УДК 004.6-024.11:004.738.5

СИСТЕМА КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ ИНТЕРНЕТ ИСТОЧНИКОВ

М.П. Батура
Научный руководитель НИЛ 8.1 БГУИР, Доктор технических наук, профессор, академик «Международной академии наук высшей школы»

И.И. Пилецкий
Доцент кафедры информатики БГУИР, кандидат физико-математических наук, доцент, старший научный сотрудник

В.А. Прытков
Проректор по учебной работе БГУИР, кандидат технических наук, доцент

Н.А. Волорова
Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент

В.Н. Козуб
Аспирант кафедры информатики БГУИР, магистр технических наук, ассистент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: bmpbel@bsuir.by, ianmenski@gmail.com, prytkov@bsuir.by, volorova@bsuir.by, kozub@bsuir.by

Аннотация. В статье приводится описание инструмента мониторинга открытых интернет-источников с целью выявления экспертов в некоторой научной области, определения тематик публикаций, оценки популярности публикаций. Описываются принятые решения при построении аналитического комплекса и полученные результаты его работы.

Ключевые слова*:* интернет-источники, Big Data, мониторинг, анализ, Machine Learning, машинное обучение, Neo4j, Hbase, Natural Language Processing, обработка естественного языка

М.П. Батура

Руководитель проекта, доктор технических наук, профессор, академик Международной академии наук высшей школы, заслуженный работник образования Республики Беларусь, автор более 150 научных работ, в том числе 4 монографий, 5 учебников и учебных пособий. Ректор Белорусского государственного университета информатики и радиоэлектроники со 02 октября 2000 года по 23 мая 2018 года. Награжден нагрудным знаком «Выдатнік адукацыі Рэспублікі Беларусь», медалью «За трудовые заслуги».

И.И. Пилецкий

Кандидат физико-математических наук, доцент, доцент кафедры информатики БГУИР, научный руководитель совместной лаборатории БГУИР — IBA и Академического центра компетенций технологий IBM. Имеет большой опыт в реализации и разработке промышленных решений в ИТ-области, являлся ведущим разработчиком, системным архитектором, руководителем и научным руководителем нескольких десятков крупных проектов, связанных с разработкой программного обеспечения и баз данных корпоративного уровня. Автор более 80 научных публикаций (в том числе монографий и учебных пособий) в области моделирования, технологии разработки программного обеспечения и баз данных. Член редакционной коллегии журнала «Baltic Journal of Modern Computing». Область научных интересов: технологии и аналитические комплексы анализа Big Data, NPL и ML алгоритмы.

В.А. Прытков

Кандидат технических наук, доцент, проректор по учебной работе БГУИР. Автор 38 научных и методических публикаций, являлся исполнителем, ответственным исполнителем или руководителем 12 научно-исследовательских проектов, член организационного комитета ряда международных научно-практических конференций. Область научных интересов: обработка изображений и текстурный анализ, синтаксические методы обработки информации, анализ слабоструктурированных данных.

Н.А. Волорова

Кандидат технических наук, доцент, заведующая кафедрой информатики БГУИР. Автор более 100 научных и методических публикаций, 19 изобретений, принимала участие в 20 научно-исследовательских работах в качестве исполнителя, ответственного исполнителя или руководителя. Член организационного комитета ряда международных научно-практических конференций. Область научных интересов: имитационные модели сложных систем, системы автоматизации моделирования.

В.Н. Козуб

Аспирант, ассистент каф. информатики БГУИР, имеет 5 публикаций тезисов докладов на научной конференции аспирантов, магистрантов и студентов БГУИР, 4 публикации в материалах международных конференций, 1 методическое пособие, ведет лабораторные занятия по курсам "Архитектурные решения для обработки больших объемов информации", "Модели и методы обработки и анализа больших объемов информации", "Технологическая платформа по управлению большими данными".

Введение

В настоящее время информация, полученная в результате анализа данных интернет-источников, является одной из базовых для принятия решений.

Как правило, это неструктурированные текстовые данные, различные мультимедийные данные. Данные могут быть получены как из социальных сетей, так и тематических сайтов (газет, журналов, библиотек, компаний и т. д.), содержащих различные публикации. Есть много работ, которые посвящены принятию решения на основании применения некоторого метода анализа данных. Результатами анализа смогут воспользоваться компании для создания систем поддержки пользователей, социологи для анализа общественного мнения, организаторы мероприятий для получения отклика участников, знаменитости для отслеживания репутации в сети, правительство для контроля настроений в обществе и др.

Пользователи интернет-ресурсов и социальных сетей могут самостоятельно выбирать интересные им направления и читать публикации интересных им людей, которым они симпатизируют (говоря языком интернета, подписываться на обновления контента). В свою очередь создатели контента могут быть со своим контентом и своими подписчиками. Такие связи, как правило, бывают достаточно сложными и представляют собой многоуровневую циклическую сеть.

Несмотря на наличие различных систем для анализа данных из интернет-источников, данная тематика не только не утрачивает актуальности, а, напротив, становится все более востребованной. В настоящее время в социальных сетях и на тематических сайтах можно найти разнообразную информацию практически о любых явлениях в мире, деятельности

организаций и людей. Проблемой является то, как разобраться в этом многообразии источников данных и самих данных (рисунок 1) как превратить данные в информацию, а информацию – в знания для принятия разумных решений.

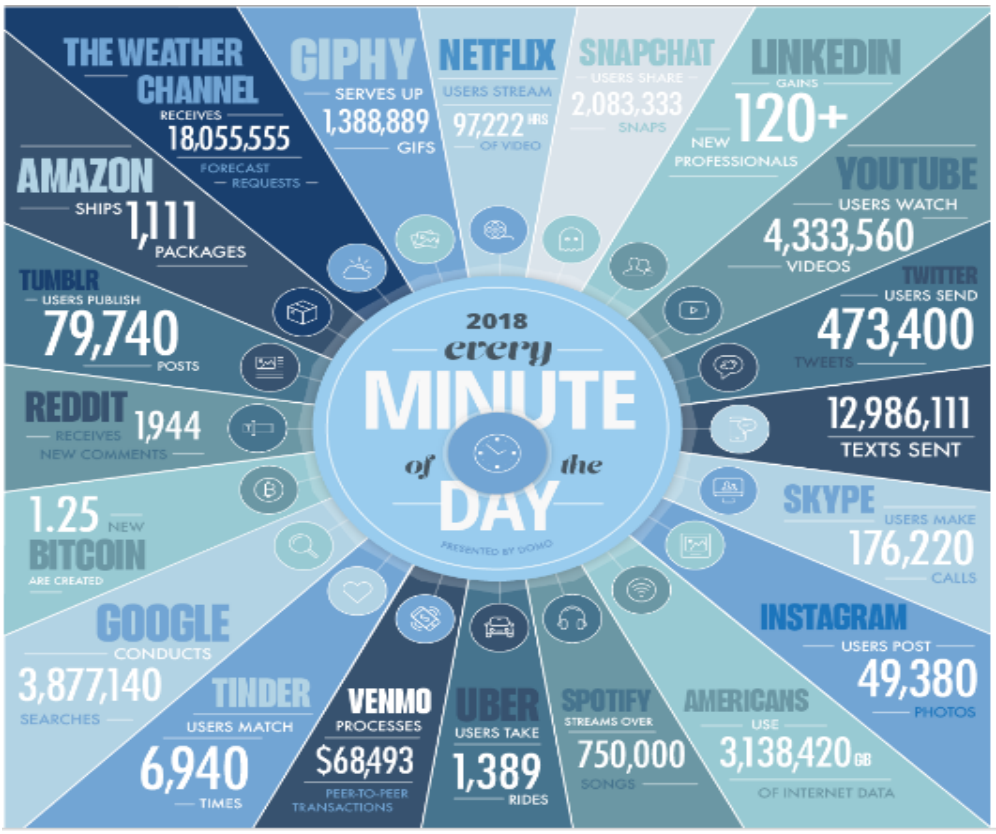

Рисунок 1. Разнообразие источников данных: источник: [1]

