Таким образом, применение методов хранения и алгоритмов обработки данных больших объёмов в современных автоматизированных системах управления дорожным движением является не просто необходимым, но и закономерным шагом для их эволюции. Благодаря этому поддерживается высокая скорость доступа и обработки данных, получение результатов в режиме реального времени, автоматическая адаптация под изменяющиеся условия на базе алгоритмов машинного обучения, преобразование форматов данных с помощью алгоритмов распознавания изображений. Ко всему этому добавляется высокая отказоустойчивость системы в результате распределённой обработки данных.

Список использованных источников:

1. Развитие автоматизированной системы управления дорожным движением Минска как части интеллектуальной транспортной системы города. [Электронный ресурс]. – Электронные данные. – Режим доступа: https://rep.bntu.by/bitstream/handle/data/27593/Razvitie_avtomatizirovannoj_sistemy_upravleniya_dorozhnym_dvizheniem_Minska.pdf
2. Автоматизированная система управления дорожным движением «АГАТ» [Электронный ресурс]. – Электронные данные. – Режим доступа: https://agat.by/wp-content/uploads/2015/05/АСУДД-АГАТ-презентация.pdf
3. АСУДД: Что висит над дорогой? [Электронный ресурс]. – Электронные данные. – Режим доступа: https://habr.com/post/124249/

ПОСТРОЕНИЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫНА ОСНОВЕ АЛГОРИТМОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Хопова А.А., Волорова Н.А.

Волорова Н.А. – к.т.н., доцент

В данном докладе рассматриваются методы тематического моделирования текстов, а также оценки качества получаемых результатов. При этом в качестве исходных данных используются тексты микроблогов, которые существенно отличаются от традиционных текстов книг, статей и пр. Так же рассматриваются система рекомендаций на основе схожести контента и с учётом реакций других пользователей (лайки, репосты и т.д.). Такая система является более эффективной, чем традиционный подход (фильтрация), благодаря использованию дополнительных метрик при формировании рекомендации. Применение такой системы позволит пользователям находить релевантные материалы, хранящиеся в социальных сетях.

Введение

В наше время количество информации постоянно растёт, большая её часть представляет собой неупорядоченные текстовые данные, например, различные WEB-ресурсы, социальные сети, блоги, форумы, новостные сайты и тому подобное. Человеку сложно самостоятельно их обрабатывать. Более того, ручной анализ неэффективен для больших объемов текста, так как он ограничен скоростью, погрешностями и ошибками, обусловленными человеческим фактором. Следовательно, требуются методы, способные автоматически обрабатывать такие данные. Методы могут основываться на извлечении определенной информации, например, списках тем, содержащихся в текстах.

Для того чтобы выделить из текста основные темы, человеку достаточно его прочитать. В условиях постоянно увеличивающегося количества информации, в частности, текстовой (так называемый, информационный бум), приходится анализировать данные такого объема, которые человек не в силах обработать. Поэтому необходимы методы, позволяющие автоматически извлекать темы из большого набора данных. Теоретически обоснованным и активно развивающимся направлением в анализе текстов на естественном языке, является тематическое моделирование коллекций текстовых документов.

Построение тематической модели может рассматриваться как задача одновременнойкластеризации документов и слов по одному и тому же множеству кластеров, называемых темами. В терминах кластерного анализа тема – это результат би-кластеризации, то есть одновременной кластеризации и слов, и документов по их семантической близости. Как правило, выполняется нечёткая кластеризация, то есть документ можетпринадлежать нескольким темам в различной степени. Таким образом, сжатое семантическое описание слова или документа можно представить в виде распределения на множестве тем. Процесс нахождения этих распределений и называется тематическим моделированием.

В 2003 году Д.Блей предложил модель скрытого размещения Дирихле (Latent Dirichlet Allocation, LDA). Это одна из первых и широко используемых вероятностных тематических моделей. Основной идеей таких моделей является наличие генеративного процесса – процесса, порождающего документы с использованием предопределенных тем. Задача заключается в том, чтобы подобрать темы таким образом, чтобы вероятность сгенерировать данный набор документов была максимальной.

В вероятностных тематических моделях темы представляются в виде распределений над словами. Оценить качество полученных тем можно вручную: можно выбрать слова с наибольшей вероятностью и понять, что они вместе означают. При большом количестве тем требуется много времени, чтобы оценить, насколько понятными для человека они получились.

Алгоритмы поиска наиболее правдоподобных скрытых параметров делятся на две

категории: на основе сэмплирования и вариационные методы. Алгоритмы первой группы пытаются собрать конечную выборку переменных, на которой ищется максимум.

