Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.43

Хопова
Александра Ашотовна

СИСТЕМА АНАЛИЗА ПОПУЛЯРНОСТИ БЛОГГЕРОВ В СОЦИАЛЬНЫХ СЕТЯХ

АВТОРЕФЕРАТ

на соискание степени магистра информатики и вычислительной техники

по специальности 1-40 81 04 Обработка больших объёмов информации

Научный руководитель
Волорова Наталья Алексеевна
кандидат технических наук, доцент

Минск 2018

АННОТАЦИЯ

Часто при анализе текстовых документов, как ручном, так и автоматическом, не нужны полные тексты исследуемых документов. Достаточно лишь небольшого количества информации о них. Примером такой информации могут быть темы, затрагиваемые в каждом документе.

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.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В данной работе были исследованы методы определения тематической направленности текстового содержимого микроблогов и реализован алгоритм автоматической оценки интерпретируемости результатов тематического моделирования текстов микроблогов.

Были выполнены следующие задачи:

– исследованы существующие методы тематического моделирования и способы оценки их качества;

– разработаны и реализованы методы автоматической оценки интерпретируемости результатов тематического моделирования по ключевым словам тем;

– выполнена экспериментальная оценка интерпретируемости методов тематического моделирования текстов микроблогов с использованием разработанных методов.

В работе было показано, что методы автоматической оценки обладают положительной корреляцией с оценками экспертов, поэтому есть основания применять их для анализа тематических моделей. Также было показано, что непараметрические модели выдают менее интерпретируемые темы, чем параметрические, при условии выбора правильных параметров.

СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из 5 разделов:

– Анализ предметной области;

– Анализ существующих решений;

– Исследование и построение решения задачи;

– Описание практической части;

– Заключение.

ЗАКЛЮЧЕНИЕ

В работе было показано, что методы автоматической оценки обладают положительной корреляцией с оценками экспертов, поэтому есть основания применять их для анализа тематических моделей. Также было показано, что непараметрические модели выдают менее интерпретируемые темы, чем параметрические, при условии выбора правильных параметров.

Дальнейшие направления развития:

–улучшение процесса нормализации текстов микроблогов;

–поиск или разработка тематических моделей, ориентированных на короткие документы;

–исследование методов объединения нескольких сообщений микроблогов в один документ для лучших результатов тематического моделирования с использованием традиционных моделей;

–исследование и разработка новых подходов для оценки интерпретируемости результатов тематического моделирования.

ПУБЛИКАЦИЯ
ПОСТРОЕНИЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ АЛГОРИТМОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

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

Хопова Александра Ашотовна, Волорова Наталья Алексеевна
Волорова – к.т.н.,доцент

В данном докладе рассматривается 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

Часто при анализе текстовых документов, как ручном, так и автоматическом, не нужны полные тексты исследуемых документов. Достаточно лишь небольшого количества информации о них. Примером такой информации могут быть темы, затрагиваемые в каждом документе.

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 одновременной кластеризации 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 может принадлежать 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 марковских цепей (Markov Chain Monte Carlo, MCMC). Примером такого алгоритма является сэмплирование по Гиббсу, которое состоит в том, чтобы на каждом шаге фиксировать все 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 этих реализаций 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 так же были проведены расчёты популярности сообщений (твиттов) при помощи метрик, на основании которых был осуществлён прогноз тенденции популярности.