ПРИМЕНЕНИЕ МОДИФИКАЦИЙ АЛГОРИТМА APRIORI ДЛЯ АСПЕКТНОГО АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

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

Иванин Н.С., Аксамит М.В., Макович Е.А.

Стержанов М.В. – к.т.н., доцент

В настоящие время значительно возросло количество пользователей, делящихся своим мнением о различных продуктах. Важной задачей является определение понравился или не понравился пользователю определенный продукт. Однако часто определения настроения о продукте в целом недостаточно. Тогда необходимо определять мнение пользователя о каждом аспекте продукта в отдельности. Для этого могут быть применены алгоритмы, основанные на алгоритме Apriori.

Алгоритмы анализа тональности текстов предназначены для определения тональности целого текста либо его фрагмента. В таком подходе предполагается, что исходный текст является мнением автора о каком-то одном конкретном объекте, например, ресторане или книге. Однако в некоторых доменах в отзыве о объекте или сущности так же содержится мнение автора о ее составляющих. Например, в отзывах о мобильных телефонах могут быть оценены такие части телефона, как экран, камера, аккумулятор, батарея. Основной задачей алгоритмов аспектного анализа является выделение аспектов и определение тональности отзыва пользователя об каждом аспекте.

Алгоритм Apriori[1] был разработан для извлечения ассоциативных правил[2]. Извлечение ассоциативных правил используется для обнаружения различных отношений между объектами в больших объемах данных. Чаще всего этот подход применяется для анализа рыночной корзины. Например, правило $\{молоко\} \to \{хлеб\}$ означает, что покупатель, купивший молоко также купит хлеб. Apriori осуществляет поиск по множеству покупок пользователей, называемому множество транзакций. Говорят, что правило $\{X\} \to \{Y\}$ имеет поддержку support, если support % транзакций из множества транзакций содержат $X \bigcup Y$ и достоверность confidence, если confidence % транзакций из множества транзакций, содержащих X также содержат Y. Алгоритм состоит из двух шагов:

1. Генерация часто встречающихся множеств элементов. Задача этого шага найти все множества элементов, поддержка которых выше определенного порога. Такие множества элементов называются часто встречающиеся множества элементов.

2. Генерация правил. Задача этого шага извлечь правила, имеющие высокую поддержку, из часто встречающихся множеств элементов, полученных на предыдущем шаге. Такие правила называются сильные правила.

В работе [3] для извлечения аспектов используется алгоритм, основанный на первом шаге CBA[4].Алгоритм CBA был разработан для извлечения ассоциативных и классификационных правил. Первый шаг называется генерация правил, он основан на алгоритме Apriori, описанном выше. Второй шаг называется построение классификатора. Шаг генерация правил состоит из двух частей:

1. Нахождение часто встречающихся множеств элементов. Делая обход данных и рассчитывая поддержку множеств элементов, алгоритм определяет, являются ли множества часто встречающимися.

2. Генерация новых множеств-кандидатов. Используя множества элементов меньшей размерности, полученные на предыдущем шаге, алгоритм генерирует множества-кандидатов, которые на следующей итерации первого шага рассматриваются как возможные часто встречаемые множества.

В работе [3] отмечается, что в качестве аспектов выступают существительные и именные группы. Длина извлекаемых именных групп при этом обычно не превосходит трех. Поэтому на первом шаге производится частеречная разметка каждой транзакции. После этого к каждому отзыву необходимо осуществить предобработку. Это необходимо для исключения слов, которые обычно не являются аспектами. Предобработка включает удаление стоп-слов, стемминг, лематизацию и исправление написания слов. На следующем шаге алгоритм извлекает часто встречающиеся множества элементов. Каждый элемент в этом множестве это возможный аспект. Для извлечения полезных и подлинных аспектов используется фильтрация. Авторы предлагают использовать два типа фильтрации: фильтрация на основе компактности (среди кандидатов длины 2 и более удаляются те, составляющие которых отстоят друг от друга на большом расстоянии) и фильтрация лишних кандидатов (среди кандидатов длины 1 удаляются те, которые определенное число раз входят в кандидаты большей длины). Затем осуществляется поиск оценочных оборотов. Для каждого отзыва, который содержит аспект, извлекается ближайшее прилагательное. Если такое прилагательное найдено, то оно рассматривается как оценочный оборот. Так же данный подход позволяет извлекать аспекты, упомянутые только несколькими пользователями. Для этого из каждого отзыва, который не содержит аспектов, но содержит оценочный оборот, извлекается наиболее близкое к оценочному обороту существительное или именная группа.

В работе [5] предлагается улучшение алгоритма Apriori. Это улучшение также может быть применено для извлечения аспектов из текста. Авторы предложили две основных техники сокращения транзакций: локальное сокращение и глобальное сокращение, которые позволяют сократить количество транзакций в

множестве транзакций, а также уменьшить размер отдельных транзакций. Другое улучшение было предложено для подсчета часто встречающихся элементов размера 2. Для этого была введена специальная хэш-функция, которая каждому такому множеству ставит в соответствие число. Это позволяет значительно ускорить поиск подмножеств каждой транзакции среди кандидатов и таким образом ускорить подсчет поддержки для кандидатов. Авторы отмечают, что именно подсчет множеств размера 2 занимает до 90 процентов времени работы алгоритма Apriori. Введенная хэш-функция позволяет также ускорить подсчет множеств большей размерности. Для этого все кандидаты, имеющий одинаковый префикс длины 2 объединяются в корзины. Далее, если мы хотим найти некоторое множество среди кандидатов, то мы, взяв префикс длины 2, сможем с лёгкостью определить нужную корзины с множествами элементов и продолжить поиск в этой корзине.

Списокиспользованныхисточников:

1. Agrawal R., Srikant R. Fast algorithm for mining association rules // VLDB'94.
2. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Pearson Addison-Wesley
2. Hu M., Liu B. Mining opinion features in customer reviews // Proceedings of the 19th national conference on Artifical intelligence, July 25-29, 2004, San Jose, California - San Jose, California, 2004- c.755-760.
3. Liu, B., Hsu, W., Ma, Y. Integrating Classification and Association Rule Mining // KDD-98, 1998.
4. Orlando S., Palmerini P., Perego R. Enhancing the Apriori Algorithm for Frequent Set Counting // Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery, p.71-82, September 05-07, 2001 – c71-82.