В настоящее время массивы информации, доступные человеку, многократно выросли благодаря развитию сети Интернет. Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией. В огромных информационных объемах имеет смысл говорить только об автоматической рубрикации.

В ходе выполнения работы был создан программный комплекс – автоматический классификатор данных с применением алгоритмов анализа естественного языка, применимый для извлечения структурированной информации из текстов. Данный программный комплекс позволяет автоматически обрабатывать поступающие материалы по выбранным тематикам.

В данной работе поставлена задача разработки программного комплекса, позволяющего автоматизировать сбор и структурирование информации на естественном языке с тематических интернет-ресурсов – классификатора данных.

Для достижения поставленной цели необходимо решить ***следующие задачи***:

1. Проанализировать современное состояние исследований в области агрегации данных и анализа естественных языков.
2. Провести анализ существующих научных и практических решений в выбранной области, изучить методы, принципы и технологии извлечения именованных сущностей, возможности их применения для агрегации данных.
3. Спроектировать компонентную реализацию программной системы для агрегации тематических данных.
4. Разработать программный комплекс.
5. Сделать вывод об эффективности созданной системы классификации данных.

Классификация или рубрикация информации: отнесение порции информации к одной или нескольким категориям из ограниченного множества, является традиционной задачей организации знаний и обмена информацией.

При применении методов машинного обучения для построения классификатора используется набор документов, предварительно отобранная человеком. Алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества текстов [1].

Машинное обучение ‒ это научное исследование алгоритмов и статистических моделей, которые компьютерные системы используют для эффективного выполнения конкретной задачи без использования явных инструкций, опираясь на шаблоны и выводы.

Входной набор материалов, который использован при отладке программного обеспечения, состоит в общей сложности из более чем 2000 документов. Документы распределены по пяти тематикам: спорт, бизнес, политика, развлечения и техника.

После импорта набора материалов выполняется предварительная обработка текста каждого документа. Текст может содержать цифры, специальные символы и ненужные пробелы. В зависимости от проблемы, решается задача удаления или не удаления этих специальных символы и цифр из текста.

Процедура обработки текста начинается с удаления всех символов, не являющихся словами, таких как специальные символы, цифры и прочее. Затем удаляются все одиночные символы (союзы, сокращения слов). На следующем шаге данные преобразуются в нижний регистр, чтобы слова, которые на самом деле одинаковы, но имеют разный регистр, могли быть обработаны одинаково. Последним шагом предварительной обработки является лемматизация. При лемматизации слово сокращается до словарной корневой формы.

После чего рассчитываются коэффициенты и частоты встречаемых слов.

Наивный Байесовский классификатор (НБК) – вероятностный классификатор, который делает предсказания, основываясь на списке определенных классов. Конечной задачей рубрикатора является определение к какому классу принадлежит документ, поэтому необходимо получить не саму вероятность, а наиболее вероятный класс. Для этого НБК производит оценку апостериорного максимума. Другими словами, рассчитывается вероятность для всех классов и выбирается тот, который обладает максимальной вероятностью.

Под моделью НБК понимается совокупность информации, которая вычисляется на основе данных, полученных во время обучения классификатора, а именно:

1. общее количество документов, участвовавших в обучении;
2. количество документов по каждому классу;
3. количество уникальных слов в обучающем наборе;
4. общее количество слов по каждому классу.

В ходе проверки результатов программного обеспечения было выявлено, что реализованный классификатор достигает точности 97,6%, что является отличным показателем.

Для проверки работоспособности программного комплекса был использован API (совокупность инструментов и функций в виде интерфейса для создания новых приложений, благодаря которому одна программа будет взаимодействовать с другой) веб-сайта NewsCatcher.com, который позволяет запрашивать популярные новости различных тематик.

Во время работы программного обеспечения с сайта NewsCatcher запрашиваются последние пять новостей, для каждой из которых последовательно определяется тема. В ходе классификации новой информации создаётся файл, в который записывается полученная новость, а сам файл сохраняется в папку с названием, соответствующим теме новости.

В програмнном комплексе реализована возможность обработки любых пользовательских текстовых файлов из входной папки (Input). Сканирование документов в папке производится непрерывно, пока программа запущена.

Таким образом, точность работы комплекса достаточно высока и позволяет решать поставленные задачи классификации.

Проблема неверно извлеченных и интерпретированных тем решается с помощью дополнения базы ранее верно определенных тем автоматически, для чего разработанный программный комплекс предоставляет необходимый инструментарий.

Разработанные модули и компоненты системы при взаимодействии осуществляют надежную и бесперебойную работу приложения. Обеспечивают удобный и комфортный интерфейс работы для пользователя.

В ходе исследования представлены подходы по оптимизации и увеличению скорости работы программы, такие как параллельное программирование и уменьшение числа термов. Проведен эксперимент по обучению на основе статей различных тематик и проанализированы его результаты.

В результате проделанной работы было разработано программное обеспечение, базирующееся на байесовском алгоритме. ПО позволяет определять тематику текста на основе данных, полученных во время обучения классификатора.

По результатам выполнения НИРС программное обеспечение позволяет производить классификацию текстов на естественном языке по пяти темам. Заложены возможности по расширению библиотек, используемых тем с целью обеспечения охвата более широкого спектра проблем.

Реализованный программный комплекс планируется использовать в подразделении для решения задачи автоматического извлечения тем документов и структурирования данных из файлов на естественном языке. Программное обеспечение позволит обеспечить оптимальную организацию процесса сбора информации и уменьшит временные затраты на поиск информации представляющий интерес.