ЮДИН М. О., КУДЕЛИН А. Г. ИНФОРМАЦИОННАЯ СИСТЕМА АВТОМАТИЧЕСКОГО ПОИСКА И АНАЛИЗА НАУЧНЫХ ПУБЛИКАЦИЙ

УДК 004.912, ВАК 1.2.2. / 05.13.18, ГРНТИ 20.53.19

Информационная система автоматического поиска и анализа научных публикаций

М. О. Юдин, А. Г. Куделин

Ухтинский государственный технический университет, г. Ухта

В статье представлена работа по проектированию и разработке «Информационной системы автоматического поиска и анализа научных публикаций» для уменьшения затрат времени на поиск статей, а также увеличения точности их анализа. Анализ предметной области выявил, что для поиска сотрудником научных публикаций по необходимой тематике для изучения и использования в научной деятельности занимает большой промежуток времени, а результаты поиска зачастую недостаточно точны, чтобы найти сразу необходимую информацию. Разработка информационной системы упростит данный процесс, позволит сократить время на поиск и изучение подходящих вариантов научных данных.

Ключевые слова: *информационная система, поиск, анализ, статья*

Information system for automatic search and analysis of scientific publications

M. O. Yudin, A. G. Kudelin

Ukhta State Technical University, Ukhta

The article presents the work on the design and development of an "Information system for automatic search and analysis of scientific publications" to reduce the time spent on searching for articles, as well as to increase the accuracy of their analysis. The analysis of the subject area revealed that it takes a long time for an employee to search for scientific publications on the necessary topics for study and use in scientific activities, and the search results are often not accurate enough to find the necessary information immediately. The development of an information system will simplify this process and reduce the time spent on searching and studying suitable versions of scientific evidence.

Keywords: *information system, search, analysis, article*

Введение

Ни для кого не секрет, что Интернет является наиболее масштабным хранилищем данных. В Интернете можно найти невообразимое количество

данных на разнообразную тематику. В большинстве своем эти данные хранятся в текстовом формате и с каждым годом количество публикаций увеличивается. В то же время, для научного сообщества прежде всего информация является ценной, если она предоставляется в виде рецензируемых статей научных журналов. Существуют различные библиографические и реферативные базы данных («SciElo», «PubMed»), которые используются такими площадками как «Web of Science» и «Google Scholar».

Основным источником рецензируемых статей для системы, описываемой в данной статье, является хранилище «Scopus».