Назначение и цели

В данной статье рассматривается проект создания «Системы комплексного анализа данных интернет-источников (СКА)», позволяющей анализировать большие объемы данных из интернет-источников в области научных исследований и предназначенной для сбора информации о научных публикациях, построения графа знаний, что дает возможность определять экспертов предметной области, тематики их работ, их взаимосвязи, а также определять передовые научные направления.

В момент подготовки данной статьи в реализации проекта СКА принимали участие студенты БГУИР кафедры информатики: Гутковский В.Н., Сернацкий В. И., Судникович К. И., Филиппович В. В., Черныш Н. Н., Шпаков Н. И.

Система должна находить экспертов (авторитетов) в предметной области и выдавать оценку их рейтинга влиятельности. Например, лучше прочитать три книги признанных экспертов в определённой области, чем десять книг дилетантов.

Перспективность проекта СКА в том, что результаты могут быть использованы как для прикладных целей, например, определения наиболее перспективных и актуальных научных направлений с учетом конкретной специфики (для конкретной организации или страны, с учетом имеющегося задела и т. д.), определения экспертов в заданных предметных областях (для приглашения специалистов или формирования команды исследовате-

лей), так и для научно-образовательной цели – подготовки специалистов по анализу больших объемов неструктурированных данных (Data Scientist) на основе разработки и оптимизации аналитических ML-алгоритмов, применяемых в системе.

Модульность СКА позволит при минимальной адаптации использовать данный программный комплекс для анализа и мониторинга различных явлений по заказу конкретных пользователей, например, изучения популярности «бренда», поиска блогеров в социальных сетях, изучения конкурентов, изучения рынка сбыта, поведения групп людей и т.д.

Интерес к системам подобного рода постоянно усиливается, поскольку создание новой техники и технологий становится все более наукоемким, причем эта тенденция характерна и для продукции массового рынка. Соответственно, компании, внедряющие новые технологии, а также придающие своей продукции интеллектуальные функции, обладают определенным конкурентным преимуществом. На глобальном уровне это проявляется в определении своей ниши в мировой экономике, ее расширении.

СКА позволит не только помочь с проведением анализа и принятия решения, но и позволит сэкономить весьма дефицитный в условиях конкуренции ресурс – время. В современной экономике важно не только дать качественный продукт с новыми свойствами, но и сделать это в числе первых.

Ценность работы заключается не только в создании «Системы комплексного анализа данных интернет-источников», но и в создании в Университете многоцелевого, модифицируемого кластера для анализа данных интернет-источников и глубокой математической подготовки специалистов по анализу больших объемов неструктурированных данных, специалистов Data Scientist, внедрении результатов исследований в научно-образовательный процесс при подготовке магистрантов в области обработки больших объемов информации.

Facebook, Google, другие популярные сервисы построены на использовании графовых моделей данных. Facebook, например, использует не только информацию о людях, их именах, профессиях и т. д., но также сведения о взаимосвязях между людьми, которые представляют ещё большую ценность. Социальные отношения могут быть явными или неявными, а социальные сети помогают идентифицировать как прямые, так и косвенные отношения между людьми, группами людей, и характер их взаимодействий. Gartner утверждает, что способность использовать эти графы обеспечивает «устойчивое преимущество в конкурентной среде».

В отличии от известных поисковых систем (например, Facebook, Google) СКА позволяет найти наиболее перспективные и актуальные научные направления и определить экспертов в заданных предметных областях.

Технология построения и функционирования СКА

Основу технологии разработки системы составляют методы и алгоритмы построения и обслуживания графовой модели социальной сети авторов и их публикаций, ссылок на их публикации и определение рейтинга конкретного автора публикаций, определение тематик публикаций и классификация их по областям знаний.

Анализируя данные из социальных сетей можно выявить как прямые, так и скрытые отношения между людьми, группами людей, а также характер их взаимодействий. На основе определенных связей между субъектами группы можно сделать выводы об индивидуальных предпочтениях субъектов и прогнозировать выбор объекта предпочтения.

Графовая модель социальной сети может быть построена на применении классической графовой модели, которая включает узлы и взаимосвязи, а также их свойства и метки.

Основным назначением графовой базы данных является применение графовых алгоритмов для обработки полученных данных, выстраивание логических взаимосвязей и подготовка и выдача информации для пользователя. Также ключевой особенностью таких БД

является формирование очень гибких запросов, наподобие следующих: ***Какие авторы наиболее часто сотрудничали с автором данной популярной статьи****, является ли данный автор писателем только в области биохимии,* ***или же он пишет еще и на темы математического анализа****, существуют ли математические публикации, которые по какой-то причине перекликаются с темой философии и т. д.*

СКА состоит из следующих компонент (рисунок 2): сбора данных, фильтрации данных и составление «мешка слов» из N-грамм (векторизации), хранилища данных, библиотеки аналитических модулей, подготовки выдачи результата, клиентского модуля.

При необходимости набор модулей и компонент может быть расширен, а некоторые модули заменены новыми. Общая технология построения многофункционального комплекса по обработке данных из интернет и работы компонент, а именно чтения данных интернет-источников, фильтрации данных и векторизации, библиотеки и хранилища приведена в более ранних публикациях [2]. В СКА апробирована технология построения многофункциональных комплексов как набор постоянно работающих компонент в виде отдельных серверов, изменена предметная область и система дополнена компонентом БД «граф знаний» и компонентом подготовки и выдачи результата.

Сами компоненты СКА состоят из набора функциональных модулей, на рис. 2 приведена логическая схема взаимодействия компонент и модулей.

