АННОТАЦИЯ
диссертационной работы Чинибаевой Т.Т. «Модели и методы управления данными с гетерогенными структурами (BigData)», представленной на соискание степени доктора философии (PhD) по специальности 6D070400 – Вычислительная техника и программное обеспечение.

Развитие современного общества и технологий связано не только с информатизацией новых сфер деятельности человека, но и с повсеместным внедрением технологий исследования и анализа данных для разработки управленческих решений.

Большое внимание уделяется развитию этого вопроса во всем мире, в частности в Казахстане. Важным документом, определяющим основные направления цифрового развития страны, является государственная программа «Цифровой Казахстан», принятая 12 декабря 2017 года. В паспорте проекта указано, что в связи со значительным увеличением объема данных государство поможет создать крупный технологический центр анализа данных и обеспечит надежную работу, сохранность, целостность национальных и государственных информационных ресурсов, в том числе на основе существующих инициатив.

Высокопроизводительные вычислительные системы, ускоряющие процесс обработки, не обладают необходимыми знаниями, полученными посредством анализа данных. Это связано с тем, что при проектировании архитектуры системы не учитывалась проблема совместимости.

Актуальность темы исследования определяется представлением моделей и методов управления большими данными с гетерогенной структурой, используемых для мониторинга и анализа информации, описывающей деятельность научных организаций.

Целью исследования является разработка программного обеспечения для поиска, систематизации, хранения, аннотирования и анализа информации, описывающей публикации ученых в этой области науки, с использованием математических моделей, алгоритмов и корпуса документа.

Объектом исследования являются гетерогенные научные данные.

Предмет исследования - модели и методы управления данными с гетерогенной структурой с целью обеспечения семантической совместимости документов.

Методы исследования. Поставленные в ходе исследования задачи решались методами анализа текстов естественного языка, классификации и программной инженерии. Результаты были представлены аппаратом математической статистики и математической логики.

Научная новизна диссертационного исследования представляется в виде интеллектуальной системы, где применяются разработки автора, а именно, алгоритмы построения онтологии отдельной области научного знания и выделения терминов-пар слов из коллекции текстов с заданным тематическим делением, а также формальное описание запросов к системе с использованием онтологий и языка SPARQL.

На защиту выноситься следующие результаты:

• математическая модель и алгоритм, технологическое и архитектурное решение, разработанные с использованием онтологии для системы анализа, передачи, заполнения и хранения востребованной информации по результатам исследования предметной области в описании результатов научной организации;

• онтология, гарантирующая дополнительные возможности в расчете запросов и эффективной верификации системного кода на всех этапах жизни, а также формальное описание системных запросов с использованием языка SPARQL;

• алгоритмы создания онтологии определенной области научных знаний и выделения пар терминов из наборов текста, отвечающих требованиям темы;

• аналитическая оценка сложности программного обеспечения, созданного с использованием математических моделей.

Теоретическая и практическая значимость работы: научная новизна и практическая значимость исследования высоки. Результаты исследования используются для объединения гетерогенных данных и использования их для дальнейшей обработки.

Апробация работы и публикация. Основные положения и научные результаты работы докладывались и обсуждались на отечественных и зарубежных международных научных конференциях. Диссертационная работа обсуждалась на научных семинарах, организованных кафедрой «Компьютерная инженерия и информационная безопасность» Международного университета информационных технологий, научных семинарах организованной университетом Гачон (Южная Корея, г. Сеул).

Основные результаты, полученные при выполнении диссертационной работы, опубликованы в 12 печатных изданиях, из них 5 статьей опубликованы в изданиях, рекомендованных ККСОН МОН РК, 7 статьей опубликованы в сборниках международной конференций (Казахстан, Южная Корея, Китай, Латвия) 1 статья опубликованы в изданиях, индексируемой базой Scopus, (процентиль 37%).

Структура и объем диссертации. Структура диссертации состоит из введения, четырех глав, заключения, списка использованной литературы и приложения. Общий объем работы составляет 119 страниц, в том числе 40 рисунков, 15 таблиц, библиография из 74 наименований, 2 приложения.

Во введении дается краткий обзор предметной области и освещаются ключевые вопросы в этой области. Обоснована значимость диссертации, сформулированы цель и требования.

Первый раздел посвящен текущему состоянию и месту на рынке технологий больших данных.

Выручка от продаж программного обеспечения и услуг на мировом рынке данных увеличится с 42 миллиардов долларов в 2018 году до 103 миллиардов долларов в 2027 году, при ежегодном темпе роста GAGR 10,48% (рисунок 1).

