ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

На сегодняшний день уже очевидно, что в будущем наиболее распространенный вид работы будет являться сочетанием человеческих и машинных ресурсов. Уже сейчас такой тандем активно практикуется в различных сферах жизнедеятельности человека. Этот симбиоз позволит значительно увеличить качество и скорость работы за счет использования вычислительной мощности техники и уникальных качеств человеческого организма.

При использовании такого подхода машина (компьютер) должна выполнять роль механизма анализа накопленных (собранных) данных и инструментом прогнозирования, человек же — корректирующей системой. Такой метод работы с информацией сможет максимально рационально использовать все лучшие качества каждого инструмента.

В дипломной работе будут освещаться вопросы разработки таких систем, их особенности и нюансы, связанные с сферой образования, в частности с распределением нагрузки кафедры. В ходе исследования и сравнительной характеристики будут определены методы и пути решения данной проблемы

Путем анализа научных трудов можно выявить новую информацию о квалификации преподавателя в различных отраслях знаний и упростить задачу работников кафедры предоставляя им более подробную информацию.

Таким образом можно получить информацию, которая позволит намного точнее рассказать об опыте субъекта в данной отрасли знаний, тем самым позволив работнику принять более взвешенное решение.

Целью исследования является оптимизация распределения нагрузки кафедры на основе анализа публикаций сотрудников кафедр, их

персональной информации и кафедральной документации. Путем анализа трудов сотрудников сферы образования можно выявить новую информацию об их знания в различных сферах, тем самым увеличить качество распределения нагрузки.

Таким образом объект исследования будет являться процесс сбора, хранения и анализа данных принимающих участие в формировании нагрузки кафедры, а предметом – подсистема сбора и анализа публикаций сотрудников кафедр системы управления кафедрой.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Изучить и проанализировать рассматриваемую отрасль знаний.
2. На основании данных полученных при решении первого пункта вывести требования к системе, ПО, и внедряемым алгоритмам
3. Проанализировать алгоритмы Data Mining, которые уже активно используются для решения подобных задач. В случае, если ни один алгоритм не сможет решить проблему в полной мере предлагается рассмотреть применение смешанных алгоритмов либо модификацию существующих.
4. Реализовать и протестировать выбранный(разработанный) алгоритм.
5. На основании полученных данных после тестирования сделать выводы позволяющие оптимизировать алгоритм и внедрить его в систему распределения нагрузки кафедры модульно.

Описание предметной области

Сфера образования является одной из наиболее важных сфер жизнедеятельности человека. Это обусловлено тем, что именно тут готовят специалистов, которые будут в будущем будут заниматься развитием человечества. Именно поэтому крайне важно поддерживать качество образования на высоком уровне.

Однако, с увеличением количества новых предметов в вузах можно увидеть, что сотрудникам данной сферы, с каждым годом, приходится оперировать все большими объемами данных, что неизбежно приводит к увеличению ошибок.

Задача распределения нагрузки является одной из наиболее трудоемких и важных в учебном процессе и, соответственно, она не может быть абсолютно защищена от человеческого фактора.

Также, стоит отметить, что за долгие годы, было накоплено множество печатной и электронной неструктурированной информации с которой крайне сложно работать.

Таким образом, если структурировать накопленные знания и упорядочить всю информацию, можно будет повысить осведомленность сотрудников об квалификации преподавателей и тем самым увеличить показатели качества обучения. Этого можно достичь если предоставлять работникам информацию в привычной для человека форме и впоследствии облегчая задачу распределения нагрузки.

Ранее уже была разработана система позволяющая выполнять операции в online режиме путем взаимодействия с системой менеджмента [1]. Однако, система только повторяла ручной процесс распределения нагрузки.

