МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра материаловедения, технологии
и управления качеством

АНАЛИЗ НАУЧНЫХ ПУБЛИКАЦИЙ С ПРИМЕНЕНИЕМ МЕТОДА «БОЛЬШИХ ДАННЫХ»

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

магистранта 2 курса 207 группы
направления 22.04.01 «Материаловедение и технологии материалов»
профиль «Менеджмент высокотехнологичного инновационного производства и бизнеса»
факультета нано- и биомедицинских технологий

Главновой Екатерины Алексеевны

Научный руководитель		
доцент, к.т.н.		И.В. Маляр
должность, уч. степень, уч. звание	подпись, дата	инициалы, фамилия

Зав. кафедрой		
профессор, д.ф.-.м.н.		С.Б. Вениг
должность, уч. степень, уч. звание	подпись, дата	инициалы, фамилия

Саратов 2019

Введение. Сегодня начинают терять актуальность прежние технологии обработки и анализа данных вследствие того, что объём данных стремительно увеличивается. Вследствие этого появляется необходимость в новых технологиях и методах её обработки. Для рассмотрения данной проблемы был введён термин Большие Данные — совокупность подходов, инструментов и методов обработки данных огромных объёмов. Необходимость обрабатывать растущие объемы данных неизбежно меняет наше восприятие мира и вещей, из которых он состоит [1].

Человеческий разум сам по себе не приспособлен для восприятия больших массивов разнородной информации, не способен улавливать более двух-трех взаимосвязей даже в небольших выборках. Но и традиционная математическая статистика не всегда справляется с рядом задач. Она оперирует усредненными характеристиками выборки, которые часто являются фиктивными величинами. Поэтому методы математической статистики могут быть полезными для проверки заранее сформулированных гипотез.

Современные технологии Data Mining перебирают информацию с целью автоматического поиска шаблонов (паттернов), характерных для каких-либо фрагментов неоднородных многомерных данных. В отличие от оперативной аналитической обработки данных в Data Mining формулирование гипотез и выявления необычных шаблонов происходит автоматизировано.

Актуальность работы заключается в том, что совокупность большого количества данных, накопленных за несколько лет, требует новых методов обработки, классификации и анализа, что в дальнейшем может стать источником дополнительной ценной информации, а именно, сведений о закономерностях, тенденциях или взаимозависимостях между данными**.**

Практическая значимость работы: направлена на повышение эффективности релевантного поиска новой информации из научных статей**.**

Цель работы – анализ методов обработки «Больших данных» для эффективной классификации научных публикаций.

Для достижения цели были поставлены следующие **задачи**:

а) рассмотрение понятия «Большие Данные» и методов их анализа, выбор оптимального метода для достижения цели;

б) обработка данных и построение дерева классификаций;

в) анализ полученных результатов.

Дипломная работа занимает 77 страниц, имеет 13 рисунков, 1 формулу и 3 таблицы.

Обзор составлен по 26 информационным источникам.

Магистерская работа содержит введение, 3 раздела, которые включают теоретическую и практическую части, заключение и приложение.

Приложение А содержит список научных публикаций.

Разделы включают в себя:

1 – Большие Данные.

2 – Анализ Больших Данных.

3 – Применение метода Data Mining.

Раздел 3 содержит практическую часть:

3.2 – Обработка данных.

3.3 – Построение дерева классификации.

3.4 – Описание построенного дерева.

3.5 – Выводы.

Основное содержание работы. Во введении обоснована актуальность работы, сформулированы цель и задачи для достижения поставленной цели.

Первый раздел магистерской работы содержит:

Рассмотрение предпосылок появления Больших Данных.

Тенденции стремительного роста информации привели к технологиям, получившие название «Большие Данные» (Big Data). Направление сформировалось из-за многократного увеличения количества информации. Многообразие форм данных, от структурированных документов и таблиц до изображений, геолокаций, аудио- и видеозаписей, требует новых способов их анализа.

Рассматривается определение Больших Данных. Большие Данные — совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети.

В качестве определяющих характеристик для Больших Данных отмечают «три V»:

- объём (volume, в смысле величины физического объёма),

- скорость (velocity, означающее скорость прироста и необходимость высокоскоростной обработки и получения результатов),

- многообразие (variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных).