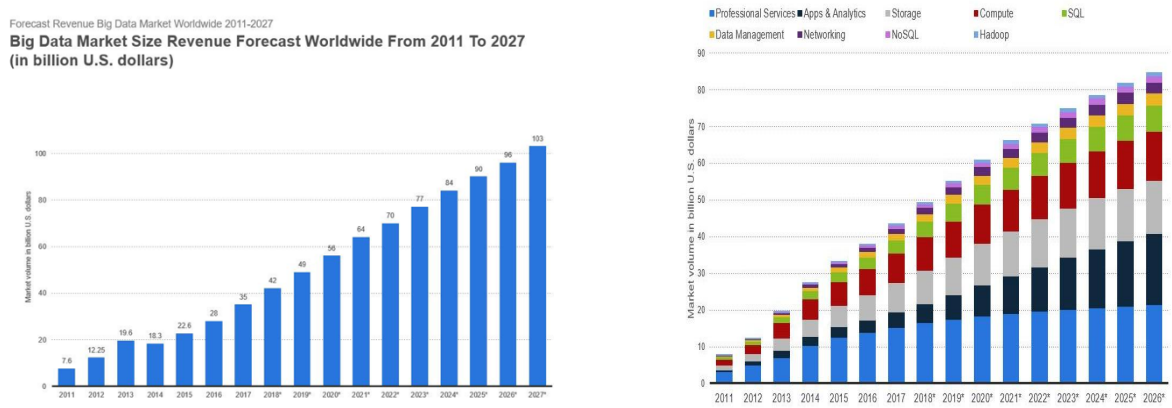

Рисунок 1 - Прогноз рынка больших данных

Технология больших данных играет особую роль в управлении научной информацией. Анализ информационных систем, используемых для решения схожих задач и представленных в сети Интернет, позволил выделить несколько групп систем, большинство из которых являются библиографическими и абстрактными базами данных, в частности Web of Science, Scopus, Google Scholar, российский портал eLibrary.ru . Они в той или иной степени сочетают в себе такие функции, как индексирование и исследование. Часть системы, например, М.В. Система ISTINA МГУ им. М.В. Ломоносова, информационно-аналитическая система Астраханского государственного университета «Результаты научной деятельности», система PURE компании Elsevier осуществляют мониторинг научной деятельности и результатов деятельности организации. Сравнение характеристик крупнейших веб-сервисов, используемых для управления научной информацией в мире, приведено в таблице 1.

В конце первой главы приведены основные недостатки известных на данный момент систем обработки и анализа научных данных. К этим недостаткам можно отнести: сложность ввода данных; сложность и невысокая способность к поиску информации; обращено внимание на использование жестких и неосведомленных моделей обучения, отсутствие гибкости систем.

Результаты исследования были представлены в виде прототипа интеллектуального программного комплекса, обрабатывающего научную информацию с неоднородной структурой.

Таблица 1 - Сравнение характеристик основных веб-сервисов

	№	Название	Уполномоченный орган	Преимущества	Недостатки	Формат данных
Большой веб-сервисы	1	Web of Science	Thomson Scientic	Статьи в системе с 1900 года	Запрос выполняется только по ключевому слову	.TXT
	2	Scopus	Elsever	Охватывает всю предметную область	Запрос выполняется только по ключевому слову	.TXT
	3	Google Scholar	Google	Принимаются во внимание статьи, которые приняты, но еще не опубликованы.	Существуют некачественные и фейковые научные публикации	.TXT
Зарубежные проекты	1	Bibster	Университет Карлсруэ, Амстердамский университет, Дрезденский банк	Выводит данные из системы в формате RDF	Информация загружается в систему через структурированный файл	BibTeX
	2	JeromeDL	Гданьский (Польша) технологический университет, Институт исследований цифровых технологий DERI (Ирландия)	Система может классифицировать и содержать электронную информацию в базе данных	Информация вводится в систему в структурированном виде или вручную. сложные запросы не выполняются, информация вводится вручную	BibTeX, Marc21, Dublin Core
	3	Flink	Амстердамский университет	Определяет область интерфейса ученых на основе ключевого слова	Собирает вручную онтологию требуемой предметной области	FOAF, SWRC
	4	AIR	Университет Вулверхэмптона (Великобритания) и Аликанте (Испания)	Система собирает информацию с веб-страниц в структуре DC	Нет сложной онтологии, моделирующей предметную область	Dublin Core
СБД		Семантикалық дерекқор	Open Sourse	Доступно любому разработчику программного обеспечения	Обеспечение логического соединения	RDF(s), OWL, SPARQL
Систем РФ	1	«ИСТИНА»	Россия, Москва	Доступно любому разработчику программного обеспечения	Обеспечение логического соединения	RDF(s), OWL, SPARQL
	2	Астрахан университетінің «ғылыми қызметінің нәтижелері»	Россия, Астрахань	Доступно любому разработчику программного обеспечения	Обеспечение логического соединения	RDF(s), OWL, SPARQL

