АННОТАЦИЯ
диссертационной работы Чинибаевой Т.Т. «Модели и методы управления данными с гетерогенной структурой (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.

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

Уровень научности исследуемой темы. В течение последних пяти-десяти лет ученые из ближнего и дальнего зарубежья вносили свой вклад в изучение больших данных. В частности: А.Ф. Тузовский, Л.В. Найханова, А.Н. Бездушный, В.А. Серебряков И С Михайлов, Ю.А. Загорулько, Дитмар Байер, К. Шахгельдян, Н. Гуарино, Н. Ной, М. Эриг, А. Маэдче, Йонг Им Чо и отечественные авторы: А.А. Куандыков Р.К. Ускенбаева Т.Г. Балова, А.А. Шарипбаев, И.Т. Утепбергенов, Р.Р. Мусабаев, У.А. Тукеев, Н.К. Мукажанов. По результатам научного анализа проблемы объединения данных с разнородными структурами, собранными из разных источников, знания в этой предметной области бессистемны, а отсутствие единого подхода к управлению большими данными определяет структуру, цель и задачи исследования.

Целью исследования является разработка моделей и методов управления научной информацией с гетерогенной структурой.

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 диссертации основана на заимствовании терминов из анонсов научных конференций, а также разработке нового алгоритма создания онтологии для конкретной области научных знаний с использованием информации, полученной из поисковых систем в Интернете. Математически обоснована оценка вычислительной сложности его реализации. Отличительные особенности разработанного алгоритма: автоматический подбор терминов по области обучения; возможность использования без изменения алгоритма создания онтологий других областей научного знания; не требует ручного труда специалиста. Также был разработан алгоритм выделения пар терминов из набора текстов по заданной теме. В отличие от других классических алгоритмов, алгоритм разделения пар терминов показал высокую эффективность при сравнении текстов с помощью классификации и кластеризации. Математически доказано, что оценка сложности расчета и основная функция веса термина в рубрике соответствуют предъявляемым к нему требованиям.

Для защиты диссертации рекомендуются следующие результаты.

- По результатам изучения предметной области при описании результатов научной организации разработка алгоритмов, используемых при работе с большими данными с неоднородной структурой с использованием онтологий;
- формальное описание системных запросов, созданных с использованием языка SPARQL, которое предоставляет необходимую информацию в онтологии;
- Доказано, что веса заголовков в заголовках, входящих в состав алгоритмов создания онтологий отдельных областей научного знания и выделения пар терминов из текстовых коллекций, соответствуют требованиям базовой функции;
- Аналитическая оценка сложности разработанного программного обеспечения.

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

Апробация работы и публикация.

Основные положения и научные результаты диссертации докладывались и обсуждались на отечественных и зарубежных международных научных конференциях: The 14th International Conference on Control, Automation and Systems, ICCAS 2014 (Южная Корея, Бусан, 2014); The 15th International Conference on Control, Automation and Systems, ICCAS 2014 (Южная Корея, Гуанджоу, 2015).

Диссертационная работа обсуждалась на научных семинарах, организованной кафедрой «Компьютерная инженерия» Международного университета информационных технологий и в научных семинарах организованной факультетом «Компьютерная инженерия» университета 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, Латвия) 1 статья опубликованы в изданиях, индексируемой базой Scopus, (процентиль 36%).

Структура и объем диссертации. Структура диссертации состоит из введения, четырех глав, заключения, списка использованной литературы и приложения. Общий объем работы составляет 105 страниц, в том числе 38 рисунков, 11 таблиц, библиография из 77 наименований, 3 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. Для получения достоверных показателей эффективности и направленности исследовательской деятельности ученого необходим инструмент, позволяющий отдельным структурным подразделениям предоставляющий сбор, анализ трудов ученого.

Таблица 1 - Инструменты и методы управления научной информацией

Методы управления научной информацией	Преимущества	Недостатки
Количественные выводы, основанные на информации отчета об исследовании	подсчитывает вручную и записывает числовой индекс	качество работы не учитывается
Экспертное заключение материалов конференции и журнала	Эксперт ведет качественную и содержательную работу	Статьи публикуются на разных языках
Анализ основных статей	Снижение нагрузки за счет аннотации	Информация, не включенная в аннотацию, будет исключена
Поиск по ключевым словам	В электронной версии отличается большой объем текстовой информации	фактический спрос не покрывается

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 2.

В заключение первой главы перечисляются основные недостатки известных на настоящее время систем обработки и анализа научных данных, которые могли бы рассматриваться как возможные решения основной задачи. К числу таких недостатков относятся: сложность ввода данных; сложность и малые возможности поиска информации; использование жестких и малоинформативных моделей области знания, нехватка гибкости систем; направленность на обработку информации из Интернет, а не на полуавтоматический ввод пользователем; недостаточное внимание к интеллектуализации алгоритмов загрузки, обработки и поиска информации.

Таблица 2 - 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 дереккор	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	«Итоги научной деятельности» Астраханского университета	Россия, Астрахань	Доступно любому разработчику программного обеспечения	Обеспечение логического соединения	RDF(s), OWL, SPARQL

Во втором разделе представлены архитектурные и технологические решения, используемые в системе автоматического управления научной информацией.

Пусть задана область научного знания D (например, «информатика»). Пусть I – множество описаний единиц (атомарных гранул) научно-технической информации в рамках этой области знания. К таким единицам относятся: научные статьи; патенты; отчеты; доклады на конференциях; тезисы докладов; монографии; учебные пособия и иные авторские разработки (рефераты, переводы). Каждый элемент множества I представляет собой некоторое текстовое описание соответствующего объекта. Основной целью системы является выполнение поисково-аналитических запросов, примеры которых представлены выше. Обозначим множество типовых запросов символом Q. Задача состоит в построении отображения $r1: Q \to 2^I$, которое сопоставляет запросу $q \in Q$ подмножество описаний единиц научно-технической информации $I_q \subseteq I$. В диссертации предлагаются методы и средства решения поставленной задачи, которое включает следующие пять этапов:

• выделение терминов, которые характеризуют область научного знания D, из текстовых описаний научно-технических конференций, посвященных этой области знания;

• построение онтологии рассматриваемой области научного знания D;

• загрузка данных о результатах научной деятельности сотрудников;

• установление связей между загруженной информацией о результатах научной деятельности и экземплярами построенной онтологии области знания;

• выполнение аналитических запросов к полученной информации.

Общая архитектура разработанной системы представлена на рис. 1.

Q. Задача задается выражением $q \in Q$ $r1: Q \to 2^I$ характеристиками блока научно-технической информации $I_q \subseteq I$.

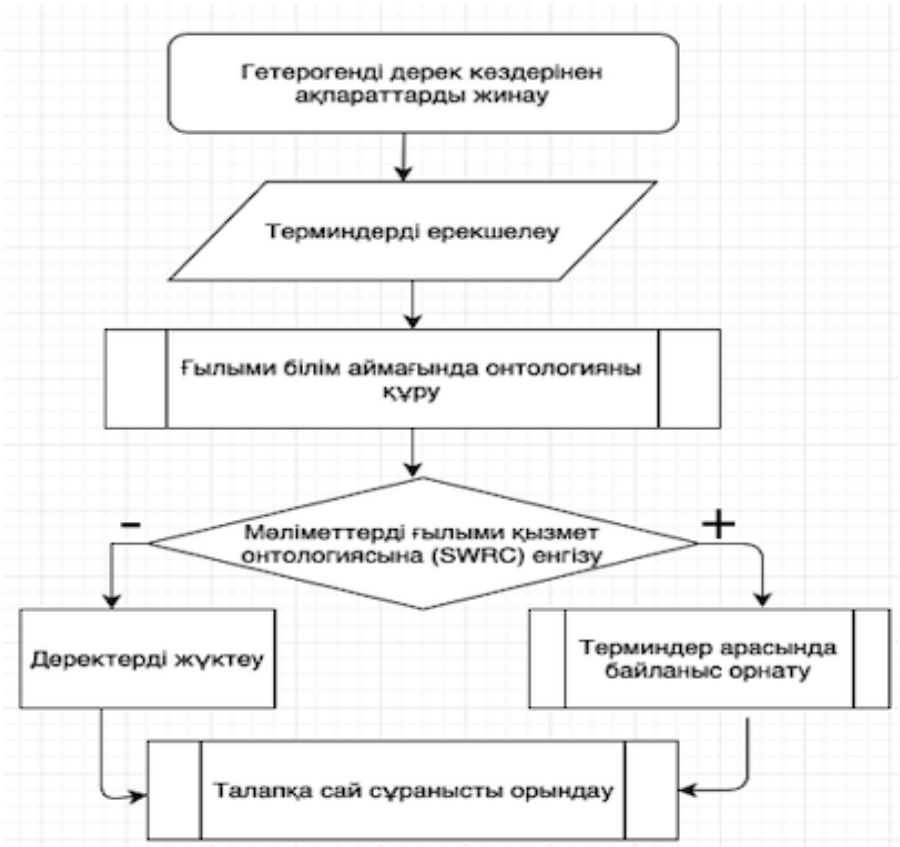

Рисунок 1 - Общая архитектура системы управления научной информацией

В соответствии с требованиями к системе в разработанном автором ее прототипе предполагается следующие три возможных способа ввода данных:

- разбор библиографических ссылок;
- разбор BibTeX-записей (BibTeX, MathML, LaTeX, FinXML);
- заполните поля вручную.

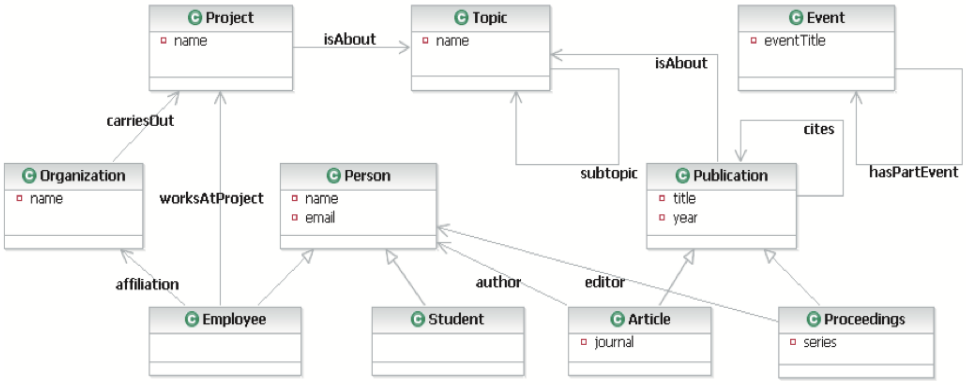

Рисунок 2 - Онтология SWRC (фрагмент)

Разбор библиографических ссылок производится с помощью программного комплекса FreeCite, разработанного в университете Брауна, США. Этот комплекс использует библиотеку CRF++, реализующую алгоритм классификации Conditional Random Fields. В ходе работ по подготовке диссертации код программы FreeCite был модифицирован в целях поддержки

русского и казахского языка. Было проведено также ее дополнительное обучение на размеченных библиографических ссылках на русском языке. В результате работы алгоритма из входной строки выделяются необходимые поля.

```
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.

Выполнение аналитических запросов к данным обеспечивается в процессе взаимодействия конечного пользователя системы с программной реализацией модели, описывающей область знания. Такая модель, построенная автором, включает как общую информацию об области знания, так и данные о результатах научных исследований сотрудников организации в этой области.

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 rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIX cs:<http://nauka.iitu.kz/ontologies/computer_science#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?pub
WHERE {
?pub a swrc:Publication.
?pub swrc:year 2020.
?pub swrc:isAbout cs:Software_Engineering.
}
```

Описание алгоритма выбора терминов из набора текстов с заданными тематическими разделами. Пусть W – множество всех слов, которые встречаются во всех документах заданной коллекции Doc, включая ε – пустое слово, а PW – множество всех упорядоченных пар слов, то есть $PW = W \times W$. Определим документ d как отображение d: N $\rightarrow W$, которое сопоставляет каждому натуральному числу n слово, стоящее на n-той позиции в данном документе коллекции. Номера позиций, на которых нет слов (после конца документа), отображаются в пустое слово. Аналогично определим абзац p как отображение p: N $\rightarrow W$, которое сопоставляет каждому натуральному числу n слово, стоящее

на n-той позиции в данном абзаце. Номера позиций, на которых нет слов, отображаются в пустое слово. Обозначим множество всех абзацев в коллекции через P. Определим рубрику r как произвольное подмножество множества документов, а именно – $r \in 2\ Doc$. Мощность рубрики, как количество документов в ней, будем обозначать через $|r|$. Обозначим множество всех заданных рубрик через R.

