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

УДК 004.65

На правах рукописи

МОИСЕЕНКО
Александр Александрович

ИНСТРУМЕНТАЛЬНОЕ СРЕДСТВО ФОРМИРОВАНИЯ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ

АВТОРЕФЕРАТ
диссертации на соискание степени
магистра экономических наук

по специальности 1-25 80 08 – Математические
и инструментальные методы экономики

Минск 2019

Работа выполнена на кафедре экономической информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **ЖИВИЦКАЯ Елена Николаевна,**
проректор по учебной работе «Белорусского государственного университета информатики и радиоэлектроники», кандидат технических наук, доцент, MBA

Рецензент: **СИНЯВСКАЯ Ольга Александровна,**
кандидат экономических наук, доцент кафедры промышленного маркетинга и коммуникаций учреждения образования «Белорусский государственный Экономический университет»

Защита диссертации состоится «26» июня 2019 г. года в $10^{\underline{30}}$ часов на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. Платонова 39, корп. 5, ауд. 806, тел. 293-89-92, e-mail: kafei@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

До начала XX века научные исследования проводились лишь небольшим кругом людей, а существенная оценка вклада проводилась только обществом основываясь на содержании данной работы.

Однако это изменилось и современный мир сложно представить без различных научных журналов, статей и даже книг. Кроме того, что сегодня научные исследования являются неотъемлемой частью жизни общества, изменились и подходы к оценке научных работ. Исходя из необходимости оценки науки и деятельности научных сотрудников в середине XX века была сформирована новая отрасль под названием «наука о науке» или же наукометрия. Основная задача, возлагаема на данную область – качественная и количественная оценка научных исследований, а также изучение эволюции науки через многочисленные измерения и статистическую обработку информации.

Но, если в XX веке еще не было возможности широкого использования вычислительных ресурсов и операции по оценки приходилось производить вручную, то в XXI веке в связи со стремительным развитием вычислительной техники и сети интернет на помощь в оценке научных работ пришли разработанные программные продукты называемые наукометрическими базами данных.

Наукометрические базы данных помогают пользователям и научным сотрудникам автоматизировать процесс качественной и количественной оценки научных работ, путем использования электронных вычислительных ресурсов, а также специализированного программного обеспечения. Данные базы предоставляют различную информацию и статистику, которая может использоваться для различных целей.

Например, одним из ключевых показателей для научных сотрудников является индекс научного цитирования, который показывает «значимость» его работы и может использоваться для получения грантов на дальнейшие исследования. На этот показатель опираются в Чешской Республики при принятии решения о выдаче грантов.

Кроме того, наукометрические базы данных предоставляют различную информацию об исследованиях по категориям, ее значимость, ипакт-фактор, которая помогает в оценке развития той или иной области науки, оценки значимости определенного журнала.

Цель данной работы – изучение наукометрических баз данных и повышение надежности при их формировании.

Задачи исследования:

– изучить основные методы формирования наукометрических баз данных и возникающие проблемы;

– провести анализ процесса формирования базы на примере Web of Science от Clarivate Analytics;

–модифицировать методы формирования базы для улучшения качества предоставляемой информации.

Объект исследования – наукометрическая база данных Web of Science от Clarivate Analytics.

Предмет исследования – методы и алгоритмы формирования наукометрических баз данных.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Сегодня наукометрия является самостоятельной областью, которая занимается качественной и количественной оценкой научных исследований и необходима для определения качества исследования и оценки научных изданий.

Из-за большого роста научных работа требуется производить постоянный мониторинг и анализ этих работ и вручную проделать данную процедуру уже невозможно. Для этого по всему миру используются известные базы данных, которые помогают решить следующие задачи:

– оценка существующих направлений развития науки и технологий;

– оценка конкурентоспособности в определенной области страны, организации или же отдельного человека;

– оценка предлагаемых учеными проектов;

– оценка результатов деятельности организации;

– выявление и оценка эффективности сотрудничества между организациями и учеными;

– формирование научных групп для проведения исследования/проекта;

– составление рейтингов организаций и авторов;

– участие в международных рейтингах;

– оценка обоснованности и определение размеров финансирования определенных научных исследований

В связи с этим наукометрические базы данных пользуются большим спросом, однако со своей стороны они должны предоставлять корректную и точную информацию по научной информации. Соответственно на эти базы накладывается большая ответственность за корректное связывание информации и ее предоставление.

Степень разработанности проблемы

Наукометрические базы данных, а также их информация были исследованы различными научными сотрудниками из США, Польши, Украины и т.д. Однако, информация по формированию таких типов баз данных не является для

них первоочередной, а проблемы, возникающие при их работе связаны с другими областями и дисциплинами, такие как математика и информационные технологии.

