УДК 004

Е.Н. Кореняко, О.В. Миролюбова, Е.И. Пермякова

ЛАТЕНТНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ КАК МЕХАНИЗМ ПОЛУЧЕНИЯ ОБРАТНОЙ СВЯЗИ ОТ ПОТРЕБИТЕЛЕЙ

Поиск новых подходов к продвижению продукции становится важнейшим в исследованиях международных компаний. Очевидно, что современный потребитель получает необходимую информацию о компании, о продукции, о вакансиях, об акциях и т.п. с сайтов компании. Чем большую полезность для потребителя несет сайт компании, тем более благоприятное впечатление он производит на потребителя и, следовательно, влияет на положительное восприятие компании в целом. Надо отметить, что потребителю также важны мнения потребительских сообществ о продукции выбранного производителя и об аналогичной продукции. Такую информацию можно найти посредством поисковых запросов в сети или на специализированных форумах. Однако качество таких отзывов может очень разниться. Для международных компаний принципиально важно проводить анализ отзывов или откликов потребителей на качество продукции в целом, иметь возможность своевременно принять меры по устранению объективных причин недовольства или развить позитивный тренд развития.

В современном мире латентно-семантический анализ (ЛСА) является одним из наиболее перспективных методов обработки информации. Метод заключается в обработке информации на естественном языке, при этом во внимание принимается взаимосвязь между коллекцией документов и терминами. Данной взаимосвязи сопоставляются некоторые факторы, которые определяют тематику документа.

Основой для возникновения латентно-семантического анализа послужили принципы факторного анализа. В результате анализа выявляются латентные связи между изучаемыми явлениями и

объектами. Кроме того, широкое применение метод получил при классификации или кластеризации документов. На первом этапе большие объемы текста подвергаются статистической обработке. На последующих этапах с помощью латентно-семантического анализа извлекаются контекстно-зависимые значения лексических единиц.

Изначально ЛСА применялся в сфере автоматического индексирования текстов и выявления семантической структуры текста. Затем этот метод был довольно успешно использован для представления баз знаний и построения когнитивных моделей.

В последние годы метод ЛСА нашел свое применение в областях, где необходимо выявление главных факторов из массива информационных данных. Наибольшее распространение метод получил в сферах поиска информации, индексации документов, классификации документов, а также в моделях понимания.

Принцип работы метода ЛСА можно представить, описав взаимосвязь между тремя слоями. Первый слой состоит из некого множества слов. Второй слой представляет собой множество документов, которые соответствуют конкретным ситуациям. Третий слой является средним слоем, который также можно назвать скрытым. Этот слой представляет собой множество узлов с различными весовыми коэффициентами, которые связывают первый и второй слои.

Метод ЛСА часто подразумевает использование матрицы, описывающей встречаемость терминов в документах. Ряды матрицы соответствуют терминам, столбцы – документам. Определение значимости элемента матрицы происходит следующим образом. Элемент матрицы пропорционален количеству его появлений в документе. В тоже время редко встречающимся терминам присваиваются высокие весовые значения, что отражает их относительную важность.

Описанная матрица во многом схожа со стандартными семантическими моделями, однако она не всегда явно выражена в виде матрицы, так как не всегда необходимо использовать свойства матриц.

После построения матрицы появлений согласно методу ЛСА необходимо произвести понижение ранга[1]. Существует ряд причин для такого понижения:

- Размер оригинальной матрицы превышает допустимые значения компьютера. В данном случае понижение представляется необходимым шагом.
- В оригинальной матрице присутствуют шумы. После понижения ранга новая матрица считается лучше отражающей действительность.
- Оригинальная матрица признана редко соответствующей документу. То есть, матрица содержит только слова, имеющиеся в документе. Однако нам интересен намного больший объем данных, содержащий все слова, относящиеся к документу, включая синонимы.

Следствием понижения ранга матрицы является комбинирование некоторых ее элементов, которые приобретают зависимость более чем от одного термина.

Это, в свою очередь, облегчает решение проблемы определения синонимов, так как в результате понижения ранга матрицы ожидается слияние размеров, ассоциированных со схожими значениями. Кроме того, понижение ранга способствует решению проблемы многозначности слов, так как компоненты группируются по принципу схожести значений.