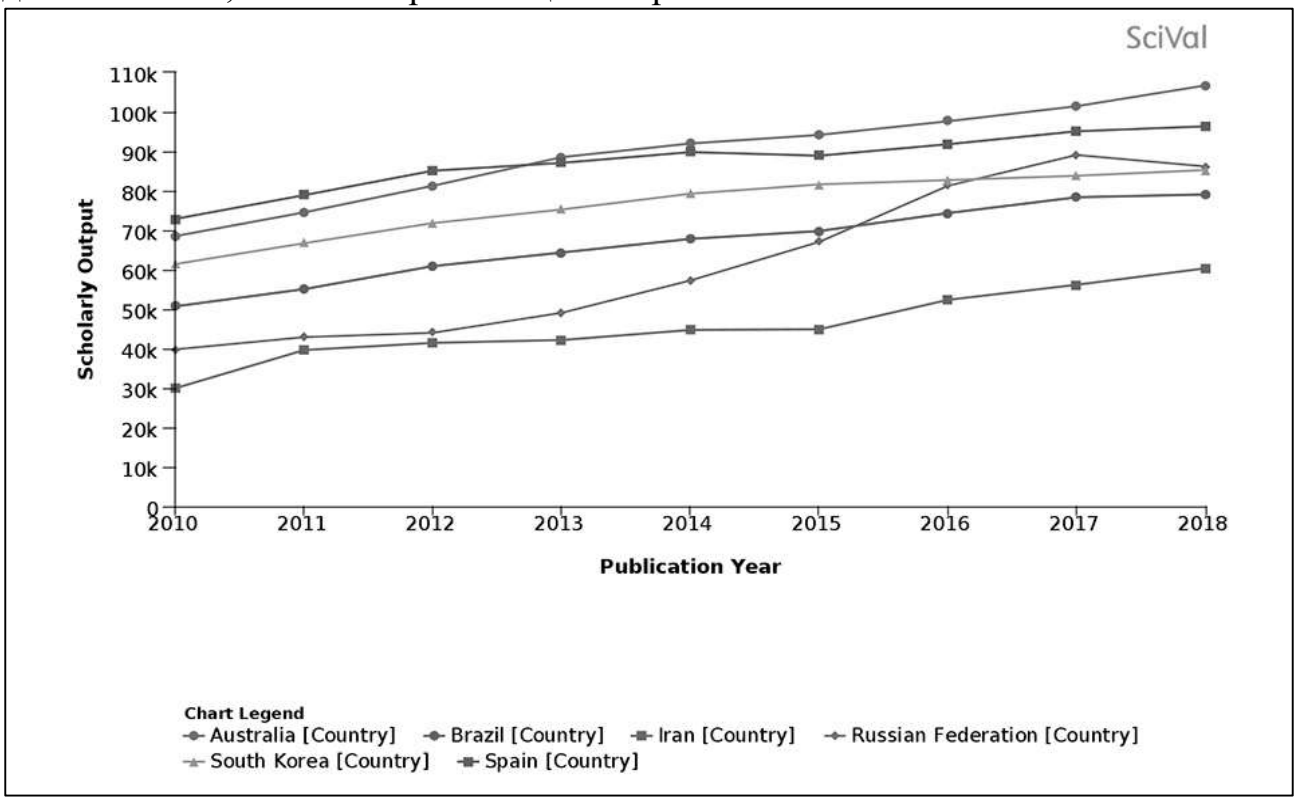

Рисунок 1. Статистика опубликованных статей из разных стран в системе «Scopus»

«Scopus» – это библиографическая и реферативная база данных и инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях [1].

Даже поиск, по ключевым словам, выводит очень много результатов. Чтобы как-то ускорить данный процесс было решено разработать информационную систему автоматического поиска и анализа научных публикаций и последующего их отображения на созданном web-ресурсе.

Проектирование программного обеспечения

Требуется автоматизировать процесс поиска и анализа научных публикаций. Для того, чтобы получить статьи, попавшие в хранилище «Scopus», необходимо их загрузить в базу данных посредством API, которое предоставляется разработчиками в издательском доме «Elsevier». «Elsevier» — это один из четырех крупнейших издательских домов мира, который ежегодно

выпускает около четверти всех статей из издаваемых в мире научных журналов. Перед использованием контракта API следует получить API ключ, который предоставляется всем, зарегистрировавшимся пользователям на сайте «Elsevier Developer». Далее необходимо определить подходящие для поставленной цели API. В данном случае были использованы три контракта:

− «Serial Title API» предоставляет список всех журналов, которые были проиндексированы в системе;

− «Scopus Search API» позволяет получить список статей, которые были опубликованы и проиндексированы в журнале, а также их уникальных идентификатор «Scopis ID»;

− «Article (Full Text) Retrieval API» возвращает полный текст статьи, включая ее абстракт, при указании ее «Scopis ID».

Полученные данные необходимо сохранить в БД для последующих действий по определению меры семантической близости между подтвержденными пользователем статьями и статьями, полученными при помощи API.

Мера семантической близости — это особая мера близости, предназначенная для количественной оценки семантической схожести лексем, например, существительных или многословных выражений. Такая мера показывает высокие значения для пар слов, находящихся в семантических онтошениях (синонимия, гипонимия, ассоциативность, когипонимия), и нулевые значения для всех остальных пар [6].

Используя семантический анализ текста, можно определить, что два текста схожи между собой по тематике, даже если эта схожесть выражена косвенно. Или, например, «лыжи» и «автомобиль» по отдельности относятся к разным категориям, но будучи использованы вместе, могут быть интерпретированы в таких категориях, как «спорт» и «отдых».

На данный момент существует несколько способов определения меры семантической близости между текстами:

- Байесовский классификатор;
- латентное размещение Дирихле;
- нейронные сети;
- векторные методы;
- деревья решений эволюционный анализ и генетическое программирование;
- латентно-семантический анализ [5].

В следствии изучения всех методов и сравнения их плюсов и минусов, был выбран метод латентно-семантический анализ (LSA – Latent semantic analysis), он же латентно-семантическое индексирование (LSI) [4].

Суть латентно-семантического анализа состоит в том, что порядок слов в тексте не имеет значения и в каких морфологических формах они представлены, важно только количество вхождений конкретных слов. Предположим, что каждую тему можно охарактеризовать определенным набором слов и частотой их появления. Если в тексте конкретный набор слов употребляется с определенными частотами, то текст принадлежит к определенной теме.

Однако в первую очередь прежде, чем переходить к оценке меры семантической близости, необходимо обработать текст – освободить текст от шумов. Для этого можно использовать: семантическое ядро и стемминг. Стемминг – это процесс нахождения основы слова для заданного исходного слова [2]. Семантическое ядро – это подборка понятий, имеющих существенное значение для данной предметной области [2].