Среди ученых, занимавшихся наукометрией, можно выделить российских ученых Налимова В.В. и Мульченко З.М. предлагавших несколько моделей, направленных на всестороннее изучение процесса развития науки, уделяя особое внимание информационной модели, которая рассматривает науку как самоорганизующуюся систему, управляющаяся своими информационными потоками. При изучении науки как информационного процесса, оказывается возможным применять количественные или статистические методы исследования. В 1969 г. в совместной монографии В.В. Налимов вводит в научный оборот термин «наукометрия» – «Будем называть наукометрией количественные методы изучения развития науки как информационного процесса». Стоит отметить, что этот термин он уже упоминал в 1966 году в своей статье «Количественные методы исследования процесса развития науки». Положительно оценивая вклад В.В. Налимова в становление наукометрии, следует отметить и негативную роль приведенного определения этого термина, поскольку оно сориентировало дальнейшие исследования в этой области на «нумерологический» путь развития.

Кроме российский ученых следует упомянуть украинского ученого Геннадия Доброва (1929- 1989 годах), посвятившего разработке данного круга проблем всю свою творческую жизнь. Опубликованная в 1966 году в Киеве его фундаментальная монография «Наука о науке: Введение в общее науковедение», которая фактически положила начало этому направлению работ, углубила интерес к науковедческим исследованиям и была переведена на многие языки мира. В ней он акцентировал внимание на необходимости систематизированного исследования тенденций и перспектив развития науки в мире. Это отражается в широком спектре вопросов, которые рассматривались: история развития науки и научных школ, состояние и тенденции развития научно-технологического потенциала, инфраструктура науки, научно-технологическая и инновационная политика, вопросы международного сотрудничества. Данное Г.М. Добровым определение науковедения «... это комплексное исследование и теоретическое обобщение опыта функционирования социальных систем в науке с целью обоснования научно-технической политики, а также рационального формирования потенциала науки и повышения эффективности научной деятельности при помощи средств социального, экономического и организационного воздействия» и сегодня актуально. Оно отражает системность науковедческих исследований и необходимость получения комплексных знаний о науке.

В качестве исключений можно привести работы А.А. Коренного, И.В. Маршаковой, С.Д. Хайтуна. В них отмечалось на первоочередность решения задач организации системы прогнозирования научных исследований, использо-

вание библиометрических показателей для определения структуры науки и отслеживания ее развития, а также на недостатки применения только количественных показателей при оценивании результативности научных исследований. Следует подчеркнуть, что И.В. Маршакова и С.Д. Хайтун считали определения наукометрии, данное В.В. Налимовым, «чересчур категоричным».

В практическом аспекте наибольший вклад в наукометрические исследования был сделан Ю. Гарфилдом. Он предложил уникальную идею по использованию научных ссылок как средства информационного поиска и изучения структуры науки. С его именем связано организация Института научной информации США и создание базы данных Web of Science с аналитическими надстройками. В тоже время сам Ю. Гарфилд неустанно призывал к осторожности в использовании данных цитирования, отмечая, что они, как и «любой инструмент – от ядерной энергии до молотка – должны быть правильно использованы».

Цель и задачи исследования

Цель данной работы –изучение наукометрических баз данных и повышение надежности при их формировании.

В работе поставлены и решены следующие задачи:

– изучить основные методы формирования наукометрических баз данных и возникающие проблемы;

– провести анализ процесса формирования базы на примере Web of Science от Clarivate Analytics;

– модифицировать методы формирования базы для улучшения качества предоставляемой информации.

Область исследования

Содержание диссертационной работы соответствует образовательному стандарту высшего образования второй ступени (магистратуры) ОСВО 1-25 80 08-2012 специальности 1-25 80 08 «Математические и инструментальные методы экономики».

Теоретическая и методологическая основа исследования

В основу диссертации легли исследования зарубежных и отечественных ученых в области баз данных, методы создания и управления высоконагруженными базами данных. При решении поставленных задач использована теории принятия решений, методы оптимизации, методы объектно-ориентированного программирования.

В качестве инструментальных средств применялись объектно-ориентированный язык программирования Java 1.8, Hibernate ORM, HikariCP и Spring Framework, Angular Framework.

Информационная база исследования сформирована на основе данных, опубликованных в журналах, а также информация предоставленная Web of Science и открытая информация из Scopus, Google Scholar.