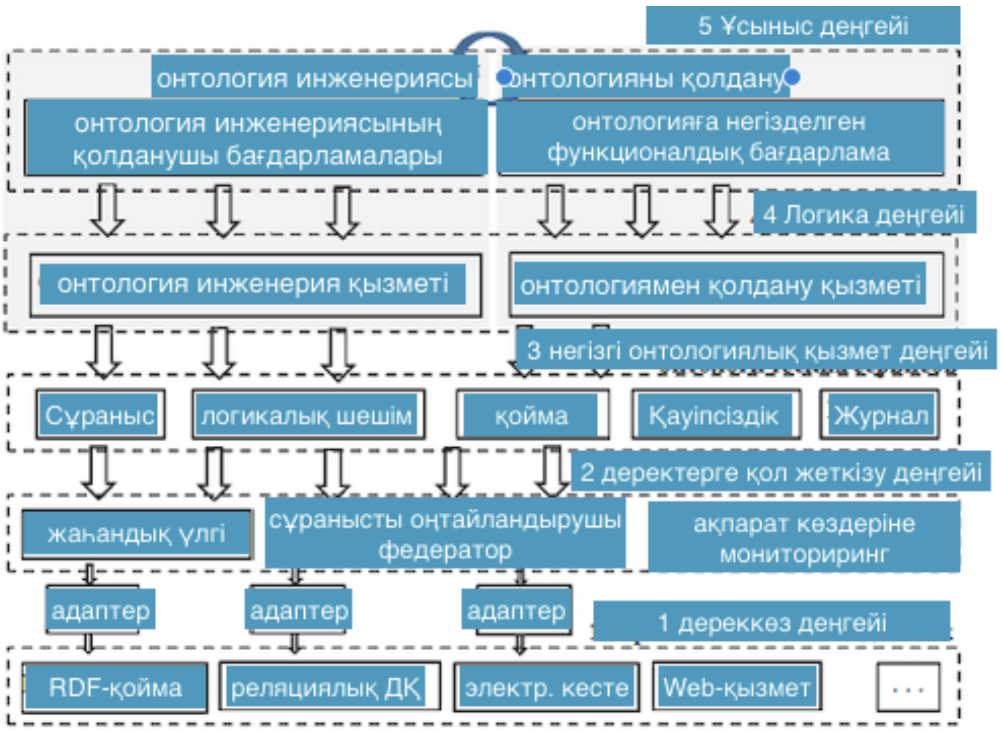

Рисунок 2 - на основе семантических данных общая структура информационной системы

Сводка научно-технической информации, которая является основным прототипом автоматизированной системы, и общая формальная модель сложного организованного вычислительного процесса.

Предположим, что D задана область научных знаний (например, информатика). Пусть I будет набором описаний единиц научно-технической информации в этой области знаний (атомарное измерение). К таким блокам относятся: научные статьи; патенты; отчеты; доклады, прочитанные на конференциях; выписки по бухгалтерскому учету; монографии; учебные пособия и др. авторские работы (рефераты, переводы). Каждый элемент множества I содержит текстовое описание соответствующего объекта.

Основное назначение системы - выполнение поисково-аналитического запроса. Обозначим множество типовых запросов символом Q. Задача задается выражением $q \in Q$ $r1: Q \to 2^I$ характеристиками блока научно-технической информации $I_q \subseteq I$.

Рисунок 3 - последовательность операций

Общая схема системы представлена на рисунке 3 и состоит из следующих моделей:

- отличать термины, описывающие область научных знаний D, от текстового описания научно-технической конференции, посвященной данной области знаний;
- D создание рассмотренной онтологии в области научных знаний;
- скачать данные о результатах научной базы работников;
- установить связь между инстанцией, собранной в сфере образования, и загруженной информацией о результатах научных исследований;
- Выполнение аналитического запроса по полученной информации;
- Общая схема состоит из следующих этапов:
- D выделить термины, описывающие область научных знаний (ключевое слово);
- Разработка онтологии области научных знаний D;
- Загрузка данных из информации различается в зависимости от области науки;
- установить связь между концепцией разрабатываемой онтологии и научными выводами пользователей;
- Образец, который отвечает на запросы, содержит сводку полученной информации.