Данные делятся на три уровня по структурированности:

Уровень первый – это структурированные данные, которые могут быть представлены отделимыми и заранее определенными полями, в которых находятся биты, имеющие различную семантику.

Уровень второй – это полуструктурированные данные. Данные такого типа имеют структурные разделители, но не могут быть представлены в виде таблицы из-за отсутствия части атрибутов у разных данных.

Третий уровень – неструктурированные данные. В них входят тексты, записанные символами различных языков, записи звуков, неподвижные изображения, видеофайлы, сообщения электронной почты, презентации и другая информация вне выгрузок баз данных [2].

Так же в разделе описываются примеры использования Больших Данных. Большей популярностью технологии Больших Данных пользуются в банковской сфере и телекоме, также востребованы в сфере добывающей промышленности, энергетике, ритейле, в логистических компаниях и госсекторе.

Второй раздел выпускной квалификационной работы содержит описание методов анализа Больших Данных. Существует множество разнообразных методик анализа массивов данных, в основе которых лежит инструментарий, заимствованный из статистики и информатики. Чем более объемный и диверсифицируемый массив подвергается анализу, тем более точные и релевантные данные удается получить на выходе. Одним их таких методов является Data Mining.

Data Mining – это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Суть и цель технологии Data Mining: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей [3].

Этапы Data Mining включают в себя:

1. Подготовка данных;

2. Извлечение паттернов (шаблонов);

3. Интерпретация результатов.

Рассмотрены задачи анализа данных, которые включают в себя:

1. Классификация (Classification) – системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

2. Кластеризация (Clustering). Задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

3. Ассоциация (Associations). В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

4. Последовательность (Sequence), или последовательная ассоциация (sequential association). Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени.

5. Прогнозирование (Forecasting). Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем. Таким образом, решение задачи прогнозирования требует некоторой обучающей выборки данных [4].

В третьем разделе магистерской работы описана практическая часть, которая включает в себя применение метода Data Mining для научных публикаций СГУ по сотрудникам факультета нано- и биомедицинских технологий за 2017-2018 год. Для обработки были использованы 358 научных публикаций, которые представлены в приложении А полного текста работы.

В результате обработки статей было выделено 7 групп (атрибутов): биологические объекты; создание образцов; не биообъекты; пленки и поверхность; свойства; общие слова; явления и воздействия.

Для классификации полученных групп был использован алгоритм дерева классификации. *Дерево классификации* – структура данных, в процессе обхода которой в каждом узле в зависимости от проверяемого условия принимается определенное решение – перемещение по той или иной ветке дерева от корня к «листьевым» (конечным) вершинам. В «листьевой» вершине дерева содержится искомое значение интересующего атрибута [5].

Любая классификация производится на основе каких-либо признаков. Для того чтобы классифицировать статьи прежде всего необходимо определить значения выбранных признаков (групп) для этих статей. Для использования математических методов признаки, как правило, должны иметь количественное

выражение. Количественным выражением признаков являются частоты появления этих слов в статьях.

Для построения дерева классификации будут использованы группы слов (описаны в пункте 3.2), которые представлены в 333 статьях.

На первом шаге необходимо найти наилучшее разбиение выборки на две части R1(j, s) = {x | xj < s} и R2(j, s) = {x | xj > s}. Найдя наилучшие значения j и s, создаем корневую вершину дерева – явления, в которую входят все статьи.

Для каждой из подвыборок повторим процедуру, построив дочерние вершины для корневой, и так далее. Каждый узел дерева содержит условие ветвления по одному из атрибутов. Для количественных атрибутов каждый узел имеет всего два ветвления – если значение атрибута меньше определенного значения, то одна ветвь, если больше или равно – то другая.

Для осуществления разбиения выборки на каждом шаге необходимо задать критерий информативности – энтропию.

В теории информации и в математической статистике энтропия – это мера неопределенности, неупорядоченности. Энтропия необходима для того, чтобы отразить способ структурирования знаний в получении узлов дерева, а так же сформулировать алгоритм, который сравним с известными алгоритмами.

Результаты энтропии представлены на диаграмме (рисунок 1).

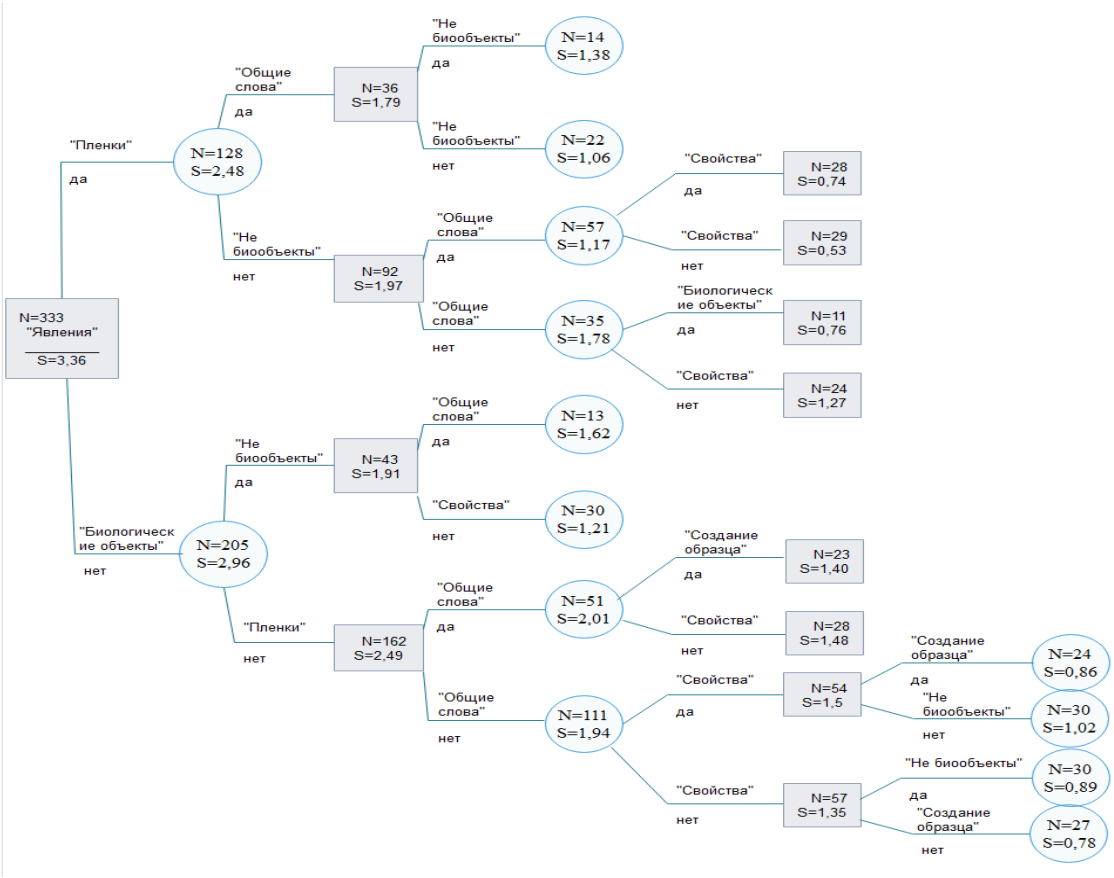

Рисунок 1 – Дерево классификации

Так же было проведено описание построенного дерева. Которое включает в себя: общее описание дерева; описание частных случаев в дереве; влияние частотности на положение; типичные статьи для листьев.

1. При разбиении необходимо, чтобы все наборы данных были однородны в плане значений классов. Для оценки однородности необходимо определить количественный критерий. Дерево классификации позволило провести разбиение статей на группы с использованием количественного критерия, что в дальнейшем должно облегчить релевантный поиск информации и выбор статей, необходимых в определенной области.

2. Существуют погрешности вызванные ограничениями, наложенными на высоту дерева и минимальный прирост энтропии. Так при ограничении высоты дерева есть вероятность, что не все статьи будут классифицированы или классифицированы, укрупнено, то есть в одной статье могут быть слова из разных выделенных групп. Например, в статье: биосенсор для обнаружения

бактериофагов на основе сверхвысокочастотного резонатора, слово «**био**сенсор» относится к группе биологические объекты, а слово «сверхвысокочастотный» относится к группе свойства.

3. Влияние параметров на построение дерева. Бывает, что небольшие изменения в наборе данных могут привести к построению другого дерева. Это связано с тем, что изменения в узле на верхнем уровне ведут к изменениям во всем дереве ниже. Так, при изменении слов в группах при классификации меняется последовательность слов на ветвях дерева, а так же количественные значения повторяемости слов.

4. Анализ авторов статей, расположенных в листьях, позволяет выявить некоторые закономерности:

- статьи некоторых авторов (коллективов авторов) попали в один лист (группу), что может свидетельствовать об узком характере исследований, то есть ограниченному выбору объектов, методов и подходов;

- есть авторы (коллективы авторов), которые встречаются в нескольких группах, что может свидетельствовать как о междисциплинарном характере исследований, так и о широте применяемых методов и подходов.

Заключение. Анализ данных больших объемов требует привлечения технологий и средств реализации высоко производительных вычислений. Основными факторами проблемы являются в первую очередь сложность и во вторую физический объем информационной коллекции.

В настоящее время в разных сферах деятельности все более нуждаются в средствах, позволяющих быстро и безошибочно перерабатывать большое количество информации. Применение таких средств позволяет существенно снизить затраты и повысить эффективность работы.

В ходе написания выпускной квалификационной работы были получены следующие результаты.

В ходе написания магистерской работы были получены следующие результаты.

В теоретической части работы:

a) рассмотрено понятие «Больших Данных» и методов их анализа;

b) выбран оптимальный метод Data Mining;

c) изучен алгоритм дерева принятия решений.

В практической части работы:

a) проведена обработка научных публикаций;

b) построено дерево классификации и проанализированы результаты:

1. Дерево классификации позволило провести разбиение статей на группы с использованием количественного критерия, что в дальнейшем должно облегчить релевантный поиск информации и выбор статей, необходимых в определенной области.

2. Существуют погрешности вызванные ограничениями, наложенными на высоту дерева и минимальный прирост энтропии. Так при ограничении высоты дерева есть вероятность, что не все статьи будут классифицированы или классифицированы, укрупнено, то есть в одной статье могут быть слова из разных выделенных групп.

3. Влияние параметров на построение дерева. Бывает, что небольшие изменения в наборе данных могут привести к построению другого дерева. Это связано с тем, что изменения в узле на верхнем уровне ведут к изменениям во всем дереве ниже. Так, при изменении слов в группах при классификации меняется последовательность слов на ветвях дерева, а так же количественные значения повторяемости слов.

4. Дерево классификаций может дать дополнительную информацию об авторах статей, а точнее об их области исследовательских интересов.

Список использованных источников

1 Введение в Big Data [Электронный ресурс] // Молодой ученый [Электронный ресурс] : [сайт]. - URL: https://moluch.ru/archive/145/40562/ (дата обращения: 15.01.2019). - Загл. с экрана. - Яз. рус.

2 Крылов, В. В. Большие Данные и их приложения в электроэнергетике / В. В. Крылов. - М. : Нобель Пресс, 2014. - 166 с.

3 Data Mining – технология добычи данных [Электронный ресурс] // Теория и практика обработки информации [Электронный ресурс] : [сайт]. - URL: http://bourabai.ru/tpoi/datamining.htm (дата обращения: 10.03.2019). - Загл. с экрана. - Яз. рус.

4 Data Mining –добыча данных [Электронный ресурс] // BaseGroup Labs – технологии анализа данных [Электронный ресурс] : [сайт]. - URL: https://basegroup.ru/community/articles/data-mining (дата обращения: 25.03.2019). - Загл. с экрана. - Яз. Рус.

5 Использование деревьев решений в задачах прогнозной аналитики [Электронный ресурс] // Прогноз [Электронный ресурс] : [сайт]. - URL: http://www.prognoz.ru/blog/platform/decision-tree-in-predictive-analytics/ (дата обращения: 26.04.2019). - Загл. с экрана. - Яз. рус.