Научная новизна

Научная новизна в работе предложены технологии и методы решения задач, возникающих при использовании наукометрических баз данных. Данные методы отличаются от остальных тем, что они основаны на реальных данных, которые используют существующие наукометрические базы данных, а также используют новые технологии, такие как G1 Garbage Collector и String De-Duplication.

Теоретическая значимость предложена гибкая логическая модель формирования хэш-сумм, обеспечивающая возможность добавления новой функциональности в рамках существующей модели. Представлены методы оптимизации текущих систем, путем оптимизации JVM.

Практическая значимость предложенные модели и методы могут использоваться при построении новых наукометрических баз данных, улучшение существующих, а также создания баз данных со смежным функционалом.

Основные положения, выносимые на защиту

1. Понятие наукометрии и наукометрических баз данных
2. Апробация модифицированных алгоритмов.
3.Модифицированное веб-приложения для управления кластерами в наукометрической базе данных.

Апробация диссертации и информация об использовании ее результатов

Результаты исследований, вошедшие в диссертацию, а также доработанный инструмент были использованы в одном из подразделений компании Clarivate Analytics, которая занимается разработкой и поддержкой наукометрической базы данных Web of Science. Результаты апробации показали, что доработанный программный продукт позволил снизить количество обращений клиентов связанных с предоставление некорректных данных, а также снизить нагрузку на существующий программный продукт.

Публикации

Изложенные в диссертации основные положения и выводы опубликованы в 2 печатных работах, представленные в виде двух статей в научных журналах.

Общий объем публикаций по теме диссертации составляет 10 страниц.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, трех глав с краткими выводами по каждой главе, заключения, библиографического списка и приложений.

В первой главе рассматриваются основные подходы реализации наукометрических баз данных, возможности, которые предоставляет каждая из них, а также достоинства и недостатки этих решений. Кроме этого, анализируются поступающие данные, а также их использование при расчете научных показателей, таких как ипакт-фактор, индекс цитирования.

Во второй главе рассматриваются методы оптимизации текущего решения на основе оптимизации формирования хэш-сумм для кластеров с учетом уже существующей системы, а также данных, которые в ней используются.

В третьей главе предложенные методики апробированы на фактическом материале; проведены численные эксперименты, использующие результаты проведенных исследований по разработке наукометрической базы данных.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность и практическая ценность темы исследования, описаны решаемые проблемы и цели исследования. Приведен свойства, которыми должны обладать современные наукометрические базы данных.

В первой главе раскрыты основные понятия и механизмы наукометрических баз данных, обусловленные постоянным ростом обрабатываемых данных. Соответственно, непрерывный и постоянный рост требует доработки программного обеспечения.

Наукометрия – область науковедения, занимающаяся исследования науки количественными методами, статистическим анализом структуры и динамикой научной информации. Она сформировалась в середине XX века в самостоятельную отрасль научного знания, предметом изучения которой стала сама наука и результаты деятельности научных работников. Эта дисциплина получила название «наука о науке», которое утвердилось в несколько ином звучании – науковедение.

Среди ученых, занимавшихся наукометрией, можно выделить российских ученых Налимова В.В. и Мульченко З.М. предлагавших несколько моделей, направленных на всестороннее изучение процесса развития науки, уделяя особое внимание информационной модели, которая рассматривает науку как самоорганизующуюся систему, управляющаяся своими информационными потоками. При изучении науки как информационного процесса, оказывается возможным применять количественные или статистические методы исследования. В 1969 г. в совместной монографии В.В. Налимов вводит в научный оборот термин «наукометрия» – «Будем называть наукометрией количественные методы изучения развития науки как информационного процесса» [7]. Стоит отметить, что этот термин он уже упоминал в 1966 г. в своей статье «Количественные методы исследования процесса развития науки» [6]. Положительно оценивая вклад В.В. Налимова в становление наукометрии, следует отметить и негатив-

ную роль приведенного определения этого термина, поскольку оно сориентировало дальнейшие исследования в этой области на «нумерологический» путь развития.

Наукометрическая база данных – библиографическая и реферативная база данных, которая служит инструментом для отслеживания цитирования научных публикаций, позволяющая производить поиск и получение статистической информации, характеризующей текущее состояние и динамику показателей востребованности, активности и индексов влияния деятельности отдельных научных сотрудников и организаций.

Данные базы помогают избавиться от таких рутинных задач, как сопоставление статьи и цитат, расчет индекса цитирования, импакт-фатора и других показателей, поиск необходимой информации.

На данный момент существуют наукометрические различные базы данных, которые используются для различных целей. Например, база PubMed используется при публикации англоязычных работ, связанных с медициной и биологией.

Но, наиболее популярными и известными базами являются: Web of Science, Scopus, РИНЦ, Google Schoolar.