Следующий шаг - описать каждый шаг. Был получен с помощью семантических, в частности, лингвистических и статистических методов для выделения терминов, описывающих область знаний. При разработке алгоритма различения терминов сформировались следующие определения.

Определение 2.13 В этой диссертации термин - это пара слов, которые описывают документ, который соответствует одной или нескольким его темам.

Формальное решение задачи реализуется следующим образом. Предположим, что многие документы Doc разделены на $r_1, \dots r_n$ Задача состоит из набора терминов. Каждый терм $A \in Terms$ состоит из двух упорядоченных

слов: $A = (A_1, A_2)$ (подробное описание формальной модели дано в разделе 3.1.1).

Определение 2.14 Задача создания онтологии $O = (I, A)$ состоит в выборе эксперта (или группы экспертов) из коллекции текстов документов, интересующей создателя онтологии предметной области, и ее формирования на на основании формальных (автоматизированных) выводов: N_C; Множество имен отношения N_R ; Набор имен N_x examples; финальный набор аксиом введения концептов I=TBox (терминологический раздел онтологии); A=ABox окончательный набор утверждения экземпляра (фактическая часть онтологии).

Построение онтологии состоит из идентификации набора абортов и названий связей, а также экземпляра этих понятий и отношений между ними. Заполнение онтологии развивает понимание и знания и направлено на поиск экземпляров концепций и отношений между ними. Согласно модели в области теории научного познания языков, которая часто используется в автоматизации, мы выявляем различные различия между этими задачами. Выводом терминологического раздела онтологии является их связь с термином как метод решения задачи, совокупность таких понятий, как область исследования, методы, задачи, решения, термины, область исследования состоит из специальных задач, термин используется в методологии, термин является ключевым понятием. Следует отметить, что создание и завершение онтологии в настоящее время важно по следующим причинам.

Онтология состоит из следующих понятий: человек, организация, статья, конференция, проект, а также отношения между ними. В этой онтологии сформированы многие концепции, необходимые для формального описания данных, содержащихся в документах, используемых в общем описательном направлении научной области (данная статья, без заключений конференции). Ключевая концепция онтологии SWRC представлена на рисунке 2.12.

Планируются следующие методы ввода данных:

- регулирование библиографических ссылок;
- настройка загружаемых метаданных (BibTeX, MathML, LaTeX, FinXML);
- заполните поля вручную.

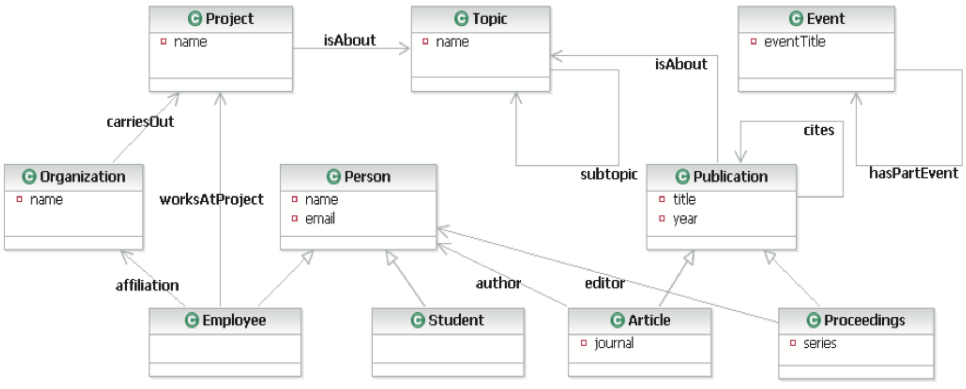

Рисунок 4 - Онтология SWRC (фрагмент)

Правила библиографических ссылок. Получение информации из библиографических ссылок - это задача получения информации из неструктурированного текста. Алгоритм условных случайных полей (CRF), показавший наибольшее влияние на результаты тестирования метода корректировки приведенных библиографических ссылок. Программный пакет FreeCite, разработанный в Университете Брауна в США, использовался в библиотеке CRF ++, которая реализует этот алгоритм.

R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143	"R.Uskenbayeva, ", "T.Chinibayeva", "Model, ", "data ", "integration ", "algorithms ", "of ", "information ", "systems ", "based ", "on ", "ontology ", "Journal ", "of ", "Theoretical ", "and ", "Applied ", "Information ", "Technology ", "E-ISSN ", "1817-3195", "ISSN ","1992-8645 ", "Vol.99 ", "Vol.99 ", "2021 ", "No 09. ", "2125-2143".

```
["R.Uskenbayeva, ",  "T.Chinibayeva "] => "R.Uskenbayeva, T.Chinibayeva";
"09: 2125-2143" => {:volume =>09, spage => 2125, :epage => 2143}.
```

Установление связи между существующей моделью в области образования и информацией, полученной из загруженных текстов, состоящих из результатов научной работы ученых, необходимо для выполнения аналитических требований. От документа, использованного до этого уровня, отличается только объем информации о научной работе сотрудника.

Следующая формула используется в этой диссертации для определения уровня семантического сходства Sim между экземпляром t∈ $t \in N_x^D$ онтологии O_D (термин в области знаний) и экземпляром $e \in N_x^S$ онтологии O_S (например, статьи):

$$Sim(e,t) = sim_{edit}(title(e),t),$$

где $title(e)$– это заголовок статьи, а $sim_{edit}(s_1,s_2) = \frac{1}{1+editDist(s_1,s_2)} - s_1$ равно количеству правок, необходимых для преобразования строки $editDist(s_1,s_2)$

Аналогичная функция линий, соединенных на основе расстояния Левенштейна. Если значение функции $Sim(e,t)$ превышает значение константы C_{sim}, то между экземпляром онтологии и научной статьей устанавливается соединение swrc: isAbout.

Онтологическое действие, связанное со знаниями, позволяет использовать текущую и прошлую апробацию алгоритмов, выполняющих аналитические запросы. В частности, переписывание запроса с использованием онтологии может выполняться автоматически с использованием механизма логического вывода.

Вот пример запроса, который позволяет вам получить публикации 2020 года для разработки программного обеспечения («Программная инженерия и информационная безопасность») для демонстрации синтаксиса языка SPARQL.

```
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIXcs:<http://nauka.iitu.kz/ontologies/computer_science#>
SELECT DISTINCT ?pub
WHERE {
?pub a swrc:Publication.
?pub swrc:year 2020.
?pub swrc:isAbout cs:Software_Engineering and IS.
}
```

Динамика интереса исследователей к тому или иному направлению исследований во времени. Запишем эту заявку так: T = {t_1,…, t_n} за последние 10 лет, сгруппированных по годам, количеству результатов научной работы по заданному направлению. Форматируем в SPARQL.

```
SELECT DISTINCT ?res ?year
WHERE {
?res a swrc:Result .
?res swrc:year ?year .
{ ?res swrc:isAbout t_1 }
UNION { ?res swrc:isAbout t_2 }
...
UNION { ?res swrc:isAbout t_n } .
FILTER ( ?year > 2006 && ?year < 2020 )
}
```

По запросу предоставляется перечень результатов всех научных работ по направлению T за последние 5 лет. Затем вам нужно будет разделить эти данные на годы, используя язык, который вы используете. Эта задача выполняется с использованием стандартной должностной функции.

Список конференций по интересующей сфере. Запишем этот запрос следующим образом: Дан список конференций, относящихся к направлению $T = \{t_1, \dots, t_n\}$. Форматируем в SPARQL.

```
SELECT DISTINCT ?conf
WHERE {
?conf a swrc:Conference .
{ ?conf swrc:isAbout t_1 }
UNION { ?conf swrc:isAbout t_2 }
...
UNION { ?conf swrc:isAbout t_n } .
}
```

Таким образом, мы формулируем следующие утвержденные правила в SPARQL.

Правило 1. Предположим, что область научного знания D или ее O_D онтология, дополнены всеми возможными терминами в предметной области, и даны отношения между ними. Предположим, что в научном поле O_S дана онтология, заполненная данными о результатах научной работы отдельных ученых. Среди этих онтологий есть все возможные связи типа swrc: isAbout, то есть каждый результат научной работы ученого имеет множество особенностей,

характеризующих его предмет. Тогда сходство языка запросов SPARQL и онтологий O_D и O_S позволяет получить гарантированный ответ на запросы.

Математическая модель алгоритма выбора терминов из набора текстов с заданными тематическими разделами.

Предположим, что W - ε - это большинство всех слов, найденных во всех документах, включая пустое слово Doc, а PW - это пара всех упорядоченных слов, то есть PW = W * W. Документ d представляет d: N → W, что соответствует каждому натуральному числу n слов в n-м направлении в данном наборе документов. Цифры без слов (после конца документа). Соответственно, новая строка p задается в указанном абзаце как p: N → W, что соответствует каждому положительному целому числу слова n в n-й позиции. Номер места, где слово не написано, помечается как пустое слово. Все новые строки в наборе помечаем буквой P. r столько документов, которые образуют заголовок, а точнее - r∈2 ^ Doc. Емкость заголовка такая же, как и размер документов в нем. будет определяться. Обозначим большинство данных заголовков R.

Мы также выделяем несколько дополнительных функций:

- $\tau_1: PW \to W, \tau_2: PW \to W$ – пар для многих других слов, в котором пара слов соответствует первому слову (соответствует второму);
- $Freq: PW * Doc \to \mathbb{N} \cup \{0\}$- d ∈Doc функция, определяющая количество пар pϖ∈PW, введенных в документ;
- $Freq: W * Doc \to \mathbb{N} \cup \{0\}$ - $d \in$ введение пары w∈W в документ функция, определяющая число
- $L(d) = |\{n \in \mathbb{N} | d(n) \neq \varepsilon\}|$ - d – длина документа;
- $id(a) = a$ – подобное изображение;

$Av(f, A) = \frac{\sum_{a \in A} f(a)}{|A|}$ - A - среднее значение функции f в последнем множественном числе.

Например, Av (|·|, R) - среднее количество документов в заголовке, $Av(L, Doc)$ - средняя длина документа, $Av(id, A)$ - среднее арифметическое множества $A = \{a_1, \dots, a_k\}$.

Набор данных, сгенерированный для алгоритма, представляет собой таблицу в соответствующих словах, которая находится в документах. Обратите внимание, что перед использованием алгоритма лемматизации - рекомендуется провести лингвистическую обработку документов, в первую очередь необходимо преобразовать словоформы в правильную (по словарю) форму. Например, для существительных в казахском и русском языках такой формой является личное местоимение. Каждый элемент в выходной таблице содержит четыре числа: номер заголовка, номер документа, номер новой строки (абзаца), номер слова. Если слово A стоит перед словом B в абзаце, то в таблице абзац слова A выше, чем абзац слова B. Таблица выбирается по первым трем столбцам. Таким образом, известно только, в каком документе, сколько раз и где это происходит.

Алгоритм состоит из четырех этапов. В каждом из них с помощью некоторых правил выбирается набор M_i и M_{i-1} полученный на предыдущем шаге. На первом этапе производится выборка из множества PW (всех пар слов),

т.е. $M_0 = PW$. Набор M_4 представляет собой пару терминов, которая удовлетворяет всем четырем измерениям.

Первый выбор пар для последующей обработки основан на предположении, что слова, обозначающие термин, значительно ближе (хотя и не обязательно близки) в тексте:

$$M_1 = \{pw \in M_0 | \exists p \in P: |p^{-1}(\tau_1(pw))| \leq MAX_{DIST}\}$$

MAX_DIST-1 - это пара двух слов в одном абзаце среди других общих слов в тексте.

Чтобы база данных была информативной, она должна содержать наиболее часто встречающееся слово в тексте. Чтобы удовлетворить это требование, все наборы используют определенные измерения, которые отличаются от большинства пар M_1, которые встречаются реже, чем MIN_FREQ:

$$M_2 = \{pw \in M_1 | \sum_{r \in R} \sum_{d \in r} Freq(pw, d) \geq MIN\_FREQ\}$$

Характерный размер - это основное измерение алгоритма. Его суть заключается в определении термина, пара должна соответствовать какому-то заголовку.

Вес пары в заголовке. Каждой паре соответствует набор чисел - вес пар в каждой секции. Вес пары pw в заголовке r определяется по следующей формуле:

$$Weight_\tau(pw) = \frac{\sqrt{\sum_{d \in r} \ln\left(\frac{\sqrt{Freq(pw,d)}}{\ln\left(\frac{r}{Av(L,Doc)} + 1\right)} + 1\right)}}{\ln\left(\frac{r}{Av|\cdot|, R}\right)} + 1$$

Соответственно определяется вес слова в заголовке $Weight_\tau(pw)$, а точнее - в приведенной выше формуле знак $Freq(pw, d)$ заменен на $Freq(w, d)$.

Общий вид деятельности. Для выбора сервисов, отвечающих установленным требованиям, выбран следующий общий вид функции $Weight_r$:

$$Weight_r(\bar{x}, \bar{y}, z) = h\big(g\big(f(x_1, y_1), \dots, f(x_k, y_k)\big), z\big)$$

Функция $f(x, y)$ определяет вес пар в документе и зависит от длины документа и количества пар в документе. Затем функция $g(x_1, \dots, x_k)$ используется для определения веса пар, что обеспечивает прямую зависимость от количества документов, найденных в парах. Затем мы используем функцию $h(x, y)$, которая показывает зависимость веса пар в заголовке от мощности относительно заголовка. Для облегчения тестирования был выбран специальный тип функции $Weight_\tau$ в частности, функция - $h(x, y)$ является производной от функции $f(x, y)$. Этот выбор основан на схожей

