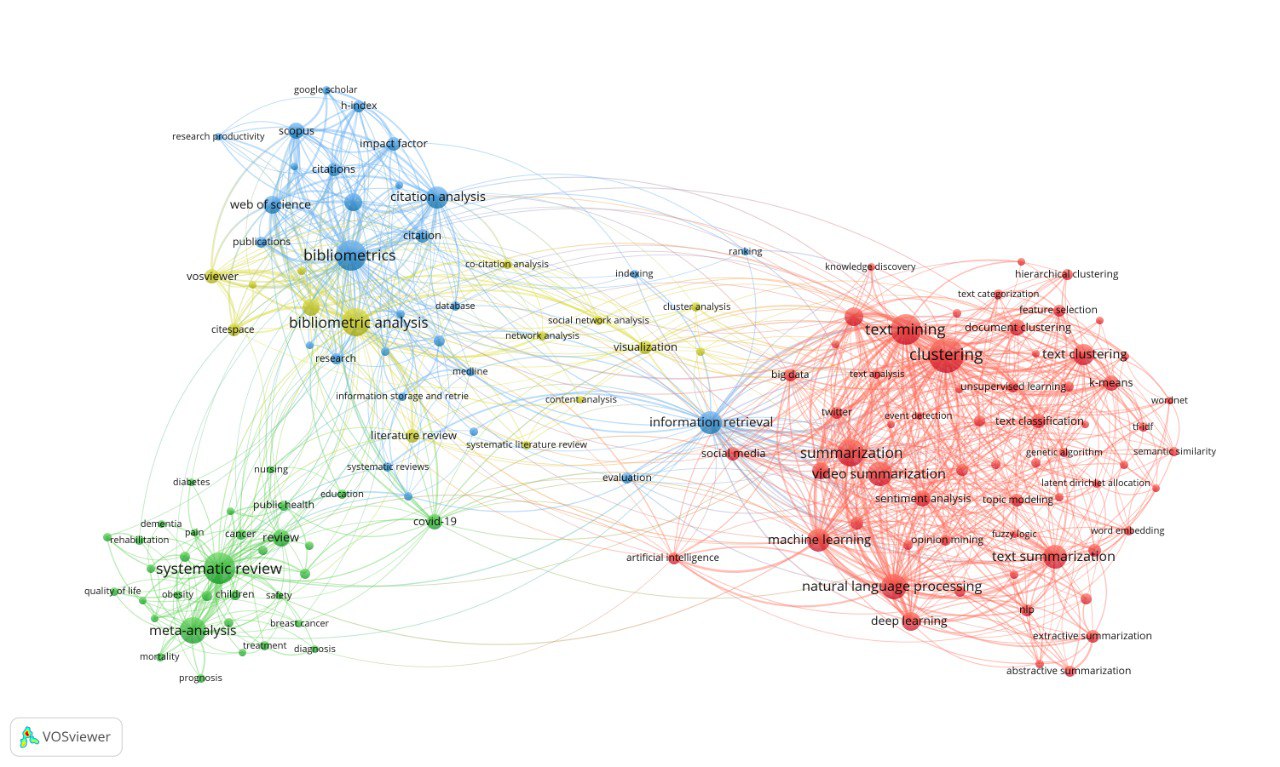
2) literature review / systematic literature review. Находит обзоры литературы. Как правило, в начале каждого обзора описывается используемая методика.

3) information extraction / information retrieval. Находит алгоритмы извлечения структурированной информации из слабоструктурированных источников.

4) text clustering. Находит алгоритмы кластеризации текстов.

5) text summarization. Находит алгоритмы создания сжатой версии длинного текста.

Построенная в VOSViewer карта ключевых слов показана на рисунке 1. Зеленый кластер — медицинские обзоры литературы. Синий и желтый кластеры — анализ баз цитирования. Красный кластер — обработка естественного языка.