Как правило, алгоритм принадлежит классу методов Монте-Карло для марковскихцепей (MarkovChainMonteCarlo, MCMC). Примером такого алгоритма является сэмплирование по Гиббсу, которое состоит в том, чтобы на каждом шаге фиксироватьвсе переменные, кроме одной, и выбирать оставшуюся переменную согласно распределению вероятности этой переменной при условии всех остальных. Методы второй группы вариационные алгоритмы. В них сначала задается параметризованное семейство распределений над скрытыми переменными, а затем с помощью EM-алгоритма ищется распределение из этого семейства, наиболее близкое к исходному апостериорному распределению.

В качестве языка программирования, на котором выполнялась реализация практических исследований, был выбран язык Java. Это объектно-ориентированный язык, который хорошо подходит для прикладных задач. Кроме того, программы, написанные и скомпилированные на Java можно запускать на любой операционной системе, где поддерживается запуск виртуальной машины Java. При проверке орфографии на этапе предобработки данных использовалась библиотека Snowball и MyStem. Для оценки интерпретируемости с помощью Google использовалась разрабатываемая в ИСП РАН утилита для скачивания веб-страниц из сети Интернет. Обе библиотеки реализованы на Java, что также является доводов в пользу данного языка программирования.

Использовались готовые реализации тематических моделей на языке C. Выбор этихреализаций обусловлен тем, что они принадлежат авторам исследуемых методов тематического моделирования. Кроме того, язык C хорошо подходит для таких задач, где производится большое количество вычислений.

Заключение

В процессе выполнения работы были исследованы методы определения тематической направленности текстового содержимого микроблогов и реализован алгоритм автоматической оценки интерпретируемости результатов тематического моделирования текстов микроблогов. Так же были исследованы существующие методы тематического моделирования и способы оценки их качества. Выполнена экспериментальная оценка интерпретируемости методов тематического моделирования текстов микроблогов с использованием разработанных методов. Разработаны и реализованы методы автоматической оценки интерпретируемости результатов тематического моделирования по ключевым словам тем. А также были проведены расчёты популярности сообщений (твиттов) при помощи метрик, на основании которых был осуществлён прогноз тенденции популярности.

ПРИНЯТИЕ РЕШЕНИЯ О ДЕНОРМАЛИЗАЦИИ БАЗЫ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Чочиева А.С.

Калугина М. А. - к.ф-м.н., доцент

В ходе проектирования баз данных может возникнуть вопрос: “Необходима ли её большая нормализация и если да, то до какой степени?” Большинство источников указывают, что всегда нужно стремиться к нормализации базы данных - “больше таблиц, меньше столбцов”, к большему разделению на сущности. Нормализация является важным процессом проектирования базы данных. Но могут ли у неё иметься недостатки?

Нормализация базы данных — это процесс реструктурирования реляционной базы данных в соответствии с нормальными формами отношения с целью снижения избыточности данных и обеспечения их целостности. Денормализация — обратный этому процесс.

Сравним эти два процесса для выделения очевидных аргументов в пользу денормализации.

Нормализация экономит место, но это место стоит дешевле [1]. В денормализованной базе данных может находиться 10000 записей, например, названия страны, что занимает гораздо больше места, чем просто внешний целочисленный ключ на одну строку в другой таблице. Но терабайтные носители сейчас дешевле, чем ранее.

Нормализация упрощает изменение строки, но чтение обычно происходит чаще, а нормализация замедляет чтение. Таким образом, если данные редко меняются, имеется смысл задуматься о денормализации.

Расчетные данные [3]. Существуют ситуации, когда есть запросы с долгими, сложными вычислениями, потребляющими много ресурсов. В таких случаях можно выделить столбцы, в которых будут храниться часто используемые расчётные значения, чтобы избежать потери ресурсов и времени на их повторное вычисление.

Нормализация снижает производительность. Это и есть наиболее часто упоминаемая причина денормализации. Сильная нормализация может привести к большому количеству соединений таблиц, что выливается в объемные составные SELECT-запросы с большим количеством JOIN операций- это будет замедлять работу базы данных. В источниках также упоминается [1], что нормализация часто становится причиной неполадок в работе программного обеспечения из-за ошибок в сложных запросах. В некоторых случаях целесообразно хранить редко изменяемые данные просто в текстовом файле, загружая их по мере необходимости.

С учетом проведенного анализа, денормализацию не рекомендуется применять в следующих случаях: