УДК 004

О.О. Філонова, С.М. Вороной
Донецкий национальный технический университет, г.Донецк
кафедра систем искусственного интеллекта

АЛГОРИТМ КЛАСТЕРИЗАЦИИ ПОИСКОВЫХ ПРОФИЛЕЙ ПОЛЬЗОВАТЕЛЕЙ ДЛЯ СИСТЕМЫ ПЕРСОНАЛИЗАЦИИ САЙТА

Аннотация

Філонова О.О. **Вороной С.М.** ***Алгоритм кластеризации поисковых профилей пользователей для системы персонализации сайта.*** *Выполнен анализ методов кластеризации больших объемов категорийных данных. Выбран эффективный метод и предложен алгоритм кластеризации поисковых профилей пользователей для систем персонализации сайтов, отличающийся небольшой вычислительной сложностью и масштабируемостью*

Ключевые слова: *методы кластеризации, категорийные данные, поисковые профили, системы персонализации, масштабируемость алгоритма.*

Постановка проблемы. Для крупных Web-порталов существует актуальная задача эффективной навигационной и поисковой поддержки пользователей. Эту задачу можно решать путем персонализации содержимого в соответствии с потребностями и особенностями поведения конечного пользователя. Недостатком существующих средств персонализации является ориентация только на текущие потребности пользователей, что снижает точность сформированных системой персонализации рекомендаций [1].

Выявление ранее не просмотренных конечным пользователем страниц удовлетворяющих его постоянным потребностям может осуществляться с учетом совокупности поисковых запросов, имевших место у того или иного пользователя с похожими профилями.Чтобы выявить такие страницы необходимо провести кластеризацию поисковых профилей и расширить поисковый профиль на основе соответствующего кластера. Универсальные алгоритмы кластеризации не эффективны при решении этой задачи из-за больших объемов и категорийности данных, высоких требований к скорости обработки. Поэтому актуальна задача разработки алгоритма кластеризации, ориентированного на использование в системах персонализации сайтов.

Цель статьи – анализ методов кластеризации больших объемов категорийных данных и разработка алгоритма кластеризации адаптированного к применению в системах персонализации.

Анализ методов кластеризации больших объемов данных. Задачу кластеризации в том или ином виде формулировали в таких научных направлениях, как статистика, распознавание образов, оптимизация, машинное обучение. Отсюда многообразие синонимов понятию кластер – класс, таксон, сгущение. На сегодняшний момент число методов разбиения групп объектов на кластеры довольно велико - несколько десятков алгоритмов и еще больше их модификаций. Рассмотрим некоторые из них с точки зрения применения для решения задачи персонификации в Web.

Алгоритм CURE (Clustering Using REpresentatives). Выполняет иерархическую кластеризацию с использованием набора определяющих точек для помещения объекта в кластер. Предназначен для кластеризация очень больших наборов числовых данных. Эффективен для данных низкой размерности, работает только на числовых данных.

Достоинством являются возможность кластеризации на высоком уровне даже при наличии выбросов, выделение кластеров сложной формы и различных размеров, линейно зависимые требования к месту хранения данных и временная сложность для данных высокой размерности. К недостаткам следует отнести необходимость задания пороговых значений и количества кластеров. Алгоритм описан в [2]

Алгоритм MST (Algorithm based on Minimum Spanning Trees). Алгоритм минимального покрывающего дерева строит граф из N-1 рёбер так, чтобы они соединяли все N точек и обладали минимальной суммарной длиной. Такой граф называется кратчайшим незамкнутым путём, минимальным покрывающим деревом или каркасом графа. Выполняет кластеризацию больших наборов произвольных данных, выделяет кластеры произвольной формы, в т.ч. кластеры выпуклой и вогнутой формы, выбирает из нескольких удачных решений самое оптимальное. К недостаткам алгоритма относятся:

- ограниченная применимость. Алгоритм наиболее подходит для выделения кластеров типа сгущений или лент. Наличие разреженного фона или «узких перемычек» между кластерами приводит к неадекватным результатам;
- высокая трудоёмкость — для построения кратчайшего незамкнутого пути требуется O(N^3) операций;
- чувствительность к выбросам.

Описание алгоритма приведено в [3]

Алгоритм CLOPE. Предназначен для кластеризации огромных наборов категорийных данных. К достоинствам относятся высокие масштабируемость и скорость работы и качество кластеризации, что достигается использованием глобального критерия оптимизации на основе максимизации градиента высоты гистограммы кластера. Отличается простотой программной реализации. Во время работы алгоритм хранит в памяти небольшое количество информации по каждому кластеру и требует минимальное число сканирований набора данных. CLOPE автоматически подбирает количество кластеров, причем это

регулируется одним единственным параметром - коэффициентом отталкивания. CLOPE предложен в 2002 году группой китайских ученых [4]. При этом он обеспечивает более высокую производительность и лучшее качество кластеризации в сравнении с многими иерархическими алгоритмами.

Алгоритм k-средних (k-means). Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции. Заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Достоинство алгоритма в простоте использования, понятности и прозрачности алгоритма. К недостаткам относятся чувствительность к выбросам, которые могут искажать среднее, медленная работа на больших базах данных; необходимость задания количество кластеров. Описан алгоритм в [5].

Анализ алгоритмов показывает, что из всех рассмотренных только CLOPE удовлетворяет необходимым требованиям.

Применение алгоритма CLOPE для кластеризации поисковых профилей. Рассмотрим алгоритм CLOPE применительно к задаче кластеризации поисковых профилей.