Все вышеупомянутые базы имееют схожий функционал, однако есть и различия, которые могут повлиять на выбор той или иной базы. Наиболее значимым является поиск аналитической и статистической информации (таблица 1.1).

Таблица 1.1 - Поиск информации по базам данных

Данные	Web of Science	Scopus	РИНЦ	Google Scholar
Собственные	+	+	+	+
По институту	-	+	-	-
По другому автору	+	+	+/- (если автор зарегистрирован в системе)	+/- (если аккаунт создан и доступ к нему открыт)

Для получения аналитической и статистической информации наукометрические базы данных включают большое количество информации, которая помогает произвести оценку исследования. Чаще всего индекс таких баз состоит из следующих данных:

– автор;
– название;
– место публикации;
– год публикации;
– том;

−номер;

−страницы;

−ЦИО (DOI);

−адресная информация, включающая в себя место работы автора/авторов, организацию, город, страну;

−ссылки на использованные источники;

Исходя из этой данных формируется статистическая информация, помогающая получить анализ научного исследования и понять значимость научного исследования.

Для авторов статей наиболее значимой информацией является индекс цитирования.

Индекс цитирования научных статей (ИЦ)— реферативная база данных научных публикаций, индексирующая ссылки, указанные в пристатейных списках этих публикаций и предоставляющая количественные показатели этих ссылок (такие как суммарный объём цитирования, индекс Хирша и др.)

Также немаловажным фактором является импакт-фактор, который может повлиять на оценку научной работы.

Импакт-фактор (ИФ) - численный показатель важности научного журнала. С 1960-х годов он ежегодно рассчитывается Институтом научной информации, который в 1992 году был приобретён корпорацией Thomson и ныне называется Thomson Scientific, и публикуется в журнале «Journal Citation Report». В соответствии с ИФ оценивают уровень журналов, качество статей, опубликованных в них, дают финансовую поддержку исследователям и принимают сотрудников на работу. Импакт-фактор имеет хотя и большое, но неоднозначно оцениваемое влияние на оценку результатов научных исследований.

Для расчета импакт-фактора берется информация о числе цитирований и количестве публикаций за определенный промежуток времени. Например, для расчета импакт-фактора журнала на 2018 год берется количество цитат за 2017-2018 года и делится на количество публикаций за этот же промежуток времени.

ИФ журнала зависит от области исследований и его типа; из года в год он может заметно меняться, например, опускаясь до предельно низких значений при изменении названия журнала и так далее. Тем не менее, на сегодня ИФ является одним из важных критериев, по которому можно сопоставлять уровень научных исследований в близких областях знаний. Например, инвестор научного исследования может захотеть сравнить результаты исследователей для оценки перспектив своих инвестиций. Для этого и используются объективные численные показатели, такие как импакт-фактор. Поэтому на подобные измерения и существует спрос.

Исходя из этих показателей можно понять, что для корректного анализа информации требуется высокое качество вносимой информации, а также оптимизированные алгоритмы для связывания статей с цитатами. Некорректная информация может повлиять на расчет значимых показателей.

Во второй главе рассмотрены существующие методы формирования кластеров в наукометрических базах данных, а также проанализирована текущая инфраструктура на предмет возможностей оптимизации ресурсов основываясь на действующей системе Web of Science от Clarivate Analytics.

Введено понятие кластера в наукометрической базе Web of Science, а также проведено описание наиболее популярных хеш-сумм.

Для анализа использовалась полная выборка по существующим кластерам, а также выборка проблемных кластеров начиная с 1900 года.

Исходя из этой информации было представлены методы оптимизации формирования кластеров путем оптимизации хэш-функций. Для этого использовалась выборка проблемных статей и цитат, с помощью которых можно было явно идентифицировать проблему, а также найти способы ее решения. Кроме этого, были предложены использования нового функционала JVM, который позволил уменьшить время задержки, а так же сократить потребление памяти до 20%.

В третей главе описаны инструменты, которые использовались при модификации базы данных Web of Science, а также базовая модель, необходимая для работы этой системы. Кроме того, были представлены результаты работы приложения после модификаций, а также сравнение с первоначальными данными.

На рисунке 1 представлена базовая модель системы Web of Science.

