***М.А. Басов, П.К. Моор***
*Тюменский государственный университет, г.Тюмень*

**УДК 004.9**

# ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ПРИ РАЗРАБОТКЕ СОВЕТУЮЩИХ СИСТЕМ

**Аннотация.** В данной статье представлено применение кластерного анализа в советующей системе. Подробно рассмотрен алгоритм кластеризации «горного» метода. Рассмотрено применение на примере.

**Ключевые слова:** кластеризация, коллаборативная фильтрация, интернет-магазин, информационная система.

В настоящее время покупки в интернет-магазинах перестали быть чем-то выходящим за пределы нормального, ведь и цены в подобных магазинах обычно ниже, и ассортимент больше. С ростом популярности интернет-магазинов растет количество магазинов и увеличивается их ассортимент. Возникают проблемы, которые негативно сказываются как на магазинах, так и на покупателях. Большой ассортимент — это не только плюс, но и минус. Покупатель, видя перед собой много товаров и их параметров, может не найти среди них то, что нужно. В отличие от обычного магазина как правило здесь нет продавца-консультанта. После покупки, клиенты обычно покидают интернет магазин, а ведь ему можно предложить докупить что-то еще к своей покупке. Большинство интернет магазинов пользуются этим, и увеличивают свою прибыль.

Эту проблему можно решить, реализовав систему подбора предлагаемых товаров, называемую Product Adviser (Product Recommendation System). Данная система является подклассом систем фильтрации информации, которые на основе характеристик предметов или социального окружения пользователя, предсказывают предпочтения, который пользователь дал бы.

В качестве решения был рассмотрен алгоритм рекомендательных систем, основанный на коллаборативной фильтрации, который представлен в [1], т.е. на анализе между пользователями и кластеризации «горного» метода, с которым

можно познакомиться в книге [2].

На *первом шаге* горной кластеризации определяют точки, которые могут быть центрами кластеров. Расстояние между точками считается с помощью евклидовых расстояний по формуле

$$D(x1,x2) = \sqrt{\sum_{k=1}^{n}(x1_k - x2_k)^2} \quad (1)$$

Где k – количество измерений.

Так, как о центрах ничего неизвестно, то в качестве потенциальных центров удобно принять сами объекты. Обозначится множество потенциальных центров кластеров $Z_h$, где $h = 1, \dots, Q$. Q может быть определено по методике, предложенной в [1] или задана экспертом.

Если центрами кластеров считаются сами объекты, то тогда $Z_h = x_k$.

На *втором шаге* для каждой такой точки рассчитывается значение потенциала, показывающего возможность формирования кластера в ее окрестности. После этого итерационно выбираются центры кластеров среди точек с максимальными потенциалами. Потенциал центров кластеров рассчитывается по следующей формуле:

$$P_1(Z_h) = \sum_{k=1}^{M} exp\left(-\alpha D(Z_h, x_k)\right) \quad (2)$$

где $M$ - количество объектов $x_i$ , exp – функция экспоненты, $\alpha$ – положительная константа, характеризующая масштаб расстояний между объектами. В простейших случаях удобно полагать, что $\alpha$ равно единице, деленной на среднее расстояние между объектами.

*На третьем шаге* алгоритма в качестве центров кластеров выбирают координаты "горных" вершин. Для этого центром первого кластера $V_1$ назначают точку с наибольшим потенциалом. Чтобы выбрать следующий центр кластера необходимо вначале исключить влияние только что найденного кластера.

Перерасчет потенциала происходит по формуле:

$$P_2(Z_h) = P_1(Z_h) - P_1(V_1) * exp(-\beta * D(Z_h, V_1)) \quad (3)$$

где $\beta$ - положительная константа, характеризующая масштаб размера одного кластера (чем значение $\beta$ меньше, тем больше размер кластера). В простейших случаях ее можно принять равной $\alpha$.

Центр второго кластера $V_2$ определяется как точка с максимальным значением потенциала $P_2(Z_h)$. Затем снова пересчитывается значение потенциалов:

$$P_3(Z_h) = P_2(Z_h) - P_2(V_2) * exp(-\beta * D(Z_h, V_2)) \quad (4)$$

Точка $x_i$ считается принадлежащим кластеру $V_k$, если расстояние до него $D(x_i, V_k)$ меньше, чем расстояние до других кластеров.

Итерационная процедура пересчета потенциалов и выделения центров кластеров продолжается до тех пор, пока максимальное значение потенциала превышает некоторый порог. В простейших случаях этот порог можно считать равным половине значения максимального потенциала $P_1(V_1)$.

Рассмотрим работу алгоритма на следующем примере. Возьмем несколько пользователей и разобьем их покупки по категориям, где К1, …К7 – категории товаров, а П1-П10 - пользователи. Рассмотрим данные о приобретении товаров пользователями, представленные в таблице 1. Количество кластеров мы определим равное трем.

Пусть пользователь П1 имеет координаты (3,3,8,2,8,2,1). Высчитаем расстояние между другими пользователями и их координатами (за измерение мы берем категорию, то есть у нас 7 измерений или осей на каждую ось по категории). Расчет расстояния вычисляется по формуле [1], где $x_1$ это массив с покупками одного пользователя и $x_2$ массив другого пользователя со своими покупками и т.д.

*Таблица 1.*

**Покупки пользователей по категории**

| Категории / Пользователи | К1 | К2 | К3 | К4 | К5 | К6 | К7 |
|---|---|---|---|---|---|---|---|
| П1 | 3 | 3 | 8 | 2 | 8 | 2 | 1 |
| П2 | 6 | 1 | 4 | 2 | 2 | 0 | 0 |
| П3 | 3 | 1 | 1 | 5 | 6 | 0 | 2 |
| П4 | 4 | 1 | 5 | 0 | 0 | 0 | 0 |
| П5 | 1 | 2 | 2 | 1 | 1 | 0 | 0 |
| П6 | 1 | 1 | 1 | 2 | 2 | 0 | 0 |
| П7 | 3 | 1 | 6 | 6 | 6 | 4 | 3 |
| П8 | 2 | 0 | 1 | 4 | 3 | 0 | 1 |
| П9 | 3 | 2 | 2 | 0 | 0 | 0 | 0 |
| П10 | 3 | 1 | 0 | 3 | 0 | 3 | 1 |

Расстояние между пользователями приведено в таблице 2: То есть в контейнере[0] хранятся покупки П1, а в контейнер[1-9], хранятся покупки остальных пользователей (П2, П3, П4 и т.д.)

*Таблица 2.*

**Расстояние между пользователями**

| | |
|---|---|
| Расстояние между точками: | контейнер[0] и контейнер[1] = 8.36660026534 |
| Расстояние между точками: | контейнер[0] и контейнер[2] = 8.42614977318 |
| Расстояние между точками: | контейнер[0] и контейнер[3] = 9.32737905309 |
| Расстояние между точками: | контейнер[0] и контейнер[4] = 9.79795897113 |
| Расстояние между точками: | контейнер[0] и контейнер[5] = 9.89949493661 |
| Расстояние между точками: | контейнер[0] и контейнер[6] = 6 |
| Расстояние между точками: | контейнер[0] и контейнер[7] = 9.59166304663 |
| Расстояние между точками: | контейнер[0] и контейнер[8] = 10.4880884817 |
| Расстояние между точками: | контейнер[0] и контейнер[9] = 11.5758369028 |

Следующей формулой найдем потенциалы центров кластеров. Так, как о центрах ничего неизвестно, то в качестве потенциальных центров удобно

принять сами объекты. Потенциал центров кластеров рассчитывается по следующей формуле (2). Центром первого кластера $V_1$ назначим точку с наибольшим потенциалом. Первым центром оказался П6, которому принадлежит значение «5.23290067537». Потому что чем плотнее расположены объекты в окрестности потенциального центра кластера, тем выше значение его потенциала.