СКА логически выполняет свои функции в несколько этапов:

– Компонент сбора данных – выполняет целевое сканирование выбранных социальных сетей, новостных порталов, сайтов и помещает полученные документы в исходном виде и специально обработанном виде в хранилище (в качестве хранилища используются HDFS и база данных документов Hbase);

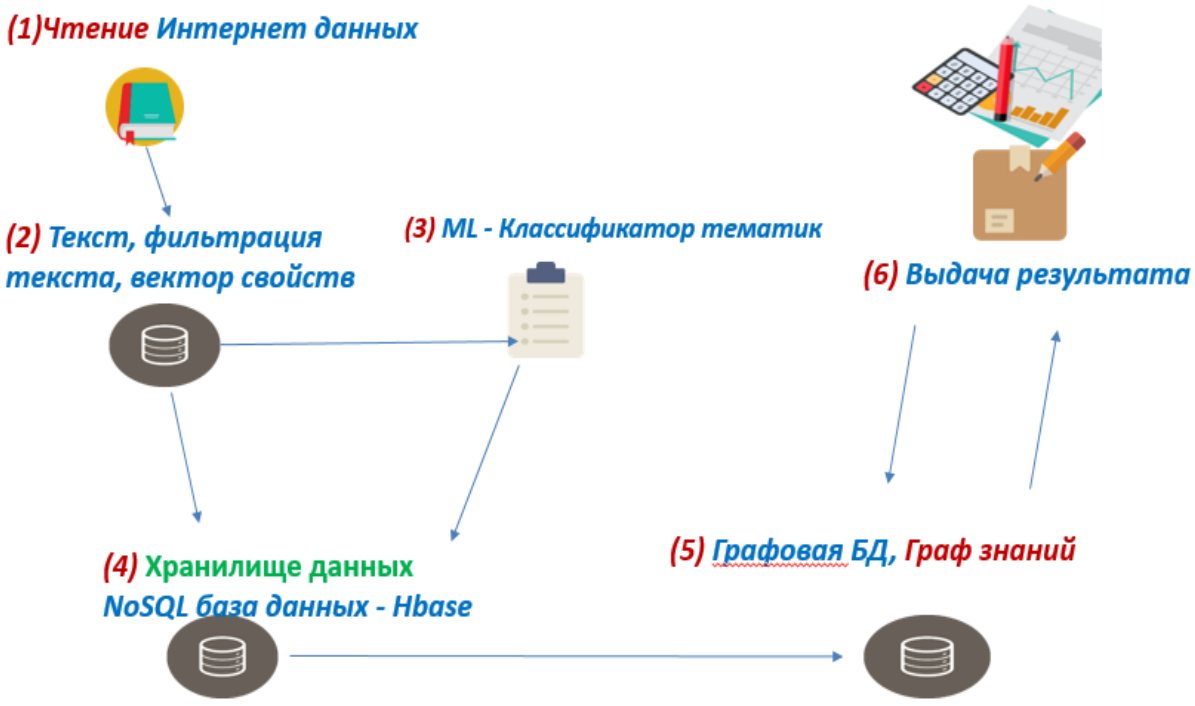

Рисунок 2. Логическая схема взаимодействия компонент

– Компонент обработки данных выполняет фильтрацию исходных данных (токенизацию, лемматизацию, стемминг, удаление стоп-слов, перевод к нижнему регистру), построение «мешка слов» и векторизацию исследуемых наборов документов («мешок слов» в специальном виде помещается в хранилище);

– Компонент библиотека аналитических модулей содержит набор модулей, которые осуществляют обработку данных, полученных из интернет-источников, с целью определения тематик публикаций, а также (в дальнейшем) поиска упоминаний о брендах, определения их тональности и формирования аналитических данных для передачи клиентскому модулю, кроме того, содержит управляющие и служебные модули. Данный компонент анализа данных может содержать модули, реализующие как самообучающиеся Natural Language Processing алгоритмы, так и алгоритмы с учителем. В общем случае компонент библиотеки аналитических модулей *может* содержать модули векторного и регрессионного анализа, модули, реализующие нейросети, модули, использующие вероятностные методы анализа данных. В настоящее время апробированы векторные [3] и вероятностные модули [4], разработаны и апробированы модули построения и анализа графовых моделей, позволяющих получить знания для принятия решений. Область применения СКА может быть расширена за счет пополнения компонента другими модулями, а применение комбинированного подхода к оценке некоторого явления позволит наиболее достоверно идентифицировать новые события и принять правильные решения;

– Компонент подготовки выдачи результата – обеспечивает взаимодействие с компонентом граф знаний, подготавливает информацию в специальном виде для пользователей СКА, а при необходимости применяет технику drill-down для уточнения выдаваемой информации;

– Графический интерфейс обеспечивает взаимодействие с пользователем системы, принимает от пользователя запросы, определяет тип запроса и вызывает нужный модуль компонента подготовки выдачи результата;

– Компонент хранилища данных – содержит данные из интернет-источников, предварительно обработанные и размеченные данные, необходимые для построения классификатора, «мешок слов», а также служебную информацию, необходимую для работы других модулей системы. В хранилище хранятся сырые данные с сайта, текст, фильтрованный текст, исходные документы, «мешок слов», тематика документов и служебная информация;

– Компонент графовая база данных и граф знаний состоят из графовой БД, моделирующей предметную область, и программных модулей, позволяющих пополнять графовую БД данными из хранилища и извлекать из нее знания о запрашиваемых объектах.

Технологическая платформа

В качестве ИТ платформы использовались VM ЦОД БГУИР, Open Source решения: Hadoop кластер, файловая система HDFS, Apache Hbase, Neo4j, библиотеки для тематического моделирования больших коллекций текстовых документов и взаимодействия с графовой БД (BigARTM Pyro, py2neo, react-d3-graph), библиотеки для языка программирования Python и Java (mining.py и др.). Все компоненты СКА реализованы как набор серверов, которые взаимодействуют по заранее определенному сценарию.

Реализация

В настоящее время компоненты СКА реализованы в опытном варианте, и в процессе работы в этом режиме ведутся доработки компонент и функциональности СКА. Ниже в данном разделе приведены результаты работы компонент СКА. Данные получены с сайта, который используется для публикаций научных работ: http://libgen.io.

Данный сайт позволяет загрузить пользователям научные работы к себе на компьютер при помощи торрентов, каждый из которых содержит 1000 научных работ, дату загрузки этого торрента на сайт и его размер.

В настоящее время в хранилище содержится информация более чем о 20 тыс. статей и документов. В дальнейшем планируется получать данные из многих сайтов и при необходимости указывать ссылки к конкретным данным.Документы, статьи читаются с сайта (сайтов), фильтруются, строится векторное представление («мешок слов») и все данные (сырые данные с сайта, текст, фильтрованный текст, исходные документы и «мешок слов») сохраняются в хранилище.

В процессе фильтрации вызываются модули компонента библиотеки аналитических модулей для определения тематики публикаций документов в корпусе документов. Для тематического анализа текста применяется модифицированный алгоритм PLSA [4]. Тематический анализ с использованием EM-алгоритма позволяет выявить N самых важных тем во всём тексте, первоначально на основе мешка слов, а в дальнейшем при пополнении корпуса документов, возможна корректировка классификатора документов.

Как уже было отмечено, все основные данные содержатся в хранилище; структура одной из записей для документа приведена ниже:

Структура записи (Hash - primary key of publication, Title - title of publication, Author - author(authors) of publication, Year - publication date, Pages - number of pages, Publisher - publisher of publication, Language - primary language of publication, Topic - topic or topics of publication, Extension - extension of publication file, Tags - array of publication tags, Locator - name of the file).

Компонент **графовая база** данных добавляет данные из хранилища и регулярно модифицирует информацию в БД и графе знаний. Базой для получения знаний данного компонента является разработанная графовая модель предметной области.

Графовая модель предметной области.

Графовая модель является развитием модели, описанной ранее [5], и в виде множества полей и обобщенного графа представлена на рис. 3.

Сущности и связи модели:

Автор – тот, кто опубликовал статью. Содержит данные о своем имени и о статьях, которые написал (отношение WROTE).