Определим еще несколько вспомогательных отображений:

• $\tau 1 : PW \to W$, $\tau 2 : PW \to W$ - проекции пары на множество слов, которые сопоставляют паре первое (соответственно, второе) слово пары;

• $Freq$: $PW \times Doc \to \mathrm{N} \cup \{ 0 \}$ - функция, которая определяет число вхождений пары $pw \in PW$ в документ $d \in Doc$;

• $Freq$: $W \times Doc \to \mathrm{N} \cup \{ 0 \}$ - функция, которая определяет число вхождений слова $w \in W$ в документ $d \in Doc$;

• $L(d) = |\{ n \in \mathrm{N} \mid d(n) =/ \varepsilon \}|$ - длина документа d;

• $id(a) = a$ - тождественное отображение;

• $Av(f, A) = \frac{\sum_{a \in A} f(a)}{|A|}$ - среднее значение функции f на конечном множестве A. Например, $Av(|\cdot|, R)$ - среднее количество документов в рубрике, $Av(L, Doc)$

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

Алгоритм извлечение терминов-пар включает четыре этапа. На каждом из них с помощью некоторого правила выбирается подмножество M_i множества M_{i-1} , полученного на предыдущем шаге. На первом этапе выбор производится из множества PW (всех пар слов), то есть $M_0 = PW$. Множество M_4 и есть термины – пары, которые удовлетворяют всем четырем критериям.

Алгоритм построения онтологии области научного знания на основе коллекции анонсов научных конференций, разделенных на рубрики, а также информации из поисковых систем в Интернет. В качестве основного источника данных для построения онтологии используются анонсы конференций, называемые в научной среде call for papers (CFP). Этот подход обладает важными достоинствами. В частности, он позволяет получить достаточно надежную, актуальную и полную информацию об области научного знания.

Для определения степени семантической близости между двумя терминами используется широко распространенная мера Normalized Google Distance (NGD). Пусть A и B – термины, а N – общее число страниц, индексируемых поисковой системой. Тогда степень семантической близости NGD между A и B определяется по формуле:

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

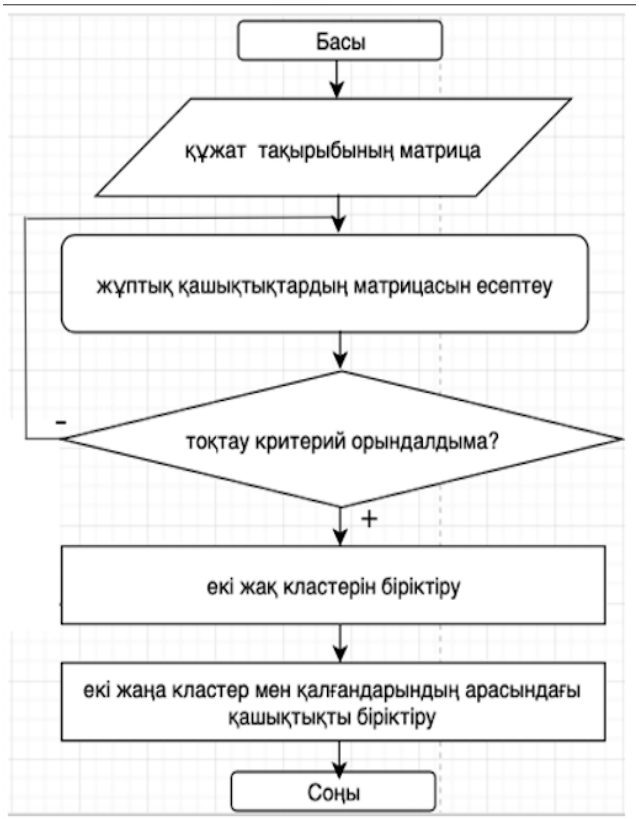

Рисунок 3 - Алгоритм построения онтологии в области научного знания

Следующим этапом алгоритма является построение иерархии терминов. Классический алгоритм построения иерархии понятий с помощью лингвистических шаблонов, разработанный Херст, оказывается неэффективным для построения иерархии научных направлений. В рамках настоящей работы специально для решения этой задачи были разработаны лингвистические шаблоны. Основной шаблон выглядит как

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

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 науки.

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