характеристике этих функций: каждая из них определяет размер заголовка $\big(h(x,y)\big)$ или размер документа $\big(f(x,y)\big)$ и вес числа. пар вошли. Таким образом, окончательное окончательное выражение функции $Weight_\tau$ выглядит следующим образом:

$$Weight_r(\bar{x},\bar{y},z) = f\big(g\big(f(x_1,y_1),\dots,f(x_k,y_k)\big),z\big)$$

Функции f и g должны иметь следующее определение и значение:

$$D(f)\colon x \in [0,+\infty), y \in (0,+\infty), E(f) = [0,+\infty),$$
$$D(g)\colon x_i \in [0,+\infty), i = \overline{1,k}, E(g) = [0,+\infty)$$

определенной области научного знания представляет собой сборник анонсов научных конференций, разработанный автором алгоритм Sonmake для создания онтологии научного знания, разделенной по темам, а также информации из поисковых систем. в интернете. Названный контактными сообщениями (CFP), он использовался в качестве основного источника данных для создания онтологии.

Рисунок 5 - Алгоритмическая схема построения онтологии в определенной области научного знания

Следующим шагом в алгоритме построения онтологии в области научного знания является фильтрация терминов $Terms_1$, которая состоит из двух уровней. На первом уровне фильтрации удаляются некоторые пары, не соответствующие размеру термина. Для этого используются четыре последовательных измерения.

На первом уровне фильтрации удаляются некоторые пары, не

соответствующие размеру термина. Для этого используются четыре последовательных измерения. Пусть $A \in Terms$– член-кандидат, состоящий из двух слов A_1 и A_2 , тогда эти критерии формируются следующим образом:

В онлайн-энциклопедии Википедия есть статья под названием A;

- $\frac{hits("A\ is\ a\ term")}{hits(A)} > C_1$;
- $\frac{hits("A\ is\ a\ concept")}{hits(A)} > C_2$;
- $\frac{hits("A_1\ AND\ A_2")}{\min\ (hits(A_1),(hits(A_1))} > C_3$.

Цель второго уровня фильтрации - удалить пару слов, не относящихся к данной области знаний D. Для этого используются следующие критерии:

$$\frac{hits("A\ AND\ D")}{hits(A)} > C_4$$

Целью следующего шага является различение пар связанных терминов, то есть семантически близких пар, связанных с возможными парами терминов в наборе N_x^D. Нормализованное расстояние Google (NGD) - это общий термин, используемый для определения уровня семантического сходства между двумя терминами. Пусть A и B - термины, а N - общее количество страниц, проиндексированных поисковой системой. Тогда уровень семантического сходства NGD между A и B определяется по следующей формуле:

$$NGD(A,B) = \frac{\max\{\log hits(A), \log hits(B)\} - \log hits("A\ AND\ B")}{\log N - \min\{\log hits(A), \log(B)\}}$$

Следующий уровень алгоритма - создание иерархии терминов. Классический алгоритм формирования концепции иерархии с помощью лингвистических шаблонов, разработанный Херстом, оказался менее эффективным для формирования научной иерархии.

В ходе исследования были разработаны лингвистические шаблоны для формирования концепции иерархии на научном уровне. Основные шаблоны следующие:

$$A\ is * \ keyword * prep(aux)?\ B$$

Классификация терминов по категориям

Чтобы определить небольшой набор классов, к которому принадлежит термин $A \in Terms_2$ уровень каждого класса $C \in N_C^D$ рассчитывается по следующей формуле:

$$score(A,C) = \frac{hits(“A\ is\ a\ C”)}{hits(A)}$$

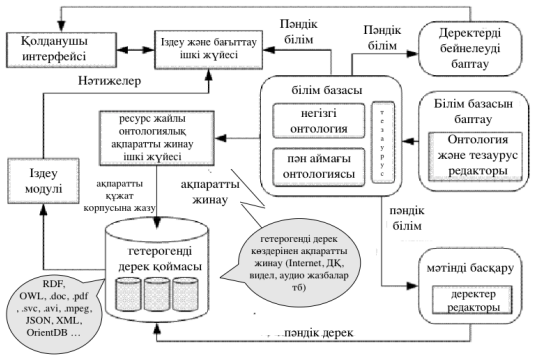

Рисунок 6 - Архитектура системы

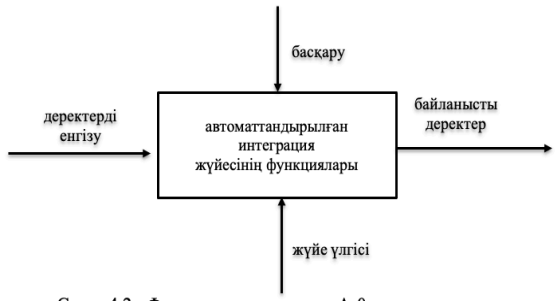

Рисунок 7 - Функциональная модель

Для реализации методов и алгоритмов, описанных в диссертации, было разработано программное обеспечение для извлечения необходимой информации из разнородных источников и представления их в виде связанных данных.