Публикация – публикация, написанная некоторым автором. Содержит в себе имя, год публикации (необходимо как минимум для генерации page-links для page_ranking алгоритмов), ID (sha256), ISBN, публикатор, количество страниц, язык, расширение файла, теги.

С публикацией соотносится следующая информация:

− ссылки на темы, к которым с определенной вероятностью она относится (вероятностная модель, отношение THEME_RELATION);

− ссылки на используемые в публикации источники (LINKS_TO, необходимо для page_ranking);

− ссылки на ключевые фразы, которые входят в ее текст (FREQUENCY, содержит количество вхождений, интерпретация мешка слов).

Тема – область знаний, к которой может относиться публикация. Определяется с помощью тематического анализа текста публикации, позволяет просмотреть все публикации, относящиеся к этой теме через THEME_RELATION.

Токен – сущность, которая представляет себя уникальным именем. Имеется возможность просмотреть все публикации, которые имеют в своем тексте вхождение ключевой фразы, также можно просмотреть количество вхождений (все это хранится в связи FREQUENCY).

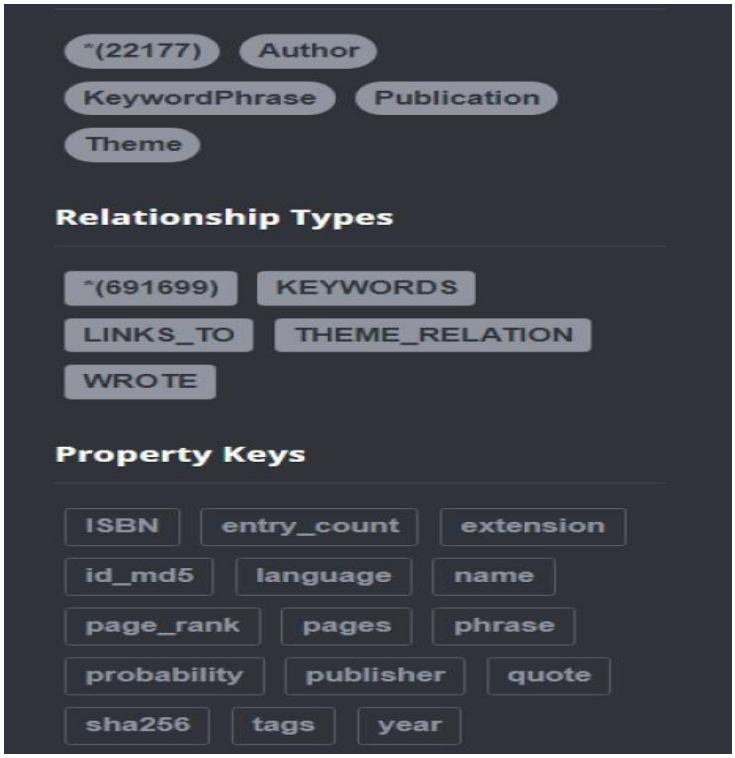

Рисунок 3. Модель БД в виде множества полей

Графовая модель позволяет получать знания о публикациях в различных аспектах, например, связанных с тематикой «биология». В таких запросах важно указывать порог вероятности тематики в статьях больше некоторой величины.

Ниже приведены примеры получения информации из БД в виде графа.

Анализируя графовую модель на рисунке 4 можно получить информацию о связях ключевых фраз (желтый цвет), к которым относится несколько публикаций, публикации с несколькими ключевыми словами (красный цвет) и авторы публикаций (зеленый цвет), а также связи между публикациями (ссылки из одной статьи на другую).

Запросы могут быть разные, например, может быть запрос о выдаче некоторых статей по теме биология или математика. Таким образом в зависимости от запросов можно получать различную информацию и интерпретировать ее.

Существуют различные алгоритмы определения наиболее важных статей, публикаций, блогеров [6], в СКА используется алгоритм page_rank [7].

В графовой БД есть узлы Publication, которые содержат атрибут page_rank, который использован для определения наиболее важной статьи и ее автора, поэтому СКА, в отличие от популярных поисковых систем, позволяет получить знание о предметной области и определить экспертов этой области.

Например, результат запроса на выдачу первых 7 статей с именами их авторов и значениями page_rank представлен на рисунке 5.

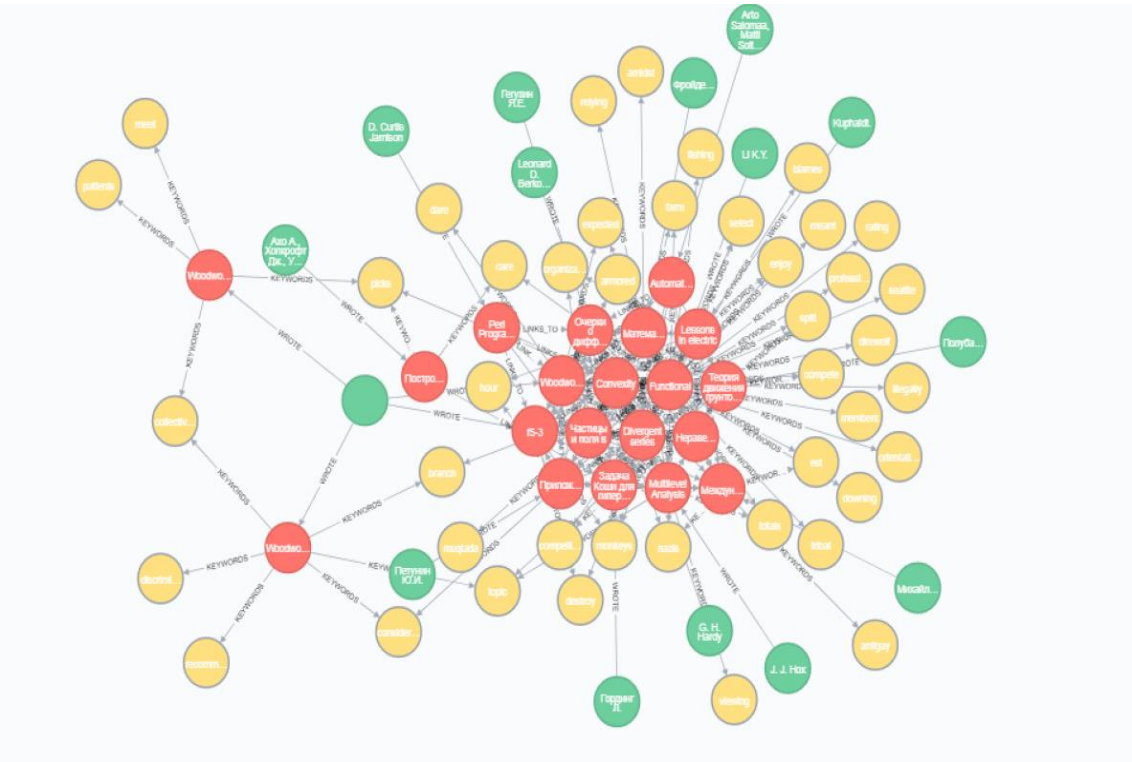

Рисунок 4. Пример получения информации из БД

"G. H. Hardy"	"Divergent series"	188.083844
"Гординг Л."	"Задача Коши для гиперболических уравнений"	144.53579399999998
"Коровкин П.П."	"Неравенства"	116.0384135
"Гегузин Я.Е."	"Очерки о диффузии в кристаллах"	88.62258999999999
"Полубаринова-Кочина П.Я."	"Теория движения грунтовых вод"	73.23886499999999
"Михайлов С.С. (ред.)"	"Международная анатомическая номенклатура (Parisiana nomina anatomica)"	62.33922999999999
"Петунин Ю.И."	"Приложение теории случайных процессов в биологии и медицине"	55.8716905