1. Визуализация ключевых слов

Можно заметить, что красный кластер крупнее трех других. Это может говорить о том, что тема обработки естественного языка является широкой и может быть разбита по более узким ключевым словам.

3.3 Анализ ключевых слов

3.3.1 Ключевое слово citation database

Анализ ключевых слов 20000 работ, найденных по ключевому слову citation database показывает следующее:

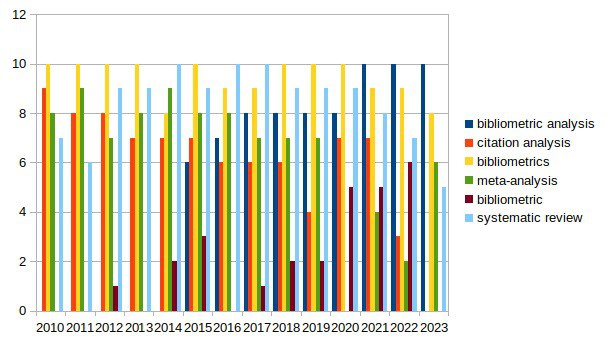
1) Ключевое слово scopus менее популярно чем web of science

2) В 2021-2022 годах был всплеск статей про vosviewer. Vosviewer это программа для графического отображения связей между статьями

3) Ключевые слова bibliometrics, bibliometric analysis, citation analyses, systematic review встречается в большинстве статей, содержащих ключевое слово citation database.

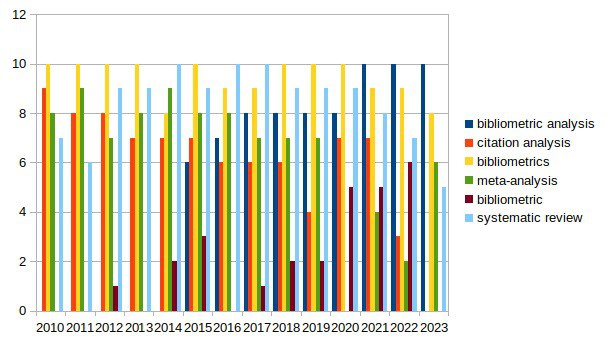
4) Ключевое слово impact factor в сочетании с citation database утратило свою популярность в 2015 году. При этом количество публикуемых статей с ключевым словом impact factor не сократилось, значит ключевое слово impact factor стал комбинироваться с другими ключевыми словами.

На рисунке 2 показана гистограмма ключевых слов, которые часто встречаются вместе с ключевым словом citation database. Высота столбца соответствует рангу ключевого слова. Ранг равен 10, если ключевое слово было самым популярным в пределах года. Ранг равен 0, если ключевое слово не попало в топ-10 ключевых слов за год.



1. Гистограмма ключевых слов, часто встречающихся вместе с ключевым словом citation database

На рисунке 3 показана гистограмма ключевых слов, которые изменили свою динамику, а также сравнение ключевых слов web of science и scopus. Высота столбца соответствует рангу ключевого слова. Ранг равен 10, если ключевое слово было самым популярным в пределах года. Ранг равен 0, если ключевое слово не попало в топ-10 ключевых слов за год.



1. Гистограмма ключевых слов, встречающихся вместе с ключевым словом citation database, которые изменили свою динамику

