ДОКЛАД

Уважаемый председатель аттестационной комиссии, уважаемые члены комиссии! Я Толкачёв Никита, студент группы 872301 инженерно-экономического факультета. Тема моей дипломной работы «Система обработки текстов нарративных источников на основе интеграции лингвистических и статистических методов анализа».

Цель дипломной работы состоит в объединение трех различных подходов в анализе нарративных источников (контент-анализ, дискурс-анализ и психоанализ) и их реализации на основе облачных вычислений.

Исходя из этого, выстраивается ряд задач, а именно:

* проанализировать классификацию текстовой информации;
* дать характеристику основных процессов обработки текстовых данных;
* ознакомиться с основными подходами в анализе нарративных источников;
* определить исходные и результирующие данные для различных подходов анализа;
* рассмотреть случаи дублирования и взаимного использования промежуточных данных для каждого из подходов;
* выявить проблемы в исследуемых методах анализа текстов нарративных источников и разработать предложения по совершенствованию их работы.

Объектом исследования являются текстовые данные нарративных источников.

Предметом исследования являются модели и методы обработки текстовых данных нарративных источников в рамках исторической информатики.

Тема дипломной работы выбрана не случайно. Дело в том, что с возникновением и развитием информационного общества главным продуктом производства становятся информация и знания. И это непосредственно влияет на возрастание роли самой информации, увеличение числа людей, занятых информационными технологиями, а вопрос о новых подходах в обработке данных стоит как никогда остро.

По результатам исследований, мировой объем информации увеличивается более чем в два раза каждые два года. Цифровая Вселенная и Большие объемы данных стали одними из движущих сил фундаментальных изменений в социальной жизни, технологиях, науке и экономике. Между тем, растут и инвестиции корпоративного сектора в эту сферу. Большие объемы данных открывают новые возможности и изменяют роль IT, но, наряду с этим возникает проблема в анализе и восприятии всей этой информации человеком с целью получить требуемые, нужные данные. Именно на этом этапе и возрастает значение автоматизации обработки данных, преимущественно текстовой информации.

Сама идея не является новой. Начиналось все с самых простых элементов, прототипов текстовых редакторов, когда к ЭВМ подключили терминалы для ввода-вывода информации. На текущий момент в области компьютерной обработки текстов сформировалось не одно направление, однако наиболее значимым является так называемый «text analysis» (грубым синонимом является «text mining», который хоть и обладает схожим смыслом, но в современном контексте не идентичен) или анализ текста.

Под анализом текста в настоящее время понимают процесс получения высококачественной информации из текста на естественном языке. Под естественным языком тут понимается язык, который используется для общения людей, а не выведенный искусственно (например, любой язык программирования или эсперанто). Полученная высококачественная информация может быть интерпретирована в различном стиле и зависит от конкретных нужд и реализаций самого анализа. Это может быть и простой список ключевых слов, и автореферат, содержащий основные части источника, и даже текст, являющийся результатом качественной оценки и интеллектуальных выводов самой системы или машины (что является на данный момент наиболее перспективным, но трудно реализуемым направлением).

Таким образом, разработанная система производит анализ текста, базируясь на трех основных подходах: контент-анализ, дискурс-анализ и психолингвистический анализ. Выбор именно этих подходов так же имеет свою обоснованность. Во-первых, они нацелены на получение не статистической информации, но на качественные, ранее неизвестные данные, связи и зависимости между ними.

Востребованность такого программного продукта была выявлена в ходе участия в нескольких научных конференциях, где были представленные ранние прототипы. Одной из них является XIII международная конференция Ассоциации «История и компьютер».

Результатом разработки стал облачный сервис, представляющий собой систему по анализу текстовых данных. Данный подход позволил применить концептуально новые решения в создании подобных систем. В рамках стандартного разделения на клиентскую и серверную часть, клиент стал практически полностью независим от сервера, что позволяет сократить количество запросов и снизить нагрузку на последний. Таким образом, пользователю достаточно один раз произвести анализ в рамках выбранного проекта и работать с ним в режиме оффлайн без внесения качественных изменений в полученных данных. Однако из этого возникли и некоторые недостатки, проявляющиеся в обратной совместимости и наличии поддержки лишь самых последних версий браузеров на момент разработки.

Процесс разработки данного программного обеспечения представлял (и представляет, ввиду продолжения разработки) собой спиральную модель. В качестве подхода к разработке программного обеспечения в рамках спиральной модели жизненного цикла, была использована получившая в последнее время широкое распространение методология быстрой разработки приложений RAD (Rapid Application Development). Это было обусловлено, в первую очередь тем, что разработка велась всего одним программистом и имела короткий производственный график. Благодаря этому, была возможность представить рабочие версии ПО на различных конференциях, в ходе которых система была дана конструктивная критика и оценка системы, а так же высказаны различные важные предложения.

Основной трудностью при интеграции и совместном использовании различных по своей природе подходов является повторное использование данных. Проблема заключается в том, как будут использованные промежуточные и конечные данные при переходе с одного этапа на другой. В противном случае, ни о какой интеграции речи идти не может. В основе решения данной проблемы в рамках нашей работы лежит модель сочетаемости нетрадиционных методов исторического познания текста нарративного источника. Она была разработана, сформулирована и описана научными сотрудниками БГУ Сидорцовым В.Н. и Приборовичем А.А. Более детально с ней можно ознакомиться в их работе «Научный дискурс историка: социальная обусловленность и методология исследования». [использование иллюстрации]

Как было сказано ранее, система представляет собой облачный сервис. Под облачным сервисом можно понимать систему, реализующую принцип SaaS (Software as a Service), в основе которой лежат облачные вычисления. Сам продукт можно условно разделить на сервер и тонкого клиента. Сервер отвечает лишь за бизнес-логику и производит предварительную обработку (структуризацию, парсинг и прочее) текста и извлечение из него необходимой информации, а так же за хранение этой самой информации. Хранение этой самой информации и словарей происходит в базе данных, к которой сервер имеет непосредственный доступ. Тонкий клиент выполняется в браузере и средствами HTML5 и JavaScript позволяет достичь большой степени автономности, убрав необходимость постоянного обмена информацией, используя лишь синхронизацию. Формат JSON обеспечивает обратную совместимость и предоставляет возможность использовать данные полученные сервером на клиенте, а так же передавать их по сети. Кроме того, HTML5 LocalStorage позволяет хранить данные на жестком диске, что обеспечивает легкую синхронизацию между вкладками или же восстановление последней сессии без необходимости обращения к серверу.

На этапе анализа предметной области и решения проблем с интеграцией различных подходов, был рарзработан алгоритм процесса контент-анализа. Как упоминалось ранее, он является одним из трех этапов анализа и является основой работы системы. Как можно увидеть из названия, контент или от английского «content» переводится как содержание или нарратив, поэтому сам текст рассматривается с точки зрения содержания. [применение иллюстрации]

Дискурс-анализ в нашем случае помимо всего использует данные, раннее полученные на этапе контент-анализа: это значимые слова и ключевые слова. Используя алгоритмы критического дискурса, мы можем выделить значимые категории и дать им анализ в соответствии со словарем. Словарь, как правило, содержит несколько трактовок, которые зависят от особенностей конкорданса, из которых выбирается наиболее подходящая.

Психолингвистический этап проводит эмоциональную оценку текста, а так же выделенных личностей дискурса. Этот метод представляет использование теории Фрейда на практике. Данный этап требует непосредственного выполнения двух предыдущих. Психолингвистический анализ состоит из двух различных направлений: анализ личности и анализ текста, причем оба направления мало связаны между собой. Для текста определяется эмоциональная окраска, его общая направленность, а так же рассматриваются отдельные предложения и слова. Личность анализируется с позиции стимулов. На основании этих стимулов дается характеристика ее действиям и словам, если таковые есть.

По видимым причинам должна быть четкая последовательность выполнения (методы могут применяться не полностью, а лишь для получения требуемых промежуточных данных): контент-анализ предшествует дискурс-анализу, после которого только следует психолингвистический анализ.

Как можно видеть, охват данных трех подходов позволяет полностью покрыть область интересов исторической информатики. Ранее не единожды говорилось, что систем, реализующих какой либо из подходов, существует достаточно большое количество, однако, основная ценность заключается именно в возможности совместного их использования. Рассмотренная гибридная модель позволяет не только совместно использовать подходы, но и использовать выходные данные одного из типов анализа в качестве входа на другом. Это можно представить в виде непрерывного потока, состоящий из нескольких этапов:

1. Первый этап. Текстовый массив поступает на обработку;
2. Второй этап. Происходит процесс контент-анализа, где будут выделены только интересующие нас сущности и вычислены различные статистические показатели для них и для данных в целом;
3. Третий этап. Обрабатываются отдельные сущности, выделенные на прошлом шаге. При помощи словарей, а так же полученной ранее статистики, выделяют другие сущности в рамках дискурс-анализа;
4. Четвертый этап. Обрабатываются данные как со второго, так и с третьего этапов. При проведении психоанализа так же используются определенные правила и словари.

Необходимо отметить, что полученный программный продукт будет интересен в первую очередь специалистам, работающим в сферах, где так или иначе используется анализ текста в ходе основной работы или же сам по себе является таковой. Это могут быть лингвисты, исследователи языка, и, в первую очередь, историки. Все объясняется тем, что сервис реализует три основных подхода (контент-анализ, дискурс-анализ и психолингвистический анализ), которые используются в историческом исследовании.

В ходе реализации системы, в качестве одной из задач, была решена проблема интеграции этих подходов. Благодаря использованию таких фреймворков, как Spring Framework и Hibernate на стороне сервера, а так же непосредственно языка Java, появилась возможность грамотно манипулировать данными. Это позволило избежать ненужного дублирования результатов, и организовать их хранение и передача на другие этапы в рамках проекта, и таким образом в полной мере реализовать модель сочетаемости нетрадиционных методов исторического познания текста нарративного источника, описанную в начале.

Хотя данный продукт и находится на стадии бета-тестирвания, он уже пригоден для использования и выполняет основные задачи. Однако это является лишь малой частью его потенциала. Система уже имеет свой вектор развития до полноценной профессиональной сети, основной задачей которой будет уже не только анализа текста, но и использование результатов этого анализа. Такая сеть призвана объединить пользователей и внедрить социальную составляющую, позволяя делиться результатами исследования, источниками, создавать отчеты, статьи и публикации. Помимо всего прочего, сам сервис, обладая огромным массивом документов, который сформирован пользователями, может составлять свой корпус языка. На данный момент реализована поддержка лишь русского языка, но в дальнейшем продукт с легкостью может быть расширен до возможности анализа текстов и других языков, а так же мульти языковых источников.

На данный же момент сервис уже имеет широкий круг применения, а в скором времени будет запущен в тестовом режим в сети интернет на выделенном сервере.

Доклад окончен. Спасибо за внимание.