Рисунок 5. Знания о самых важных документах

Комбинируя подобные запросы с вероятностью тем и ключевыми словами можно находить наиболее популярные статьи и авторов статей по указанному запросу на заданную тему. Сейчас решается задача по выбору параметров при формировании комбинации данных о распределении и page_rank и возможной модификации алгоритма, но продемонстрировать работу можно уже на данном варианте проекта. Например, запрос ниже (рисунок 6) иллюстрирует топ-7 отношений Author->Publication, где вероятность темы больше, чем 0.4, и Publication page_rank которых являются максимальными в данной области.

Рисунок 6. Выдача результатов по специфическому запросу с использованием page_rank и probability distribution

Компонент подготовки выдачи результата и интерфейса взаимодействия с пользователем СКА выполняет:

–поиск публикаций, авторов и областей знаний по части названия;

–поиск наиболее цитируемых экспертов (авторитетов) в предметной области (отображается на панели управления);

–поиск новых зарождающихся областей исследований, также как и тех, к которым интерес уже пропал (отображается на панели управления);

–просмотр различных индивидуальных параметров профиля автора публикаций и параметров его индивидуальной работы (количество публикаций, количество ссылок на них, дата последней публикации);

–просмотр различных индивидуальных параметров некоторой предметной области;

–просмотр различных индивидуальных параметров конкретной публикации (предметные области, количество цитирований);

–произвольный запрос к графу знаний в терминах языка запросов графовой базы данных, используемой в СКА;

–просмотр состояния баз данных проекта;

–разделение доступа пользователей к информации.

Доступ к ресурсам СКА регламентирован на основании профиля пользователя, это администратор или зарегистрированный непривилегированный пользователь. Для доступа к ресурсам СКА непривилегированный пользователь может воспользоваться готовыми страницами доступа или конструктором запросов на языке Cypher. На риунке 7 приведена первоначальная страница пользователя СКА.

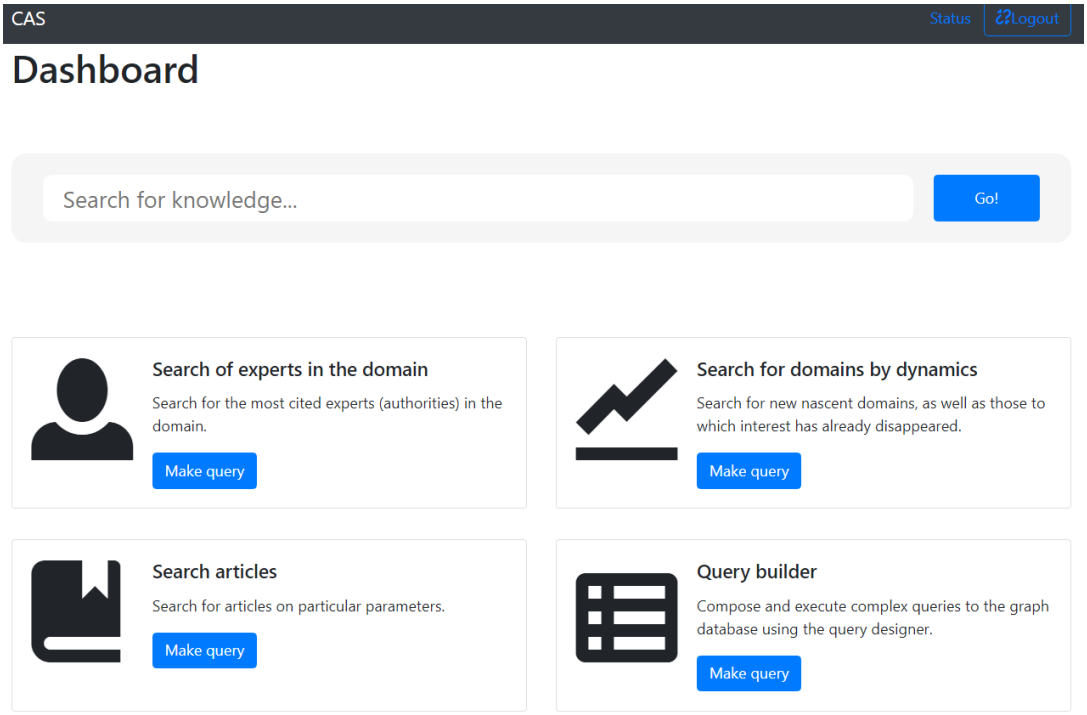

Рисунок 7. Первоначальная страница выдачи

Ниже на рисунке 8 – 15 приведены примеры выдачи знаний СКА о предметной области.

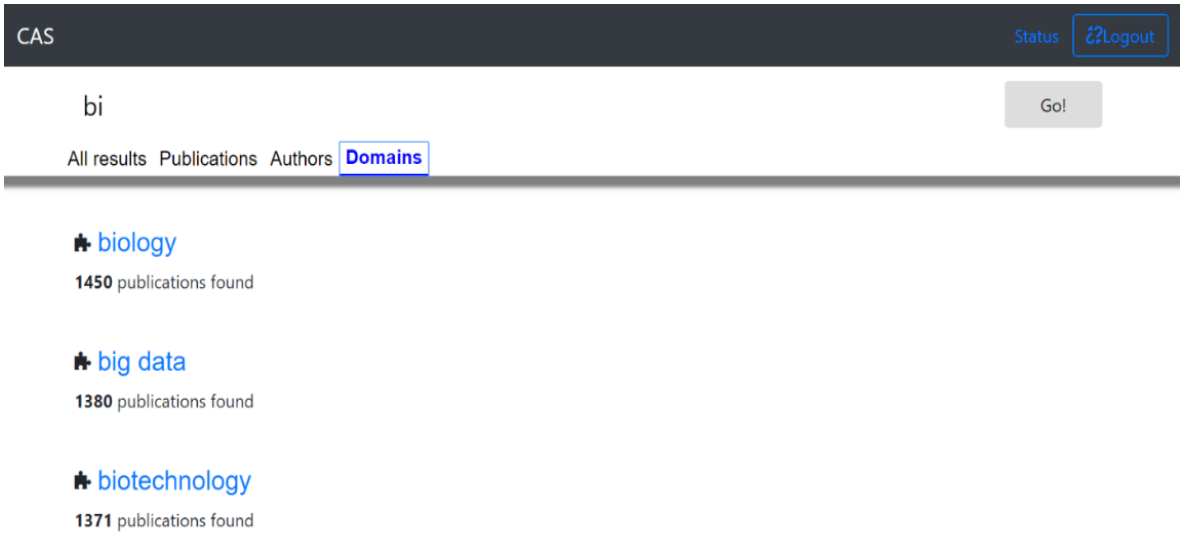

Рисунок 8. Только области знаний

Для каждой «сущности» (автора, публикации, области знаний) создана страница подробной информации (рисунок 9).

