ИСПОЛЬЗОВАНИЕ ПРИНЦИПОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ СОЗДАНИЯ СИСТЕМЫ АНАЛИЗА ТЕКСТА НА ПРЕДМЕТ ЗАИМСТВОВАНИЙ

В.В. Останин, студент группы 17В20,
научный руководитель: Захарова А.А.
Юргинский технологический институт (филиал) Национального исследовательского Томского политехнического университета
652055, Кемеровская обл., г. Юрга, ул. Ленинградская, 26

В свете последних событий Россия как никогда раньше сильно нуждается в новых научных исследованиях и разработках. Отставание нашей страны во многих отраслях экономики очевидно, перед нами стоит ряд острых проблем, которые требуют незамедлительного решения. Особенно актуально это стало перед лицом нависшей опасности, санкций, угроз во внешней и внутренней политике, и возникшей необходимости в импортозамещении. Для решения данных проблем требуются прорывные исследования в различных областях науки. С недавних пор государство различными способами поощряет научные исследования, особенно по приоритетным направлениям, были учреждены различные гранты, активизировались правительственные фонды, увеличено финансирование научных учреждений. К сожалению, данный факт кроме положительного воздействия имеет и негативный эффект.

Некоторые учёные говорят о том, что количество информации с каждым годом увеличивается примерно в 2 раза. Информация порождает новую информацию, а структура исходной информации усложняется. Но не всякая информация является полезной, информация имеет свойство копироваться, распространяться, изменяться, фальсифицироваться.

Такие свойства информации приносят значительный вред обществу и экономике нашей страны. Увеличение финансирования научных исследований вызвало всплеск плагиата научных публикаций и исследований. В результате дополнительные денежные средства могут пойти не на поддержку инновационных проектов, а на проекты, которые лишь создают видимость научной значимости. К сожалению, использованием чужих трудов занимаются не только школьники и студенты, но и аспиранты, и даже доктора наук. Конечно, заимствование является неотъемлемой частью научных исследований, т.к. используя результаты чужих исследований в качестве фундамента возможно достичь каких-то совершенно новых и значительных результатов. Но очень часто дело обстоит иначе. Научная работа или её часть копируется полностью или немного видоизменяется и выдаётся за собственную. К сожалению, это встречается повсеместно. В последнее время в связи с расширением возможностей сети Интернет «написание» научной статьи очень сводится всего лишь к нажатию комбинаций клавиш «Ctrl + C» и «Ctrl + V», которые соответствуют операциям «скопировать» и «вставить», и иногда к поиску синонимов в интернете с целью обмануть программу «антиплагиат».

Такая псевдонаучная деятельность наносит серьёзный ущерб экономике страны, т.к. денежные средства в результате расходуются не по назначению, а так же и имиджу страны и научных учреждений. На самом деле, проблема плагиат - это проблема не только научных исследований. Простейший пример плагиата – списывание на экзаменах у «соседа». В результате оба обучающихся скорее всего получат одну и ту же оценку, при том что реальную работу проделал лишь один из них. По окончании учебного заведения с одинаковыми оценками об успеваемости оба выпускника с большой долей вероятности смогут устроиться на одинаковые работу, при том что один из них не будет обладать необходимыми компетенциями. Таким образом очевидно, что плагиат подрывает не только научную деятельность страны, но и её кадровый потенциал. В качестве доказательства распространённости данного исследования можно привести результаты опроса, проводившегося в США: 8о% студентов признаются, что хотя бы раз в жизни списывали, 36% студентов списывают регулярно, 90% учащихся убеждены, что их плагиат никогда не будет обнаружен. 58% американских школьников в 1969 году уже давали свои работы для списывания своим соученикам, 97,5% - в 1989 году. 54% студентов признаются в том, что они списывали из Интернета. 74% студентов признают, что они достаточно регулярно списывали [1].

На самом деле, плагиат сам по себе вне зависимости от того, в какой сфере деятельности он проявляется, довольно неприятен автору информации. Очень многие ресурсы в интернете дублируют друг друга, используя чужой контент и выдавая его за свой. Это сказывается на рейтинге и посещаемости ресурса источника информации и в конечном итоге наносит экономический урон ему.