Метод ЛСА обладает многочисленными достоинствами. Во-первых, метод является наилучшим инструментом для выявления скрытых зависимостей во множестве разнородных документов. Во-вторых, для применения метода необязательно ручное обучение машины. Для определенных целей, таких как кластеризация, достаточно автоматизации. В-третьих, в методе используются значения матрицы близости, которые основаны на частотных характеристиках документов и лексических единиц. В-четвертых, благодаря методу ЛСА частично снимаются такие проблемы распознавания, как полисемия и омонимия.

Несмотря на свою перспективность, у метода ЛСА наблюдается ряд недостатков. Во-первых, существенным недостатком метода является значительное снижение скорости вычисления при увеличении объема входных данных. Во-вторых, вероятностная модель метода не соответствует реальности. В основе метода лежит предположение, что слова и документы имеют нормальное распределение. Однако, как показывает накопленный опыт, ближе к действительности оказывается распределение Пуассона. В связи с этим для практических применений наилучшим образом подходит усовершенствованная модель метода – Вероятностный латентно-семантический анализ, которая основана на мультиноминальном распределении.

Задача продвижения сайта компании это привлечение на него посетителей из выдачи поисковых систем. Поэтому важно составить семантическое ядро так, чтобы оно соответствовало популярным запросам пользователей при поиске товаров и услуг в наиболее популярных поисковых системах Яндекс и Google.

В зависимости от того, насколько часто те или иные слова используются при поиске, выделяют: запросы общего характера, которые не позволяют выявить потребность пользователя («водонагреватель»); уточненные запросы («купить водонагреватель»); максимально точные запросы («купить водонагреватель плоский 80 л»). Несмотря на то, что такие запросы совершают всего лишь около 100-200 человек в месяц, они могут обеспечить наибольшую конверсию, поскольку пользователи максимально заинтересованы в предмете поиска.

Такая же система действует в систематизации отзывов по конкретной продукции. Выделяются наиболее важные ключевые слова, которые интересны производителю для реальной оценки продукции. Перечень ключевых таких слов, отражающих направленность и тематику, составляет семантическое ядро для сайта, для анализа данных. Для крупных сайтов данный список может насчитывать несколько тысяч слов. Определение семантического

ядра – это основа для формирования стратегии продвижения. В зависимости от смыслового поля выполняется техническая оптимизация, выбираются целевые страницы, формируется наполнение сайта и т. д. Семантическое ядро обеспечивает результативность продвижения, помогает получить целевых посетителей. Для составления семантического ядра, во-первых, необходимо определить все возможные запросы, которые описывают содержание сайта или удовлетворяют критериям аналитика направленности отзыва, во-вторых, удалить запросы с низкой частотностью, анализ по которым не принесет желаемого результата, в-третьих, распределить получившийся список между страницами сайта. По наиболее конкурентным запросам, как правило, продвигают главную страницу и страницы с наибольшим статическим весом (на них больше всего ссылок с внутренних страниц и других сайтов). Другие запросы группируются и распределяются между остальными страницами сайта.

При создании семантического ядра необходимо учитывать следующие моменты:

- в составе ядра должны присутствовать как общие, так и «узкие» запросы. Использование исключительно низкочастотных запросов приведет целевых посетителей на сайт, но объем трафика будет намного меньше, чем по высокочастотным ключевым словам. Преобладание общих запросов негативно отразится на поведенческих факторах;
- обязательно при составлении семантического ядра нужно использовать ассоциативные ключевые слова (по смежным темам). Это сделает тексты более привлекательными для поисковых систем и посетителей сайта;
- нельзя пренебрегать ключевыми словами с ошибками, которые пользователи могут сделать по невнимательности («покупка афтомабиля»). Поисковые системы находят ответы и для таких запросов;

- количество ключевых слов зависит от объема текста на странице. Чтобы тексты были читаемыми и привлекательными для поисковых систем и посетителей, частота их использования не должна превышать 7% .

Таким образом, компьютерное понимание текста достигается за счёт погружения текста в единую среду знаний, представления смысла в памяти компьютера и возможности операций надонтологическим смыслом. Технологию семантического анализа можно использовать для формирования баз данных, архивирования электронных документов, их индексирования, классификации и поиска в Интернет. Также представляется возможность создавать интеллектуальные банки данных, работающие в единой среде знаний.

Еще одной проблематикой бытия информационного века является то, что в настоящее время в интернете существует множество тематических форумов и блогов, содержащих кладезь важной для определенных предметных областей информации. В рамках задач сбора информации для контроля и определения качества услуг или товаров – т. е. для задач сбора и ранжирования отзывов в интернете важно выстроить определенного рода базу данных или индекс, позволяющий провести анализ открытых текстов на предмет того какой именно окрас отзыва носит этот текст если мы определили этот текст как отзыв.