На сегодняшний день становятся все более актуальным использование различных инструментов интеллектуального анализа для оптимизации человеческого труда. Такой подход значительно увеличивает качество работы. Анализируя предметную область, выявляя зависимости и обучаю машину работать в паре с человек такой подход получил широкое распространение в сферах медицины, рекламы и бизнеса. Сбор и анализ данным помагает выявить общие тенденции и выявлять зависимости еще на ранних этам процессов, что значительно ускоряет работу и ускоряет принятие решений. Однако, стоит отметить, что данный подход, в связи с сложностью исполнения, все еще является весьма трудоемким и дорогим в разработке.

В скором будущем человеку придется обрабатывать в миллионы раз больше информации чем мы уже накопили сейчас. Объем информации возрастает ежегодно на 30% что около $2{,}5 \cdot 10^8$ байт на человека [2]. На основании статистики, собранной за долгое время, уже давно был выведен и освещен закон роста информации в работах Дерека Прайса и описан как экспоненциальная функция (1).

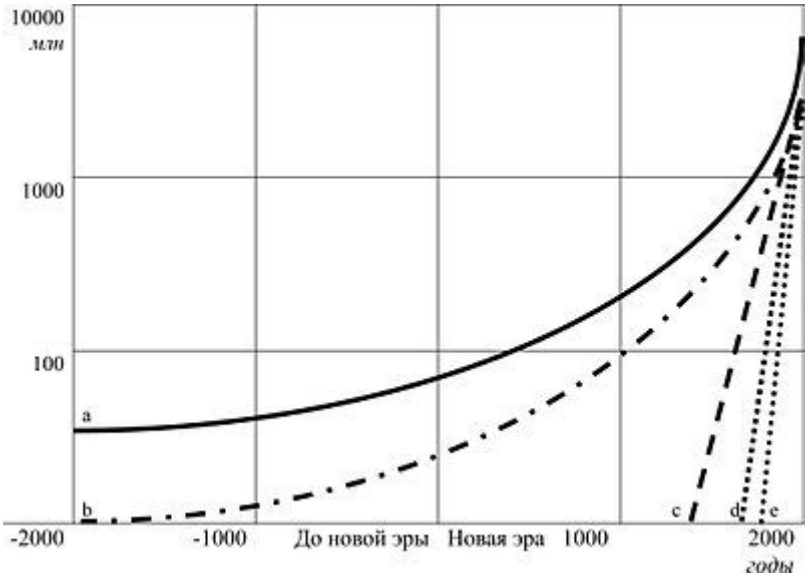

Рисунок 1. График увеличения накопленной информации относительно времени.

Очевидно, что при таком росте информации необходимы инструменты ее обработки, иначе ее рациональное использование становится невозможным.

Существует множество методов анализа данных, а вся отрасль, занимающаяся анализом данных, называется Data Science [3]. Этот раздел знаний активно занимается применением различных методов и подходов к обработке и анализу данных. Сочетая использование вычислительных мощностей компьютеров так и законы линейной алгебры и статистики становиться возможным уже сейчас производить анализ терабайтов данных.

Одним из направлений Data Science является Data Mining. Объединяя в себе совокупность методов для решения различных аналитических задач, этот подход уже активно используется для интеллектуального анализа различными крупными компаниями. Например Google для индексации страниц в своём поисковике используют метод

Page Rank, а Amazon используют наборы методов для рекламы своих продуктов [4].

В целом, методы Data Mining делятся на два основных типа по их применению: прогнозирующие и описательные [5].

Каждая группа состоит из методов которые обладают рядом определенных свойств. Например, к прогнозирующим методам относятся методы Дерева Решения, Нейронные Сети, Методы Регрессии, а к описательным методы кластеризации и классификации.

Для поставленной задачи нам будут необходимы методы второй группы. Методы классификации применяются для отраслей в которых изначально заданы точные модели данных, относительно которых будет работать алгоритм (Рис. 1). Благодаря этому, алгоритм работает весьма быстро но имеет недостаток связанный с жесткостью задания условий.