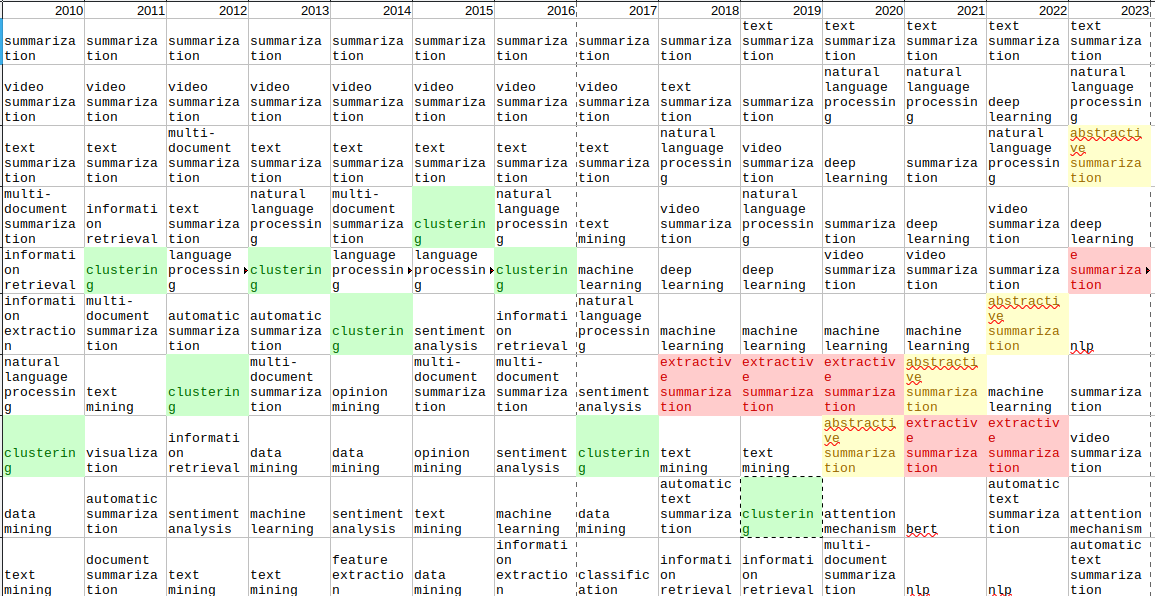
3.3.2 Ключевое слово information retrieval и literature review

Ключевое слово literature review используется часто — на Scopus нашлось 600.000 статей с этим ключевым словом. При этом количество статей, которые можно выгрузить одним запросом намного меньше и составляет 20.000. Поэтому при поиске публикаций ключевое слово literature review было сгруппировано с ключевым словом information retrieval.

Анализ ключевых слов, идущих вместе с information retrieval и literature review не показал интересной информации.

3.3.3 Ключевое слово summarization

На рисунке 4 показана таблица ключевых слов, встречающихся вместе с ключевым словом summarization. Чем выше ключевое слово в таблице, тем чаще оно встречается.



1. Таблица ключевых слов, встречающихся вместе с ключевым словом summarization

Из рисунка можно сделать следующие предположения:

1) Ключевое слово clustering ушло в пользу ключевых слов abstractive summarization (генерация нового, сжатого текста) и extractive summarization (сжатие текста путем выбора значимых фрагментов).

2) Ключевое слово multi-document clusterization имеет такую же динамику как и ключевое слово clustering.

3) Ключевое слово abstractive summarization имеет такую же динамику как и ключевое слово extractive summarization.

3.3.4 Ключевое слово text clustering

На рисунке 5 представлена диаграмма ключевых слов, встречающихся вместе с ключевым словом text clustering. Каждый столбец соответсвует определенному году (от 2010 до 2023). Чем выше позиция ключевого слова в строке, тем популярнее ключевое слово.



1. Таблица ключевых слов, встречающихся вместе с ключевым словом text clustering

Из карты ключевых слов можно сделать следующие предположения:

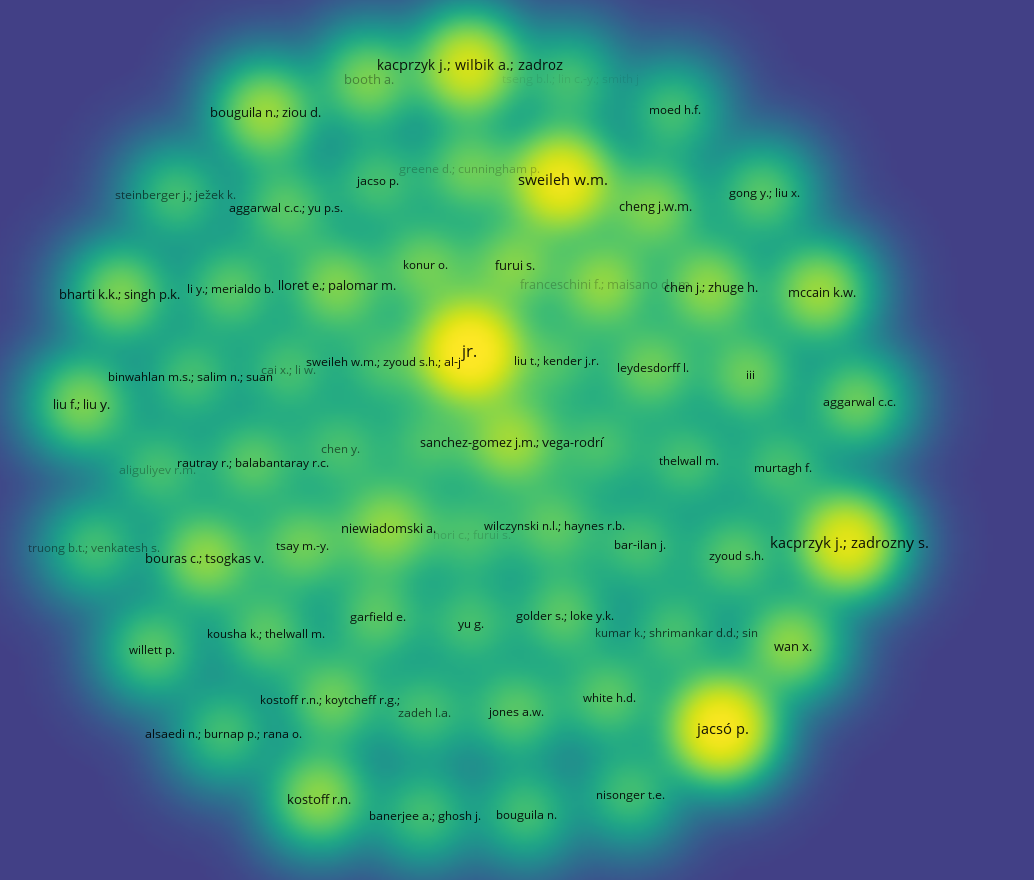
1) Алгоритм k-means сохраняет популярность

2) С 2019 года набирает популярность topic modeling

3) Совместно с ключевым словом text clustering часто используются ключевые слова text mining, document clustering

3.4 Ключевые авторы

При помощи VOSViever были визуализированы связи авторов статей, найденных по ключевым словам из пункта 3.2. Наиболее ярким цветом отмечены авторы с наибольшим количеством публикаций. Результат визуализации представлен на рисунке 6.



1. Тепловая карта рейтинга авторов по количеству публикаций

Наиболее ярко подсвечиваются следующие авторы:

1) jr. Вероятно это сокращение слова journal и отображается из-за того, что набор данных не был должным образом предобработан.

2) sweileh waleed. Писал литературные обзоры на медицинские темы

3) jacso peter. Проводил сравнительные анализы баз цитирования.

4) kacprzyk janusz. Кластеризация текстов

Заключение

В ходе работы был проведен обзор предметной области поиска научной информации. Определена структура научной публикации, описаны различия баз цитирования, приведены методы ранжирования авторов и журналов.

В ходе изучения литературы обнаружены следующие направления развития области поиска научной информации:

1) Развитие метрик оценки научной значимости публикаций

2) Развитие методов кластеризации научных публикаций

3) Развитие методов извлечения структурированной информации из слабоструктурированных источников

4) Развитие методов автоматического создания аннотаций к публикациям.