*Таблица 3.*

**Потенциалы центров кластеров**

| Пользователи | Потенциалы |
|---|---|
| П1 | 3.31798555946 |
| П2 | 4.68569826173 |
| П3 | 4.27460598932 |
| П4 | 4.58737072086 |
| П5 | 5.17643020288 |
| **П6** | **5.23290067537** |
| П7 | 3.55266255988 |
| П8 | 5.02615007849 |
| П9 | 4.99427959778 |
| П10 | 4.50584152312 |

После этого итерационно выбираются центры кластеров среди точек с максимальными потенциалами. Чтобы выбрать следующий центр кластера необходимо вначале исключить влияние только что найденного кластера по формуле (3).

Теперь нужно распределить пользователей по кластерам, чтобы это сделать, нужно определить наименьшее расстояние между пользователем и центром кластера. Для этого воспользуемся формулой (1). Берется центр первого кластера и потенциал первого пользователя и с помощью формулы (1) рассчитывается расстояние между ними. После берем потенциал этого же пользователя и находим расстояние с другими центрами кластеров. Найденные

расстояния сравниваются между собой и выбирается минимальный из них. Это и будет означать, что данный пользователь был определен в тот кластер, расстояние с которым был наименьший. Таким образом, мы найдем расстояния между каждым пользователем и каждым центром кластеров, определим, к какому кластеру относится пользователь, и распределим их.

В итоге при распределении пользователей по кластерам, получили такую таблицу:

*Таблица 4.*

**Распределение пользователей по кластерам**

| Кластер 0 | Кластер 1 | Кластер 2 |
|---|---|---|
| Центр – (1,1,1,2,2,0,0) (П6) | Центр – (6,1,4,2,2,0,0) (П2) | Центр – (3,1,6,6,6,4,3) (П7) |
| Элементы | | |
| (3,1,1,5,6,0,2) (П3) | (6,1,4,2,2,0,0) (П2) | (3,3,8,2,8,2,1) (П1) |
| (1,2,2,1,1,0,0) (П5) | (4,1,5,0,0,0,0) (П4) | (3,1,6,6,6,4,3) (П7) |
| (1,1,1,2,2,0,0) (П6) | | |
| (2,0,1,4,3,0,1) (П8) | | |
| (3,2,2,0,0,0,0) (П9) | | |
| (3,1,0,3,0,3,1) (П10) | | |

На завершающем этапе распределим предложения товаров. Сначала в предложение попадут товары из той категории, из которой пользователь не покупал, но покупали внутри кластера, в который его определил алгоритм. Если на предыдущем шаге достаточное количество категорий не набралось для

формирования предложения, то в силу вступает следующий метод. Выбираем ту категорию, в которой разница между купленными товарами пользователя и в среднем купленными товарами внутри кластера этой категории максимальна. Например, пользователь купил 2 мыши и 3 клавиатуры, а внутри кластера, в который был определен пользователь, мышей было куплено 5 и клавиатур 5. Высчитываем разницу между мышами: 5-2 = 3. Теперь высчитаем разницу между клавиатурами: 5-3 = 2. Берем максимум из разниц, а значит, будет предложена мышь.

В результате работы были:

выявлена основная проблематика предметной области;

изучены свойства и особенности существующих алгоритмов кластеризации;

разработан локальный сайт, позволяющий имитировать деятельность интернет-магазина, использующий в советующей системе рассмотренный выше алгоритм.

Данный алгоритм можно дополнить другими различными рекомендательными системами, например, Content-based (основанные на контенте), с которой можно ознакомиться [1]. Эта рекомендация базируется на собранных данных о каждом конкретном товаре. Начиная от простых: «Книги того же жанра или автора», «Вещи того же производителя» и т.д.

## СПИСОК ЛИТЕРАТУРЫ


1. M. Jones Introduction to approaches and algorithms [Электронный ресурс]. Режим доступа: https://www.ibm.com/developerworks/opensource/library/os-recommender1/index.html?S_TACT=105AGX99&S_CMP=CP (дата обращения: 01.02.2017)
2. Yager R., Filev D. Essentials of Fuzzy Modeling and Control. USA: John Wiley & Sons. - 1984. - 387p. (дата обращения: 03.02.2017)