Методы кластеризации применяются в отраслях где заведомо неизвестны особенности каждой новой группы (Рис. 1) Алгоритм распределяет данные по группам схожести основываясь на каком-то коэффициенте. Данные методы требуют больше временных затрат, но являются более гибкими и расширяемыми.

Также, для создания внутренней иерархии кластеров используется метод иерархической кластеризации, благодаря которому можно получить данные связанные между собой родительно-наследственными связями, что позволит сделать алгоритм точнее.

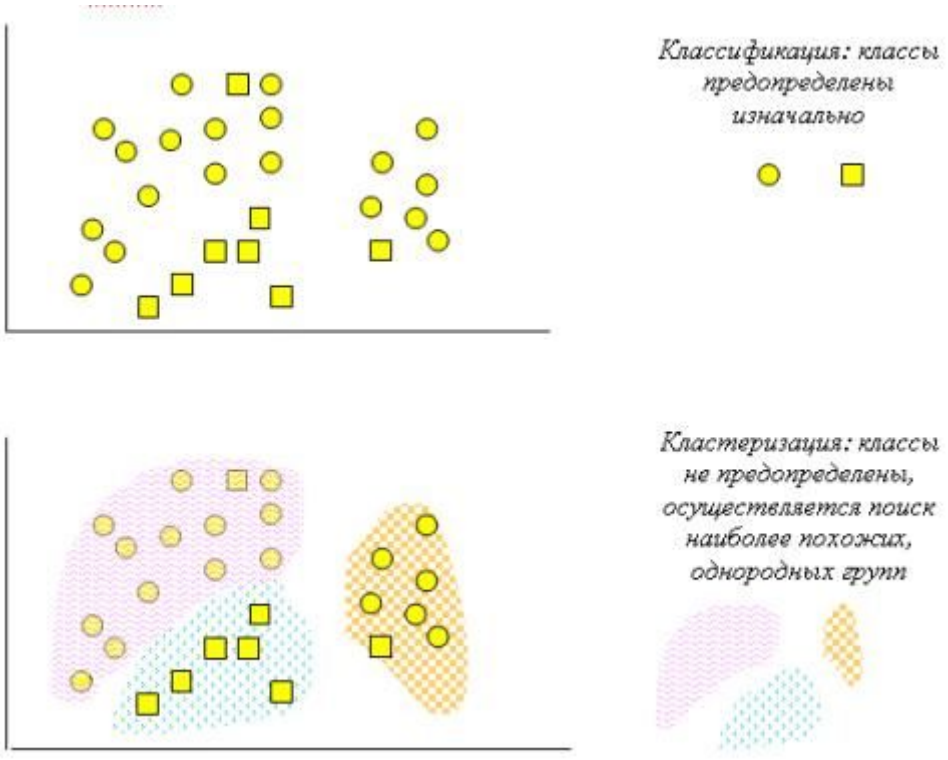

Рис. 1. Иллюстрация различий выборки методов кластеризации и методов классификации.

В каждом из методов Data Mining можно использовать алгоритмы самообучения, которые позволят определять новые данные с высокой точность. С внедрением такого алгоритмы в метод иерархической кластеризации можно, в ходе его работы, получать новые кластеры и делать разбиения точнее с каждой новой итерацией.

Для реализации подобных алгоритмов обычно используются микросервисы благодаря которым, в зависимости от необходимости, можно, в любую единицу времени, увеличивать или уменьшать машинную мощность, что позволяет сделать подсистему более гибкой и удобной.

Произведя анализ области знаний было выявлено, что сотрудники занимающиеся нагрузкой кафедры не всегда и не точно знать об квалификации преподавателей в различных сферах, что затрудняет

распределение нагрузки и тратит время. Однако, каждый научный деятель постоянно пишет множество статей и трудов, анализ которых мог бы значительно облегчить задачу распределяющих нагрузку.

В рамках поставленной задачи можно использовать методы кластеризации для распределения научной литературы преподавателей по различным кластерам, тем самым увеличивая осведомленность работников, занимающихся нагрузкой, о достижениях их коллег.