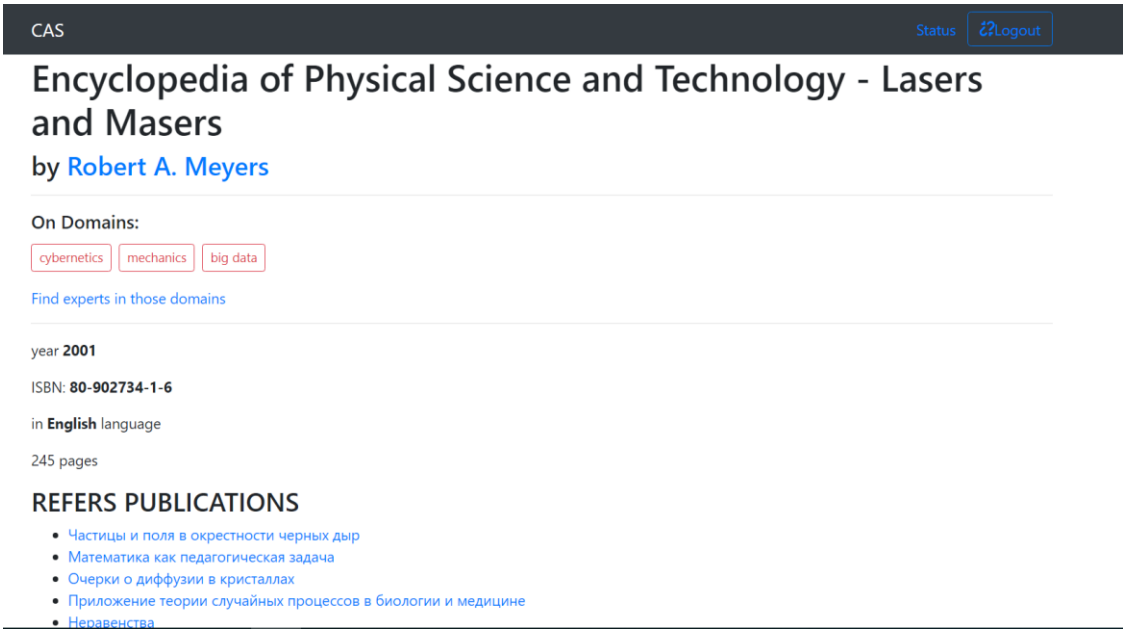

Рисунок 9. Конкретная публикация

На странице публикации присутствуют ссылки на страницу информации об авторе, ссылка на поиск экспертов в тех областях, к которым относится данная публикация, список публикаций, на которые она ссылается (рисунок 10). Поиск авторов-экспертов в области (областях) выдается на основе количества публикаций у автора в обозначенных сферах науки и количестве ссылок на эти публикации.

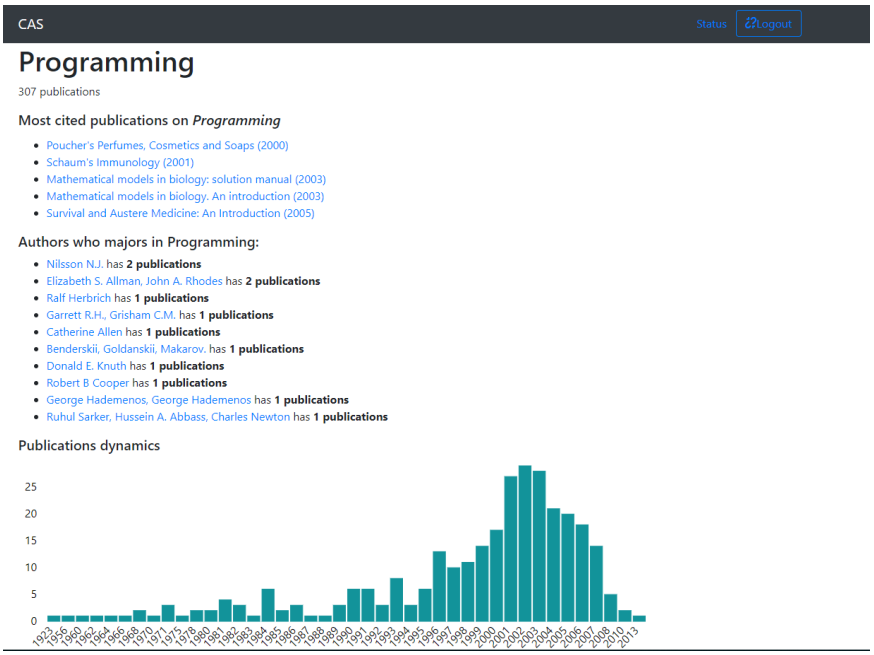

Рисунок 10. Страница области знаний «bIOTechnology»

Поиск авторов-экспертов в области (областях) свои результаты выдает на основе количества публикаций у автора в обозначенных сферах науки и количестве ссылок на эти

публикации. Примеры выдачи экспертов – рисунок 11, а публикации конкретного автора – рисунок 12.