Такой анализ под силу провести системам, использующим в своей работе онтологию. Онтология – это спецификация концептуализации предметной области, т. е. это отображение какой-либо предметной области, состоящее из словарей терминов и аксиом, логически описывающих отношения между терминами.

Формирование онтологии требует понимания анализируемой предметной области, т. к. предполагает формирование использование больших словарей с терминами предметных областей. Формирование таких словарей вручную занимает довольно приличный объем времени, что повышает важность задачи по автоматизированному построению онтологии.

Онтологию можно построить, используя следующие подходы:

- Представление онтологий в виде конечного автомата
- Построение семантической карты ресурса
- Подход на основе лексико-синтаксических шаблонов
- Автоматическое построение онтологии по коллекции текстовых документов

В рамках статьи по ЛСА мы рассмотрим только четвертый метод построения, т.к. он тесно связан с применением одной из составляющих ЛСА – словарем или концептуальным индексом, сформированным по результатам работы ЛСИ.

Концептуальные индексы создаются с помощью латентной семантической индексации (ЛСИ), задача которой – построение выводов о семантической близости терминов, собранных в словарь. Рекомендуется использовать в процессе формирования концептуальных индексов именно статистические подходы, т.к. они не зависят от лингвистических особенностей языка анализируемого текста.

Автоматическое построение онтологии по коллекции текстовых документов разделено на 3 этапа:

- Подготовка текстовых данных;
- определение классов онтологии;
- определение отношений «is-a» и «synonym-of», построение иерархии классов.

Качество текстовых данных как исходного материала определяет качество результата работ, поэтому важно обеспечить отсутствие как минимум орфографических ошибок в анализируемых текстах.

Кластеризация (выделение классов по признаку) документов по общей тематике сократит время формирования онтологии с учетом работы по каждому кластеру в отдельности. В случае с LSA кластеризация проводится по терминам из состава концептуального индекса определенной предметной области.

Таким образом, мы видим, что применение ЛСА в комбинации с другими алгоритмами дает вполне осязаемый практический результат: анализируя кластер документов – мы можем дать ответ практически на любой вопрос по их содержанию. В частности при анализе отзывов потребителей о товарах и услугах мы знаем, что анализируемые тексты являются именно отзывами и можем составить на основе этого кластера концептуальный индекс (словарь терминов) и, самое важное – автоматизировать построение связей между терминами словаря, что позволит нам понять содержится в тексте отзыва положительный или отрицательный отзыв.

Латентно-семантический анализ (ЛСА) является наиболее изученным и доступным в настоящее время алгоритмом анализа текстовой информации сети Интернет. Этот алгоритм используют в работе ведущие поисковики, что подтверждает его действенность и гибкость.

Применение ЛСА в маркетинге и контроле качества услуг и товаров принесет огромную пользу за счет анализа уже накопившихся в сети массивов информации и отзывов по продуктам, услугам и компаниям. Анализ этого массива информации предоставит максимально адекватную систему рейтингов товаров и услуг, т.к. не будет зависеть от конкретных ресурсов интернет. Такая независимость достигается за счет анализа как открытых, так и ангажированных ресурсов. Влияние ангажированных ресурсов на оценку можно значительно снизить добавлением в комплекс алгоритмов ЛСА еще один алгоритм фильтрующий список неправдоподбных ресурсов, составленный заранее оператором или взятый из других баз данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК:

1. Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы. Учебное пособие. Казань, Москва. 2006;

2. Рабчевский Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска. // Труды 11-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL 2009. – Петрозаводск, 2009. – С. 69–77.

3. Tamara G Kolda, Dianne P O'Leary A Semi-Discrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval. 1996

4. William B.Frakes and Ricardo Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cli_s, New Jersey, 1992.

5. Berry M.W., S.T. Dumais& G.W. O'Brien Using Linear Algebra for Intelligent Information Retrieval 1994

6. Dumais S. Improving the retrieval of information from external sources. BehaviorResearchMethods, Instruments, &Computers, 23(2):229--236, 1991

7. Todd A Letche, M.W. Berry, Large-Scale Information Retrieval with Latent Semantic Indexing. 1996

8. Markovsky I. Low-Rank Approximation: Algorithms, Implementation, Applications, Springer, 2012.