Далее для определения меры семантической близости использовался следующий алгоритм:

− выявляется коэффициент каждого слова относительно общего количества слов;

− после происходит выявление синонимов внутри списка. Если слова синонимичны между собой, то берется только то слово, у которого больший коэффициент;

− потом происходит нормализация полученных коэффициентов по формуле нормализации;

− данный процесс повторяется со всеми статьями, которые были найдены в журналах;

− последним этапом является поиск близости статей: коэффициент близости – это сумма перемноженных совпадающих слов между примером и найденными статьями.

Данный был выполнен на языке Python, поскольку на данном языке программирования имеются все необходимые библиотеки для работы с большими объемами текстовых данных.

После определения меры семантической близости несколько статей с наибольшей мерой семантической близости с ранее выбранными пользователям статьями помещаются в отдельные таблицы базы данных, которые в последствии отображаются пользователю, как наиболее подходящие для изучения и опубликования на сайте.

Для разработки обеспечения помощи научных сотрудников необходимо было выявить наиболее подходящую под поставленные задачи CMS (Content Management System), которая помогла бы упросить и значительно ускорить выполнение задачи по разработке и наполнению контентного сайта. Среди множества представленных на данный момент систем управления содержимым (CMS) нами была выделена одна наиболее подходящая под наши задачи – WordPress.

WordPress – это самая популярная среди всех остальных CMS с открытым кодом, которая в основе своей использует язык разработки PHP. Была выбрана эта система управления контентом, поскольку она имеет открытый исходный код, с возможностью разработки собственных плагинов для web-приложения, кроме того, в ней имеются все необходимые инструменты для разработки новостного сайта, а также богатейший выбор все возможных тем и плагинов, которые помогают решить большинство задач по разработке Web-портала.

Изначально при проектировании сайта важно выбрать подходящую тему, которая будет правильно отражать суть новостного портала. Выбор пал на ColorMag, эта адаптивная тема, заточенная на публикацию новостей, газет,

журналов и прочих видов сайтов. При помощи данной темы был предопределен первичный вид сайта: выставлена шапка сайта с меню и логотипом сайта, основная часть сайта была разделена на подзаголовки, которые разделяют новости на под темы. Также в нижней части сайта расположился футор с дополнительной информацией о компании, представляющей данный портал. Далее необходимо было найти все нужные плагины, помогающие добавить функционал и корректно настроить его работу. Elementor – средство редактирования страниц на сайте, позволяющее легко выставить контент в нужные блоки на сайте. Ultimate Member – плагин, добавляющий окно регистрации на сайт, при помощи него администратор сможет разделить пользователей по ролям, выделив простых читателей, писателей и пр.

Поскольку на сервере находится помимо разрабатываемого нами сайта другие web-ресурсы, было принято решение о распределении необходимых для работы файлов в контейнеры, при помощи программного обеспечения docker.

Функции системы

Основными функциями разрабатываемой системы являются:

– Аутентификация и авторизация пользователя при входе в систему.

– Сохранение введенных данных пользователем при работе с приложением.

– Сбор статей из хранилища данных.

– Формирование списка статей.

– Назначение данных для примера к поиску похожих статей.

– Определение списка журналов для проведения в них поиска статей.

– Формирование, отфильтрованного по критерию близости, списка статей, с возможностью скачать статью.

Результат разработки системы

На данном этапе реализации «Информационной системы автоматического поиска и анализа научных публикаций» были реализованы основные функции веб-приложения, отвечающие поставленным требованиям.

Для сайта был разработан плагин, использующий возможности API «Scopus» для сбора данных о журналах и статьях, которые в них публиковались и последующего отображения пользователю в удобном для чтения формате.

Рисунок 2. Страница списка журналов

Рисунок 3. Страница с текстом статьи в журнале

Кроме отображения журналов и статей было разработано приложение по поиску и анализу статей, полученных из «Scopus». Пользователь информационной системы для проведения поиска в первую очередь выбирает из каких журналов загрузить статьи, для формирования примера перед анализом. Для более точной загрузки пользователю доступен ввод дополнительных параметров: года публикации статей и ключевых слов статьи (Рисунок 4Рисунок).

Рисунок 4. Выбор журнала

Далее пользователю необходимо выбрать, какие статьи будут использоваться для формирования примера (Рисунок 5), либо он может ввести свой текст (Рисунок 6).