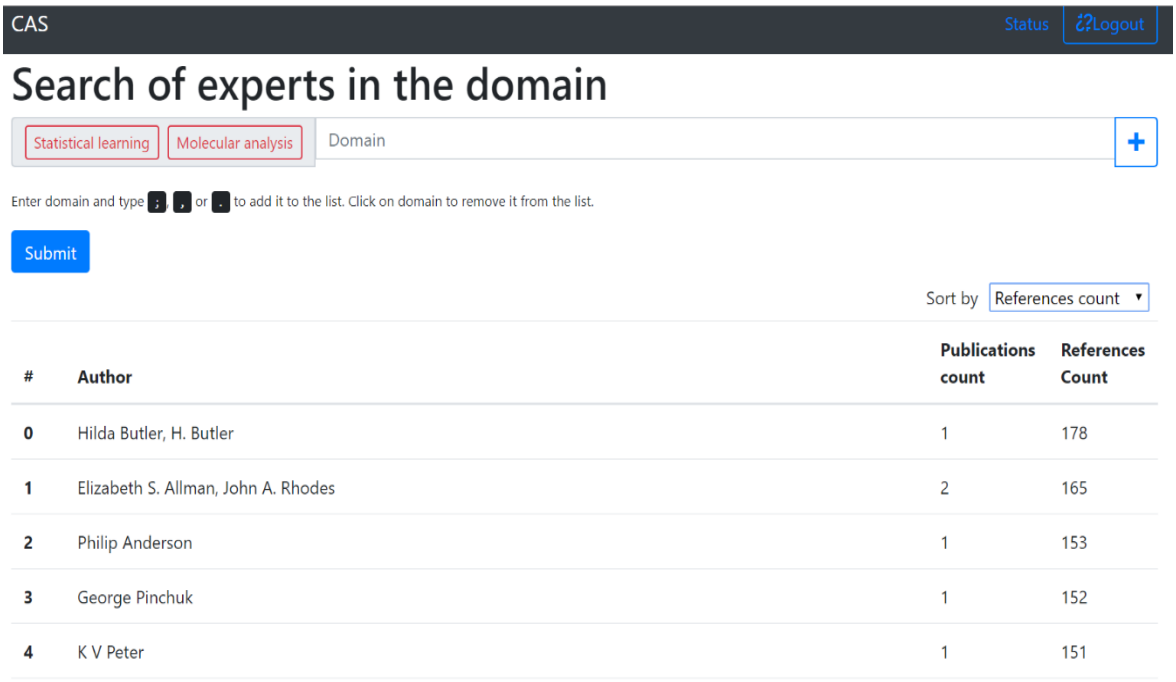

Рисунок 11. Поиск экспертов

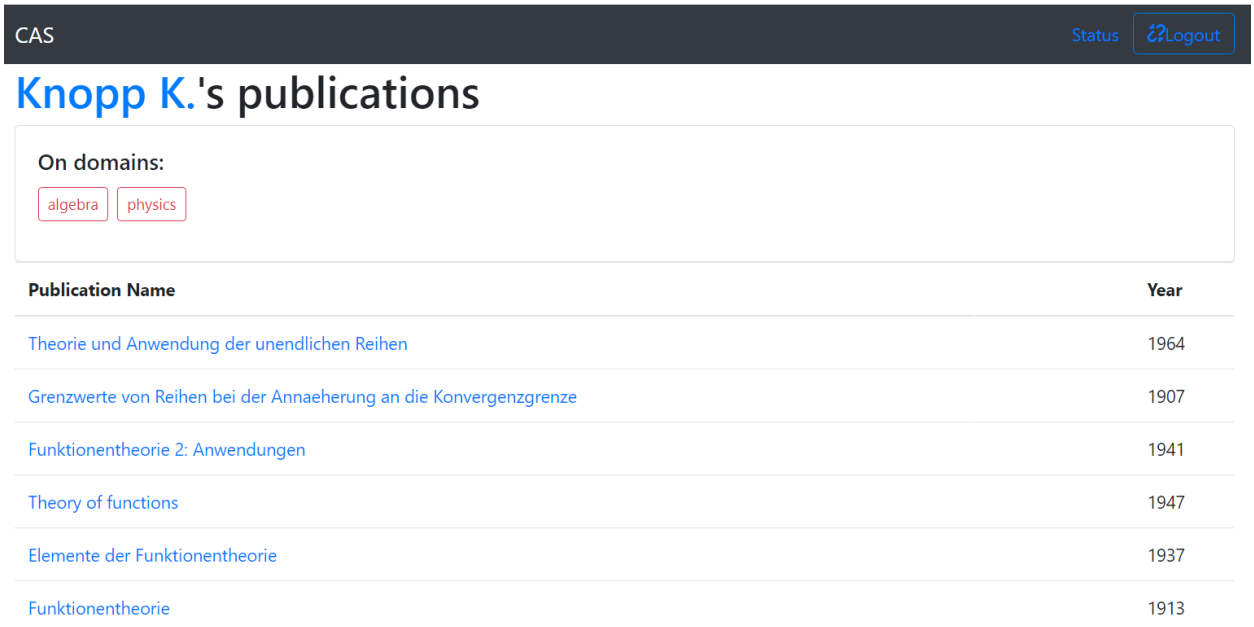

Рисунок 12. Публикации конкретного автора в обозначенных областях

Вывод областей знаний на основе вычисленного индекса популярности.

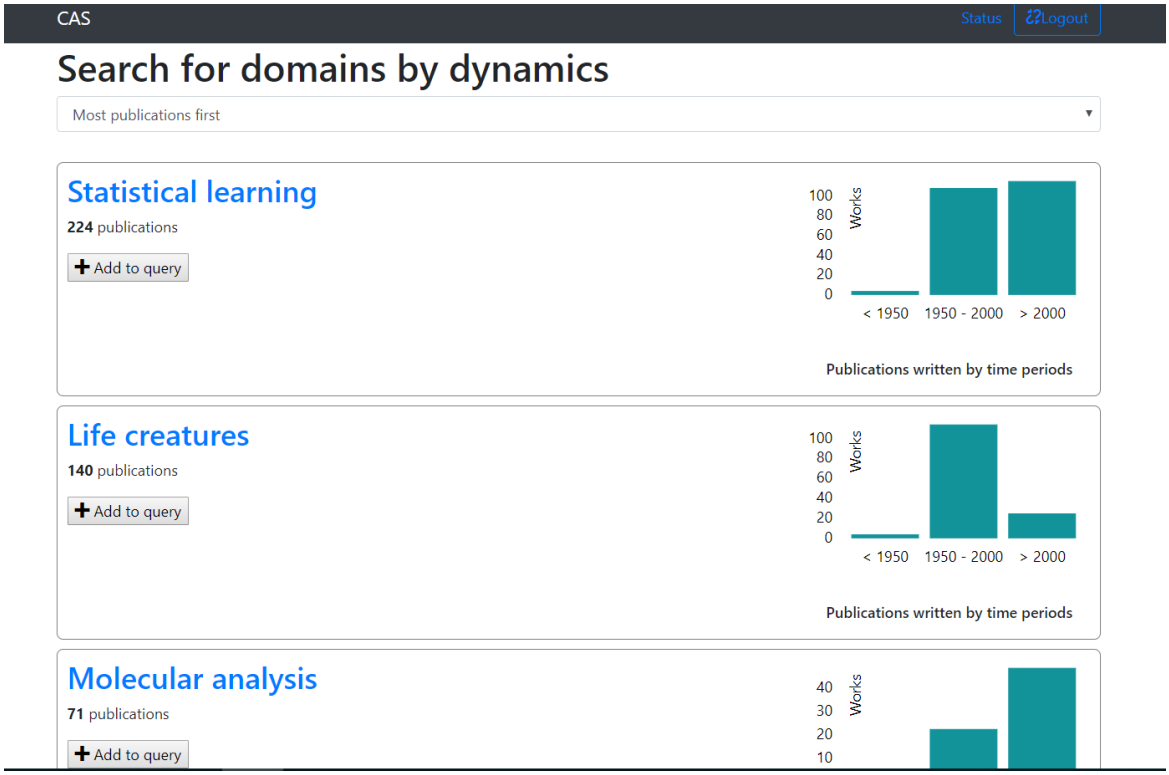

Рисунок 13. Области знаний по их популярности

Кроме просмотра списка можно, нажимая на кнопку «+» справа, выбрать некоторые области и осуществить поиск экспертов в них (рисунок 13.).