Список литературы

1. Jinha, A. E. Article 50 million: an estimate of the number of scholarly articles in existence [Electronic resource] / A. E. Jinha // Learned Publishing. — 2010. — Vol. 23, iss. 3. — P. 258-263. — Available from: <https://doi.org/10.1087/20100308>.

2. Сибирякова, Т. Б. НАУЧНАЯ ПУБЛИКАЦИЯ: ОСНОВНЫЕ ТРЕБОВАНИЯ И МЕТОДИКА ПОДГОТОВКИ / Т. Б. Сибирякова. — Волгоград : Индивидуальный предприниматель Ващенко Александр Николаевич, 2017. — 80 c. — ISBN 978-5-600-01785-6.

3. ГОСТ Р 7.0.7-2021. СИСТЕМА СТАНДАРТОВ ПО ИНФОРМАЦИИ, БИБЛИОТЕЧНОМУ И ИЗДАТЕЛЬСКОМУ ДЕЛУ. СТАТЬИ В ЖУРНАЛАХ И СБОРНИКАХ. ИЗДАТЕЛЬСКОЕ ОФОРМЛЕНИЕ : дата введения 01.10.2021. – URL: https://protect.gost.ru/v.aspx?control=8&baseC=6&page=1&month=9&year=2021&search=&RegNum=1&DocOnPageCount=15&id=230783 дата обращения: 24.12.2022). – Текст: электронный.

4.  Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук (по состоянию на 01.02.2022 года). – URL: https://vak.minobrnauki.gov.ru/uploader/loader?type=19&name=91107547002&f=11305 дата обращения: 24.12.2022). – Текст: электронный.

5. What have we learned from Environmental Kuznets Curve hypothesis? A citation-based systematic literature review and content analysis [Electronic resource] / A. Naveed [et al.] // Energy Strategy Reviews. — 2022. — Vol. 44. — P. 100946. — Available from: <https://doi.org/10.1016/j.esr.2022.100946>.

6. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses [Electronic resource] / M. E. Falagas [et al.] // The FASEB Journal. — 2007. — Vol. 22, iss. 2. — P. 338-342. — Available from: <https://doi.org/10.1096/fj.07-9492lsf>.

7. Методы автоматизированного извлечения метаданных научных публикаций для библиографических и реферативных баз цитирования / А. Н. Герасимов, А. М. Елизаров, Е. К. Липачев, Ш. М. Хайдаров // Информационное общество: образование, наука, культура и технологии будущего : сборник научных статей. Труды XIX Международной объединенной научной конференции "Интернет и современное общество", Санкт-Петербург, 22–24 июня 2016 года. – Санкт-Петербург: Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, 2016. – С. 41-48. – EDN UWOLXF.

8. Franceschini, F. Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? [Electronic resource] / F. Franceschini, D. Maisano, L. Mastrogiacomo // Journal of Informetrics. — 2014. — Vol. 8, iss. 3. — P. 751-765. — Available from: <https://doi.org/10.1016/j.joi.2014.07.003>.

9. Pan, R. K. Author Impact Factor: tracking the dynamics of individual scientific impact [Electronic resource] / R. K. Pan, S. Fortunato // Scientific Reports. — 2014. — Vol. 4, iss. 1. — Available from: <https://doi.org/10.1038/srep04880>.

10. Sicilia, M. Comparing impact factors from two different citation databases: The case of Computer Science [Electronic resource] / M. Sicilia, S. Sánchez-Alonso, E. García-Barriocanal // Journal of Informetrics. — 2011. — Vol. 5, iss. 4. — P. 698-704. — Available from: <https://doi.org/10.1016/j.joi.2011.01.007>.

11. McGrath, J. M. Before You Search the Literature: How to Prepare and Get the Most Out of Citation Databases [Electronic resource] / J. M. McGrath, R. E. Brown, H. A. Samra // Newborn and Infant Nursing Reviews. — 2012. — Vol. 12, iss. 3. — P. 162-170. — Available from: <https://doi.org/10.1053/j.nainr.2012.06.003>.