Рисунок 5. Выбор статьи для создания примера

Рисунок 6. Ввод собственного текста для создания примера

После формирования примера, пользователю остается выбрать журналы, в которых будет проводиться поиск (Рисунок 7) и получить результаты поиска и анализа научных публикаций (Рисунок 8Рисунок).

Рисунок 7. Выбор журналов, в которых будет проводиться поиск

Рисунок 8. Список статей полученный после анализа

Рисунок 9. Результата анализа

Заключение

В данной статье дано краткое описание работ по разработке «Информационной системы автоматического поиска и анализа научных публикаций». Помимо вышеописанных пунктов, процесс разработки системы включил в себя следующие этапы:

– Выполнено предпроектное исследование, выделены границы системы с помощью контекстной диаграммы, на которой были выявлены две основные

сущности пользователь и информационная система, проведена декомпозиция основного процесса – поиск и анализ научных публикаций.

– осуществлен выбор средств проектирования, рассмотрены их основные характеристики и преимущества при реализации системы. Для проектирования были использованы API, которые обращаются к базе данных «Scopus», языки программирования Python и PHP, а также система управления содержимым «WordPress»;

– выполнена разработка технического задания на выполнение работы;

– разработана логическая модель данных, на которой были отражены все таблицы необходимые для работы пользователя с информационной системой и необходимые для сохранения всех выбранных пользователем полей на страницах ИС, а также результат работы автоматического поиска и анализа статей. На основе логической модели данных была построена физическая модель базы данных;

– организована информационная безопасность системы.

– выполнена реализация всех функциональных требований (аутентификация и авторизация пользователя при входе в систему; сохранение введенных данных пользователем при работе с приложением; сбор статей из хранилища данных; формирование списка статей. назначение данных для примера к поиску похожих статей; определение списка журналов для проведения в них поиска статей; формирование, отфильтрованного по критерию близости, списка статей, с возможностью скачать статью) для достижения поставленной цели.

Дальнейшая доработка системы включает в себя реализацию полного функционала и внедрение в комплекс автоматизированных информационных систем УГТУ.

Список использованных источников и литературы

1. Scopus: сайт Википедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Scopus (дата обращения 14.03.2021).

2. Бондарчук Д. В. Определение семантической близости термов с помощью контекстного множества [Электронный ресурс]. – Режим доступа: https://elar.urfu.ru/bitstream/10995/43751/1/cai-2016-41.pdf (дата обращения 14.03.2021).

3. Российская наука в Scopus и WoS: количество или качество: сайт Indicator [Электронный ресурс]. – Режим доступа: https://indicator.ru/engineering-science/rossijskaya-nauka-v-scopus-i-wos-kolichestvo-ili-kachestvo.htm (дата обращения 15.03.2021).

4. Латентно-семантический анализ: сайт Habr [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/110078/ (дата обращения 15.03.2021).

5. Бондарчук Д. В. Алгоритмы интеллектуального поиска на основе метода категориальных векторов [Электронный ресурс]. – Режим доступа: https://www.susu.ru/sites/default/files/dissertation/dissertation.pdf (дата обращения 15.03.2021).

6. Мера семантической близости: сайт NLPub [Электронный ресурс]. – Режим доступа: https://nlpub.ru/ (дата обращения 14.03.2021).

List of references

1. Scopus: site Wikipedia [Electronic resource], https://ru.wikipedia.org/wiki/Scopus (accessed 14.03.2021).

2. Bondarchuk D. V. Determining the semantic proximity of terms using a contextual set [Electronic resource] // Ural State University of Railway Transport. URL: https://elar.urfu.ru/bitstream/10995/43751/1/cai-2016-41.pdf (accessed 14.03.2021)

3. Russian Science in Scopus and WoS: quantity or quality: Indicator website [Electronic resource] // Indicator.ru. Updated: 08.02.2019. URL: https://indicator.ru/engineering-science/rossijskaya-nauka-v-scopus-i-wos-kolichestvo-ili-kachestvo.htm (accessed 15.03.2021)

4. Latent Semantic Analysis: Habr website [Electronic resource] // Habr.com. Updated: 20.12.2010. URL: https://habr.com/ru/post/110078/ (accessed 15.03.2021)

5. Bondarchuk D. V. Algorithms of intellectual search based on the method of categorical vectors [Electronic resource] // Ural State University of Railway Transport. 2016. URL: https://www.susu.ru/sites/default/files/dissertation/dissertation.pdf (accessed 15.03.2021)

6. Measure of semantic proximity: NLPub site [Electronic resource] // NLPub.ru: a catalog of resources for natural language processing. Date of update: 13.10.2017. URL: https://nlpub.ru (accessed 14.03.2021)