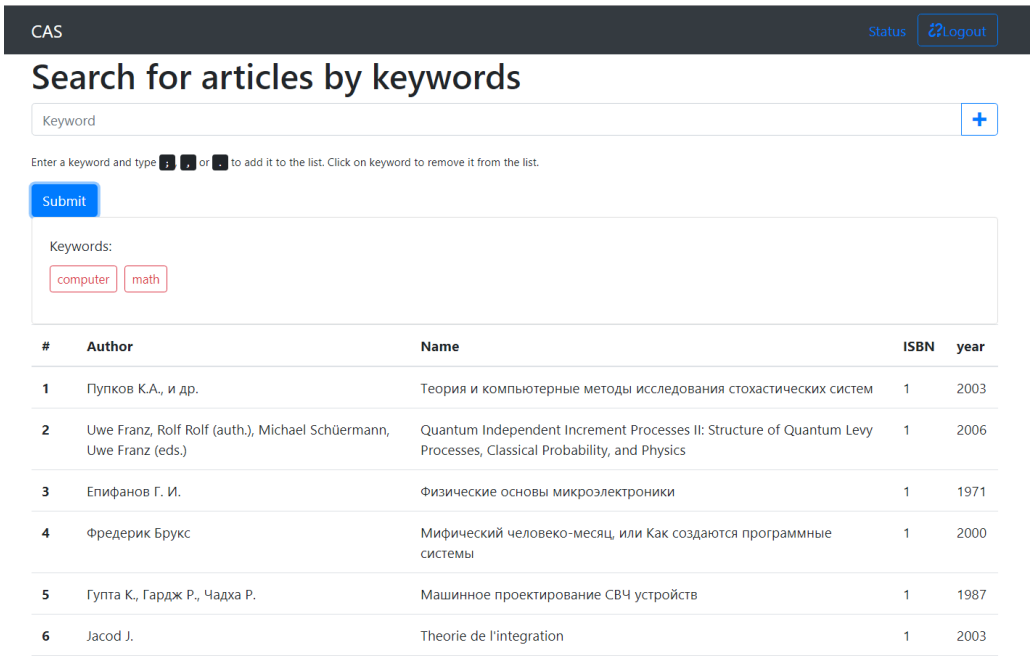

Рисунок 14. Поиск публикаций по ключевым словам

С помощью конструктор запросов пользователь может построить любые другие запросы на языке запросов Cypher (например, по запросу: *match (a) – [r] – (b), return a, b, r limit 100* визуализация результата с помощью спец-библиотеки на рисунок 15).

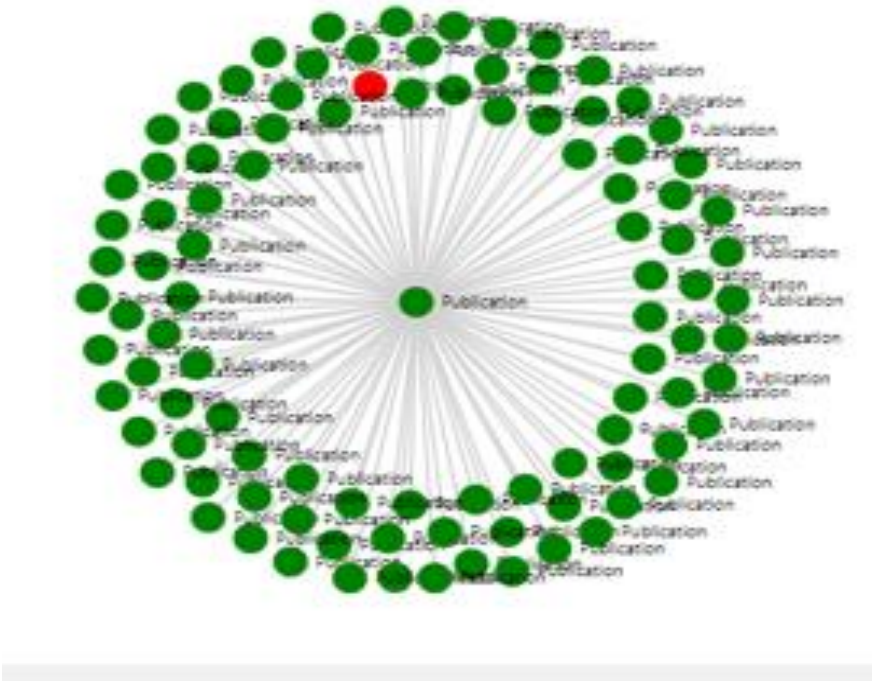

Рисунок 15. Визуализация результата

Заключение

«Система комплексного анализа данных интернет-источников» – это инновационный научно-образовательный проект БГУИР. Результаты выполнения проекта будут использоваться при обучении магистрантов по тематике «Обработка больших объемов информации», подготовке специалистов «Data Scientist», а также для получения экспертных данных при проведении исследовательских работ в университете. В настоящее время над проектом работает коллектив студентов, которые уже получили навыки работы в команде над большим проектом и знания по извлечению данных из интернет-источников, их обработке и анализу с помощью NPL и ML алгоритмов.

Выбранная архитектура СКА позволяет её модернизировать и функционально наращивать в процессе эксплуатации. Так, планируется нарастить компонент «хранилище» и дополнить компонент «библиотека модулей» векторными алгоритмами и алгоритмами нейровычислений, что существенно расширит область применения СКА и позволит получать информацию в интересах сотрудников университета, государственных и частных организаций.

Литература

[1] Data Never Sleeps 6.0 [Электронный ресурс] / Режим доступа: https://www.domo.com/learn/data-never-sleeps-6 Дата доступа: 18.01.2019.

[2] Пилецкий, И. И. Аналитический комплекс анализа данных из открытых интернет источников / И. И. Пилецкий, В. А. Прытков, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 – Minsk, BSUIR, 2018. – P. 193 – 199.

[3] Романов А.А., Пилецкий И. И. Классификация тональности текстовых документов с помощью метода опорных векторов. Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2017 -06 мая 2017.

[4] Чугаинов К. В., Пилецкий И. И. Методы тематической кластеризации новостных статей. Научно-практические исследования №2 (ISSN 2541-9528) – Омск: Дельта, – 2017 с. 295 – 298.

[5] Прытков, В. А. Анализ репозитория университета с использованием графовой базы данных / В. А. Прытков, И. И. Пилецкий, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018. – Minsk, BSUIR, 2018. – P. 177 – 183.

[6] В. Н. Козуб, И. И. Пилецкий. Использование алгоритмов обработки естественного языка и графовых баз данных для построения рекомендательной системы / // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 / editorial board: M. Batura [etc.]. – Minsk, BSUIR, 2018. – P. 274 – 277.

[7] Pagepank algo neo4j. [Электронный ресурс] – Режим доступа: https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/ Дата доступа: 22.01.2019

SYSTEM FOR COMPLEX ANALYSIS OF DATA FROM INTERNET SOURCES

M.P. BATURA
Scientific director of the lab 8.1 of the BSUIR, Doctor of Engeneering Sciences, Full Professor,Member of the International Higher Education Academy of Sciences

I.I. PILETSKI, PhD
Associate Professor of Informatics Department of the BSUIR

V.A. PRYTKOV, PhD
Vice-rector for education BSUIR, Associate Professor

N.A. VOLARAVA, PhD
Head of the Informatics Department of the BSUIR, Associate Professor

V.N. KOZUB
Postgraduate student of the BSUIR,Master of Technical Sciences, Assistant of Informatics Department the BSUIR

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: bmpbel@bsuir.by, ianmenski@gmail.com, prytkov@bsuir.by, volorova@bsuir.by, kozub@bsuir.by

Abstract. This article describes the monitoring tool for open Internet sources in order to identify experts in a certain scientific field, determine the topics of publications, and assess the popularity of publications. It also describes the decisions taken in the construction of complex analytical and results of his work.

Keywords*:* Internet sources, Big data, monitoring, analysis, Machine Learning, Natural Language Processing, Neo4j, Hbase.