Заключение. В диссертации описаны методы и приемы построения системы управления научной информацией. Теоретическая основа действия - онтология. Версия, предложенная автором системы, включает в себя такие шаблоны, как выполнение запроса, онтологию и результаты научной работы ученых, шаблон для загрузки информации, шаблон, формирующий формальную модель в сфере образования.

В ходе работы над диссертацией были получены следующие основные результаты.

- На основе изучения предметной области разрабатываются технологические и архитектурные решения на основе онтологии, математических моделей и алгоритмов для создания систем выводов, хранения и пополнения информации, описывающей результаты научной организации. Используя онтологию и язык SPARQL, предоставляется формальное описание системных запросов, которое обеспечивает дополнительные функции и вычислительные гарантии эффективности системного кода на протяжении всего его жизненного цикла.

- Разработан алгоритм создания онтологии в области научных знаний, основанный на использовании информации в поисковых системах в Интернете, выявлении терминов в рекламе научных конференций. Получена аналитическая оценка сложности реализации его программы.

Апробация работы:

1. R. Uskenbayeva, Y. Chinibayev, A. Kassymova, T. Temirbolatova, K. Mukhanov. Technology of integration of diverse databases on the example of medical records//Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS 2014) - Gyeonggi -do, Korea, 2014. P 282-285. ISSN: 2093- 7121.

2. R.Uskenbayeva, T.Temirbolatova, Young Im Cho, Z.Uskenbayeva, G.Bektemyssova, A. Kassymova. Recursive decomposition as a method for integrating heterogeneous data sources//Proceedings of the 15th International Conference on

Control, Automation and Systems (ICCAS 2015). – Busan, South Korea. October 13-16, 2015 – P.2076-2079. ISSN: 2093 - 7121

3. Р.К. Ускенбаев, Т.Т.Темірболатова, А.Б. Касымова. Бұлттық есептеуде mapreduce технологиясымен үлкен деректерді өңдеу - // Вестник КазНТУ имени К.Сатпаева No5 (111). – 2015. С.50 - 53. ISSN 1680- 9211

4. Р.Ускенбаева, Г. Бектемысова, Т.Темірболатова. Интеграция больших неоднородных данных с использованием языка R и HADOOP - Вестник КазАТК - №4 2015-11-01

5. Ускенбаева Р.К., Аманжолова С.Т., Темірболатова Т.Т. Анализ и локализация инцидентов снижения работоспособности распределенных вычислительных систем. Труды международного форума «инженерное образование и наука в XXI веке: проблемы и перспективы», посвященного 80-летию Каз НТУ им. К.И. Сатпаева

6. T. Temirbolatova, D. Beisenov Automatic asynchronous exchange of business object between heterogeneous systems - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia

7. T. Temirbolatova, A.Khamitov, A. Keldybay, T.Sembayeva Manage different-structured Big Data - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia

8. Temirbolatova T. Jarmukhambetov Y., Temirbolatova U. The method of extracting semantic meta descriptions from databases//2nd International scientific conference «Information Technologies in Science &Industry» International IT University, May 19, 2016 Almaty, Kazakhstan. ISBN 978-601-7407-33-9

9. T. Chinibayeva Security semantic database problems // Herald of the Kazakh-british technical university ISSN1998-6688. V INTERNATIONAL CONFERENCE "DIGITAL TECHNOLOGY IN SCIENCE AND INDUSTRY - 2019» (DTSI-2019), 10th Anniversary INFORMATION TECHNOLOGY INTERNATIONAL UNIVERSITY Vol.16, No.3 (2019), pp. 168-174

10. R.Uskenbayeva, T.Chinibayeva. Algorithm for the construction of an ontology in the field of scientific knowledge//The Bulletin of Kazakh Academy of Transport and Communications named after M. Tynyshpayev ISSN 1609-1817. Vol. 107, No.4 (2018), pp. 259-266

11. R.Uskenbayeva, T.Chinibayeva. Method of extracting meta description from databases//Herald of the Kazakh-british technical university ISSN1998-6688. Vol.15, No.4 (2018), pp. 116-123

12. R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143