Отображая полученную информацию в удобной и привычной форме, можно еще на этапе распределения нагрузки увеличивать качество обучения, так как на каждую дисциплину будет автоматически рекомендоваться преподаватель с наиболее высокими оценками в выбранной дисциплине.

Обзор методов кластеризации

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами[6]. Внутри каждой группы должны оказаться объекты с близкими параметрами схожести, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Применение кластерного анализа в общем виде сводится к следующим этапам:

1. Отбор выборки объектов для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.
3. Вычисление значений меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).
5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

В рамках поставленной задачи

Общий подход к кластеризации объектов

Для определения объектов в методах кластеризации используются специальные метрики, или функции расстояния. Эти расстояния могут определятся в одномерном или многомерном пространстве и являются описательной характеристикой схожести объектов.
В основном, используется 3 наиболее популярных видов измерений:

$$P = VALUE \left| A_i \neq B_i \right|$$

Задача выявления ключевых слов в различных задачах анализа текста является одной из самых важных. Это обусловлено тем, что именно ключевые слова позволяют определять смысл и природу текста. Ключевое слово — слово в тексте, способное в совокупности с другими ключевыми словами дать высокоуровневое описание содержания текстового документа, позволяющее выявить его тематику [1].

В анализе научной литературы ключевыми словами являются языковые слова и словосочетания которые относятся к определенной области знаний. Именно эти слова позволяют определять какая предметную область текста.

Необходимо установить, что будет являтся keyword, а что нет. Для этого достаточно рассмотреть части речи в языке который выбран для анализа. Можно отметить, что только имена существительные могут относиться к ключевым словам, так как именно эта часть речи полностью называет объект. Таким образом, необходимо разделять существительные и другие части речи.

Также, стоит отметить, что основной частью слова, отвечающей за его смысл, является его корень и в некоторых случаях может иметь смысл проверки глаголов и прилагательных на наличие корней в базе уже существующих ключевых слов. Таким образом можно повысить точность определения новых ключевых слов, синонимов и производить более детальный анализ текста

Другим оценочным фактором будет плотность частоиспользуемых ключевых слов (академическая тошнота). Академическая тошнота – наиболее употребляемые слова и словосочетания в тексте, которое характерны для любых документов[2]. Для улучшения показателя

определения тематики текста необходимо исключать подобные слова и словосочетания, что позволит более узко воспринимать текст.

Стоит отметить, что во многих языках встречаются коллизии в словах и их необходимо определять. Одно и то же ключевое слово в разном контексте может означать разный смысл и не относиться к другой предметной области [3]. Для решения этой проблемы можно использовать полиморфную ассоциативную связь между предметной областью и ключевыми словами, что позволит на лету определять значение слова в конкретном контексте.

Таким образом исключив из выборки все слова которые не являются именами существительными (глаголы, прилагательные, местоимения и т. д.), отфильтровав часто используемые слова в научной документации мы получим набор слов описывающих суть текста.

Высчитав количество ключевых слов в тексте и определив соответсвие между словами и темами в которых они чаще всего фигурируют можно определеить тему текста и отнести его к какой-то категории — кластеру. При анализе научных трудов преподавателей можно выявить самые популярные темы и основные направления в исследованиях и работах. Такая информация позволит более качественно распределять кафедральную нагрузку.

ЛИТЕРАТУРА

1. Диплом Царюк А.О. 2016 стр. какая-то
2. Lyman P., Varian H.R. How much information. Release of the University of California. Oct.27, 2003.
3. Data Science from Scratch Joel Grus Publisher: O'Reilly Media Release Date: April 2015 Pages: 330 / 9 старница
4. Matthew Richardson, Amit Prakash, Eric Brill. Beyond PageRank: Machine Learning for Static Ranking. — 2006.
5. Электронный ресурс. http://intellect-tver.ru/?p=165
6. Мандель И. Д. Кластерный анализ. — М.: Финансы и Статистика, 1988.