12. Snyder, H. Literature review as a research methodology: An overview and guidelines [Electronic resource] / H. Snyder // Journal of Business Research. — 2019. — Vol. 104. — P. 333-339. — Available from: <https://doi.org/10.1016/j.jbusres.2019.07.039>.

13. Патент № 2729224 C2 Российская Федерация, МПК G06F 16/338, G06F 16/31, G06F 16/93. Способ извлечения информации и корпоративная система поиска информации : № 2019100812 : заявл. 10.01.2019 : опубл. 05.08.2020 / М. С. Багаев, О. Л. Багаева ; заявитель Общество с ограниченной ответственностью "МАКСИОЛ". – EDN AIZXZE.

14. How do authors select keywords? A preliminary study of author keyword selection behavior [Electronic resource] / W. Lu [et al.] // Journal of Informetrics. — 2020. — Vol. 14, iss. 4. — P. 101066. — Available from: <https://doi.org/10.1016/j.joi.2020.101066>.

15. Successful Keyword Searching: Initiating Research on Popular Topics Using Electronic Databases [Electronic resource] // The Journal of Academic Librarianship. — 2001. — Vol. 27, iss. 5. — P. 420. — Available from: <https://doi.org/10.1016/s0099-1333(01)00252-x>.

16. Information retrieval from scientific abstract and citation databases: A query-by-documents approach based on Monte-Carlo sampling [Electronic resource] / F. Lechtenberg [et al.] // Expert Systems with Applications. — 2022. — Vol. 199. — P. 116967. — Available from: <https://doi.org/10.1016/j.eswa.2022.116967>.

17. Ridley, D. D. Citation searches in on-line databases: possibilities and pitfalls [Electronic resource] / D. D. Ridley // TrAC Trends in Analytical Chemistry. — 2001. — Vol. 20, iss. 1. — P. 1-10. — Available from: <https://doi.org/10.1016/s0167-2940(01)90065-2>.

18. Sesagiri, A. R. Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems [Electronic resource] / A. R. Sesagiri, S. Foo, N. Pang // Information Processing & Management. — 2017. — Vol. 53, iss. 3. — P. 577-594. — Available from: <https://doi.org/10.1016/j.ipm.2016.12.006>.

19. Assessing Scientific Research Papers with Knowledge Graphs [Electronic resource] / K. Sun [et al.] // Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. — 2022. — Available from: <https://doi.org/10.1145/3477495.3531879>.

20. Separating the wheat from the chaff: A topic and keyword-based procedure for identifying research-relevant text\*✰ [Electronic resource] / A. Eads [et al.] // Poetics. — 2021. — Vol. 86. — P. 101527. — Available from: <https://doi.org/10.1016/j.poetic.2020.101527>.

21. Automatic text summarization: A comprehensive survey [Electronic resource] / W. S. El-Kassas [et al.] // Expert Systems with Applications. — 2021. — Vol. 165. — P. 113679. — Available from: <https://doi.org/10.1016/j.eswa.2020.113679>.

22. Automatic classification of literature in systematic reviews on food safety using machine learning [Electronic resource] / L. d. van [et al.] // Current Research in Food Science. — 2022. — Vol. 5. — P. 84-95. — Available from: <https://doi.org/10.1016/j.crfs.2021.12.010>.

23. Веремьянина, А. О. Построение лексико-семантических ресурсов для обеспечения извлечения информации из новостных текстов по энергетике / А. О. Веремьянина // Материалы 54-й Международной научной студенческой конференции МНСК-2016: Прикладная лингвистика, Новосибирск, 16–20 апреля 2016 года. – Новосибирск: Новосибирский национальный исследовательский государственный университет, 2016. – С. 17-18. – EDN XSKFSF.

24. Крутиков, Н. О. Разработка системы извлечения информации изтекстов на русском языке в областикриминалистики / Н. О. Крутиков, Н. Подаков, В. А. Жилякова // Проблемы информатики. – 2016. – № 3(32). – С. 70-84. – EDN XBSQOH.

25. Система PullEnti - извлечение информации из текстов естественного языка и автоматизированное построение информационных систем / О. В. Золотарев, М. М. Шарнин, С. В. Клименко, К. И. Кузнецов // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности (SCVRT2015-16) : Труды Международной научной конференции: в 2-х томах, ЦарьГрад, Московская область, Россия, 21–24 ноября 2016 года. – ЦарьГрад, Московская область, Россия: Автономная некоммерческая организация "Институт физико-технической информатики", 2016. – С. 28-35. – EDN XSUFET.

26. Свидетельство о государственной регистрации программы для ЭВМ № 2020616818 Российская Федерация. Программа для извлечения данных из неструктурированной информации Webiomed.NLP : № 2020615953 : заявл. 17.06.2020 : опубл. 23.06.2020 ; заявитель Общество с ограниченной ответственностью «К-Скай». – EDN DKJQDF.

27. Большакова Е. И., Носков А. А. Анализ текста на основе лексико-синтаксичеких шаблонов с сокращением многовариантности // Новые информационные технологии в автоматизированных системах. 2010. №13. URL: https://cyberleninka.ru/article/n/analiz-teksta-na-osnove-leksiko-sintaksichekih-shablonov-s-sokrascheniem-mnogovariantnosti (дата обращения: 26.12.2022).

28. Патент № 2646386 C1 Российская Федерация, МПК G06F 17/27. Извлечение информации с использованием альтернативных вариантов семантико-синтаксического разбора : № 2016147965 : заявл. 07.12.2016 : опубл. 02.03.2018 / С. Е. Мацкевич ; заявитель Общество с ограниченной ответственностью "Аби Продакшн". – EDN KRIANL.

29. Патент № 2691855 C1 Российская Федерация, МПК G06K 9/66, G06F 17/27, G06F 17/28. Обучение классификаторов, используемых для извлечения информации из текстов на естественном языке : № 2018110387 : заявл. 23.03.2018 : опубл. 18.06.2019 / С. Е. Мацкевич, И. А. Булгаков ; заявитель Общество с ограниченной ответственностью "Аби Продакшн". – EDN TVVYQH.

30. Брюхов, Д. О. звлечение информации из разноструктурированных данных и её приведение к целевой схеме / Д. О. Брюхов // Аналитика и управление данными в областях с интенсивным использованием данных: XVII Международная конференция DAMDID/RCDL'2015. Обнинск, 13-16 октября 2015 года. Труды конференции / НИЯИ МИФИ. - 2015. - Обнинск: ИАТЭ НИЯУ МИФИ. - С.81-90. - ISBN: 978-5-9530-0398.

31. Jeon, J. W. SNA Approach for Analyzing the Research Trend of International Port Competition [Electronic resource] / J. W. Jeon, Y. Wang, G. T. Yeo // The Asian Journal of Shipping and Logistics. — 2016. — Vol. 32, iss. 3. — P. 165-172. — Available from: <https://doi.org/10.1016/j.ajsl.2016.09.005>.

32. Манучарян, Л. А. Применение систем извлечения информации в наукоемких индустриях / Л. А. Манучарян // Экономика и социум. – 2011. – № 1(1). – С. 197-199. – EDN WBFYXJ.

33. Tang, X. Contrastive author-aware text clustering [Electronic resource] / X. Tang, C. Dong, W. Zhang // Pattern Recognition. — 2022. — Vol. 130. — P. 108787. — Available from: <https://doi.org/10.1016/j.patcog.2022.108787>.

34. He, Y. Mining a Web Citation Database for author co-citation analysis [Electronic resource] / Y. He, S. H. Cheung // Information Processing & Management. — 2002. — Vol. 38, iss. 4. — P. 491-508. — Available from: https://doi.org/10.1016/s0306-4573(01)00046-2.