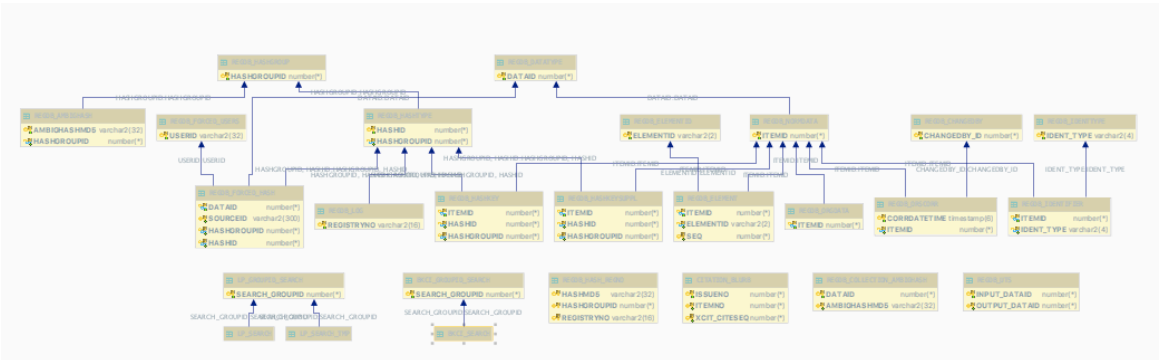

Рисунок 1 – Модель базы Web Of Science

.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведен анализ существующих наукометрических баз данных. Сформулированы требования к разрабатываемой системе.

2. Предложен подход к оптимизации работы системы и инфраструктуры с учетом увеличивающегося объема данных и высокой нагрузки.

3. Предложен ряд алгоритмов формирования кластеров для наукометрических баз данных, а также параметры необходимые для оптимизации JVM.

4. Рассмотрен функционал существующих база данных и приведен сравнительный. Приведена важность целостности данных и отказоустойчивости в работе существующих решений путем анализа информации предоставленной компанией Clarivate Analytics.

Рекомендации по практическому использованию результатов

Предложенное исследование может использоваться компанией Clarivate Analytics, в частности для системы Web of Science, а также смежными компаниями, которые занимаются качественным и количественным анализом научной информации.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. Моисеенко, А. А. Наукометрические базы данных / Моисеенко А. А. // Центр научных публикаций сборник научных публикаций «ВЕЛЕС», Киев, 17 июнь 2019 г. – Киев: Украина 2019. – С. 79 – 82.

2. Моисеенко, А. А. Формирование кластеров в наукометрических базах данных / Моисеенко А. А. // International independent scientific journal, Краков - Краков: Польша 2019. - С. 19 – 23.

РЭЗЮМЭ

Маісеенка Аляксандр Аляксандравіча

Інструментальны сродак фармірования навуковаметрычнай базы даных

Ключавыя словы: навукаметрыя, статыстыка, базы даных, размеркавальныя сістэмы, Web of Science, алгарытмы, хэшаванне.

Мэта работы: даследаванне і аптымізацыя алгарытмаў звязвання даных шляхам аптымізацыі хэшавання ў існуючых сістэмах, аптымізацыя выкарыстання рэсурсаў, аптымізацыя структуры сістэмы.

Атрыманыя вынікі і іх навізна: у працы прапанаваны тэхналогіі і метады рашэння праблем, якія ўзнікаюць пры выкарыстоўванні існуючых навуковаметрычных сістэм. Гэтыя прапановы дазволяць вырашыць шэраг праблем ад некарэтнага звязвання дадзеных да аптымізацыі выкарыстоўванных рэсурсаў і прадухіленне уцечак памяці.

Вобласць ужывання: навукаметрыя, навуковыя публікацыі.

РЕЗЮМЕ
Моисеенко Александра Александровича
Инструментальное средство формирования наукометрической базы данных

Ключевые слова: наукометрия, статистика, базы данных, распределенные системы, Web Of Science, алгоритмы, хеширование.

Цель работы: исследование и оптимизация алгоритмов связывания данных в существующих системах, оптимизация использования ресурсов, оптимизация структуры системы.

Полученные результаты и их новизна: В работе предложены технологии и методы решения проблем, которые возникают при использовании существующих наукометрических баз данных. Эти предложения позволят решить ряд проблем от некорректного связывания данных до оптимизации использования ресурсов и предотвращения утечек памяти.

Область применения: экономика, дистанционное обучение.

SUMMARY
Moiseenko Alexander Alexandrovich
Investment Portfolio Formation and Management Support System

Keywords: scientometrics, statistics, databases, distributed systems, Web Of Science, algorithms, hashing.

The object of the study: research and optimization of data linking algorithms by optimizing hashing in existing systems, use optimization of resources, structure optimization of the system.

The results and novelty: this work proposed a technologies and methods that arise while using existing scientometric databases. These suggestions can solve a number of problems from incorrect data linking to optimizing the use of resources and prevention of memory leaks.

Sphere of application: scientometrics, scientific publications.