Таким образом, плагиат является серьёзной угрозой не только для отдельного автора, но и для всего государства в целом. К сожалению, эта проблема достаточно плохо контролируется. Сущест-

вуют различные системы предназначенные для поиска плагиата, но суть их работы в конечном счёте сводится к одному и тому же – поиск информации в интернете и проверка на наличие совпадений частей текста. Данные системы достаточно просто обходятся благодаря богатству русского языка. Достаточно заменить слово на его синоним или же просто заменить букву в слове, и система перестаёт замечать плагиат. Можно так же пересказать исходную информацию совершенно другими словами, полностью изменив и количество слов, и их порядок, но при этом сохранить основную идею. Благодаря таким действиям практически невозможно обнаружить плагиат обычными средствами.

В последнее время произошло довольно серьёзное увеличение вычислительных мощностей. Увеличивается скорость передачи информации по сети, увеличивается количество информации, которую можно обработать за единицу времени, однако, увеличивается и объём хранимой информации. Выделить плагиат из уже существующей информации представляется практически невозможным, а ведь количество информации непрерывно увеличивается.

Уже сейчас современный человек за месяц получает и обрабатывает столько информации, сколько человек 17-го века — за всю жизнь. В одной только сети Facebook ежемесячно выкладывается 30 млрд новых источников информации.

В 2011 году общий мировой объем сгенерированных человечеством данных составил более 1,8 зеттабайт (1,8 трлн Гб), что в 57 раз больше, чем песчинок на всех пляжах Земли.

Проблема обработки «больших данных» (BigData) состоит не столько в их объеме, сколько в отсутствии адекватного инструмента для работы с ними. Количество внешних и внутренних источников информации непрерывно растет, а сами данные становятся и сложнее, и разнообразнее — структурированные, неструктурированные и даже квазиструктурированные. В-третьих, они поразному индексируются. При этом далеко не все данные ценны — по оценкам IDC, к 2020 году доля полезной информации составит лишь 35% от всей сгенерированной [2].

Одним из возможных решений проблемы плагиата является искусственный интеллект, а точнее один из подразделов теории искусственного интеллекта – машинное обучение. Суть машинного обучения заключается в построении алгоритмов, способных к обучению. В общей формулировке задача машинного обучения выглядит так. Дано конечное множество прецедентов (объектов, ситуаций), по каждому из которых собраны (измерены) некоторые данные. Данные о прецеденте называют также его описанием. Совокупность всех имеющихся описаний прецедентов называется обучающей выборкой. Требуется по этим частным данным выявить общие зависимости, закономерности, взаимосвязи, присущие не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые ещё не наблюдались (рис. 1). Так например, уже сейчас такие алгоритмы активно используются для фильтрации электронной почты от спама или в различных поисковых системах, в системах распознавания рукописного ввода, образов, жестов и речи, в системах прогнозирования и многом другом [3].

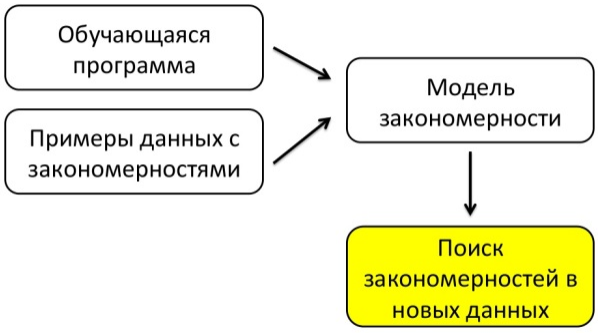

Рис. 1. Схема работы алгоритма

Преимущества использования машинного обучения очевидны – с помощью таких алгоритмов станет возможен поиск плагиата при использовании синонимов, перестановки слов и даже поиск плагиата по смыслу. Каждое новое использование такого алгоритма повышает его точность при оценке информации за счёт выявления новых закономерностей на основе имеющихся или благодаря действиям пользователя.