Пусть имеется база поисковых профилей D, состоящая из множества поисковых профилей $\{t_1,t_2,\ldots,t_n\}$. Каждый профиль t_i есть набор поисковых запросов $\{i_1,\ldots,i_m\}$. Множество кластеров $\{C_1,\ldots,C_k\}$ есть разбиение множества $\{t_1,\ldots,t_n\}$, такое, что $C_1\cup\ldots\cup C_k=\{t_1,\ldots,t_n\}$ и $C_i\neq\varnothing\wedge C_i\cap C_j=\varnothing$ для 1<=i, j<=k. Каждый элемент C_i называется кластером, n, m, k – количество профилей, количество запросов в базе профилей и число кластеров соответственно.

Каждый кластер C имеет следующие характеристики: D(C) – множество уникальных поисковых запросов; Occ(i,C) – частота вхождений запроса i в кластер C; W(C) = |D(C)|; H(C) = S(C)/W(C);

$$S(C)=\sum_{i\in D(C)}Occ(i,C)=\sum_{t_i\in C}|t_i|$$

Гистограммой кластера C называется графическое изображение его расчетных характеристик: по оси OX откладываются объекты кластера в порядке убывания величины Occ(i,C), а сама величина Occ(i,C) – по оси OY (рис. 1).

Очевидно, что чем больше значение H, тем более "похожи" два профиля. Поэтому алгоритм должен выбирать такие разбиения, которые максимизируют H. Для более качественного разбиения вместо H(C) можно использовать градиент $G(C)=H(C)/W(C)=S(C)/W(C)^2$.

Задача кластеризации сводится к нахождению такого разбиения множества поисковых профилей на кластеры, при котором глобальная функция стоимости имеет максимальное значение:

$$profit(C_i, r) \to \max,$$

$$profit(C_i, r) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} |C_i|},$$

где $|C_i|$ – количество объектов в i-том кластере, k – количество кластеров, r – положительное вещественное число большее 1.

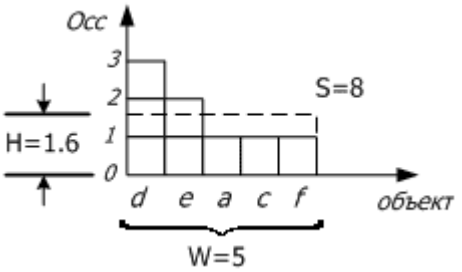

Рисунок 1. Пример гистограммы кластера

С помощью параметра r, названного авторами CLOPE *коэффициентом отталкивания* (repulsion), регулируется уровень сходства транзакций внутри кластера, и, как следствие, финальное количество кластеров. Этот коэффициент подбирается пользователем. Чем больше r, тем ниже уровень сходства и тем больше кластеров будет сгенерировано.

Рассмотрим реализацию алгоритма. Пусть поисковые профили хранятся в таблице базы данных. Для построения начального разбиения, определяемого функцией Profit(C,r) требуется первый проход по таблице профилей. После этого требуется незначительное (1-3) количество дополнительных сканирований таблицы для повышения качества кластеризации и оптимизации функции стоимости. Если в текущем проходе по таблице, изменений не произошло, то алгоритм прекращает свою работу.

При построении начального разбиения из таблицы читается очередной профиль и создается новый кластер (отдельная таблица) или помещается в уже существующий кластер, который дает максимум Profit(C,r).

На итерационном этапе просматривается таблица профилей и для каждого профиля решается задача определения кластера, если новый кластер

C_j максимизирует Profit(C,r), то профиль переносится в этот кластер. В начале каждого цикла устанавливается индикатор перемещения moved := false. Если в цикле происходит перемещение профиля индикатор перемещения изменяется moved := true. Итерации завершаются, если значение moved = false не изменится. После завершения итераций удаляются все пустые кластеры.

Алгоритм CLOPE является масштабируемым, поскольку способен работать в ограниченном объеме оперативной памяти компьютера. Во время работы в оперативной памяти хранится только текущая транзакция и небольшое количество информации по каждому кластеру, которая состоит из: количества транзакций N, числа уникальных объектов (или ширины кластера) W, простой хэш-таблицы для расчета Occ(i,C) и значения S площади кластера.

В результате кластеризации поисковый профиль конечного пользователя i окажется в определенном кластере C*.

Для предъявления пользователю не просмотренных страниц, соответствующих постоянной информационной потребности, производится расширение его поискового профиля. Для этого поисковые запросы, входящие в состав кластера C*, ранжируются по частоте их вхождения в кластер. В расширенный поисковый профиль выбирается некоторое количество lm поисковых запросов с наибольшей частотой вхождения.

Выводы. Выполнен анализ методов кластеризации больших объемов данных и выбран метод подходящий для решения задачи кластеризации поисковых профилей. Разработан масштабируемый алгоритм, ориентированный на применение в системах персонализации для Web порталов. В результате кластеризации устанавливаются близкие поисковые профили пользователей и на основе этого выявляются ранее не просмотренные пользователем страницы соответствующие его постоянным информационным потребностям.

Список литературы

1. Koren, Y.; Bell, R. & Volinsky, C., "«Matrix Factorization Techniques for Recommender Systems.»", Computer (IEEE) . — T. 42 (8): 2009, P. 30-37

2. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim «CURE: An Efficient Clustering Algorithm for Large Databases». SIGMOD '98 Proceedings of the 1998 ACM SIGMOD international conference on Management of data P. 73-84

3. Daniel Fasulo «An Analysis Of Recent Work on Clustering Algorithms». Department of Computer Science & Engineering. Box 352350. University of Washington.1999

4. Yang, Y., Guan, H., You. J. CLOPE: A fast and Effective Clustering Algorithm for Transactional Data In Proc. of SIGKDD’02, July 23-26, 2002, Edmonton, Alberta, Canada

5. Hartigan, J. A., and Wong, M. A. “A K-means clustering algorithm,” *Applied Statistics*, (1979), 28, 100–108.