Более широкой задачей этого алгоритма может стать некое цензурирование информационного наполнения рунета, т.к. проблемы информации не ограничиваются одним лишь плагиатом. Самообучающийся алгоритм подобного рода можно было бы использовать для фильтрации избыточной и дублирующей, оскорбительной, недостоверной информации, тем более, что такой подход достаточно успешно работает в почтовых сервисах для фильтрации спама.

Конечно, машинное обучение, а именно обучение ранжированию уже применяется в большинстве современных поисковых систем, к которым относится и система «антиплагиат», но в этих системах она применяется для поиска совпадений в сети, а не для анализа самого текста. Новизна идеи этой работы заключается в том, чтобы использовать обучающийся алгоритм не только для поиска совпадений информации в сети, но и для поиска и анализа закономерностей в самой проверяемой информации с повторным поиском совпадений в сети. Безусловно, такой подход значительно увеличит требования к ресурсам системы и продолжительность процесса проверки, но даже при анализе большого объёма информации использование распределённых или облачных вычислений позволит эффективно использовать данный алгоритм проверки.

Литература.

1. Plagium vulgaris: как предотвратить плагиат в науке [электронный ресурс] - режим доступа : http://rian.ru/online/326470019.html
2. Объем информации в мире будет удваиваться каждые пару лет / Полит.ру [электронный ресурс] – режим доступа: http://polit.ru/news/2013/05/14/jump_bigdata/
3. Машинное обучение / MachineLearning.ru [электронный ресурс] – режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение

ОЦЕНКА И ВЫБОР КАНАЛА СБЫТА МЕТАЛЛУРГИЧЕСКОГО ПРОИЗВОДСТВА

М.С. Рыльцев, студент группы 17В10,
научный руководитель: Важдаев А.В.
Юргинский технологический институт (филиал) Национального исследовательского Томского политехнического университета
652055, Кемеровская обл., г. Юрга, ул. Ленинградская, 26

Реализация продукции является завершающим этапом в цепи поставок. Данный этап имеет решающее значение в достижении главной цели товародвижения. Сбыт – это процесс реализации произведенной продукции с целью превращения товаров в деньги и удовлетворения запросов потребителей.

Отправка продукции конечному потребителю может производиться разными способами, путем использования различных каналов сбыта. Канал сбыта — цепь фирм, участвующих в покупке и продаже товаров по мере их продвижения от изготовителя к потребителю. Различают несколько уровней канало сбыта:

1. Канал нулевого уровня — прямой метод продаж от производителя к потребителю (используется, когда сбыт продукции осуществляется крупными партиями).

2. Одноуровневый канал. В его состав сходят производитель, представитель розничной торговли, потребитель.

3. Двухуровневый уровень. Основные звенья: производитель, оптовый посредник, мелкий посредник, потребитель (используется, когда предприятие не вкладывает средства в формирование сбытовой системы и сотрудничает с оптовыми и розничными посредниками, составляющими независимую сбытовую сеть).

4. Трехуровневый (состоит из оптового посредника, мелкооптовой и розничной торговли).

5. Многоуровневый (имеет множество посредников в сбытовой сети).

От конкретного состава и количества участников, составляющих канал сбыта, зависит эффективность реализации продукции, и, в конечном счете, прибыль предприятия. Как быстро готовая продукция будет доставлена потребителю, какова ее конечная стоимость, какие затраты и риски несет производитель и потребитель при использовании определенного канала и другие важные моменты необходимо принимать во внимание при выборе канала распределения. Под эффективностью при оценке канала сбыта будем понимать степень достижения результата, т.е. доставка качественной продукции потребителю в срок за оговоренную цену с минимальными затратами.

В практике производственных предприятий каналы сбыта продукции часто складываются стихийно. Поэтому для эффективной работы предприятия необходимо время от времени проводить комплексную оценку их эффективности с целью выбора каналов и участников товародвижения, сотрудничество с которыми является наиболее выгодным с точки зрения производителя.