УДК 004.021

ИСПОЛЬЗОВАНИЕ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ ЭЛЕКТРОННОЙ КОММЕРЦИИ

***Д.Н. АЛЕКСЕЕВ*[1], *А.И. БАЛЕНКО*[2*]**

[1] *магистрант кафедры вычислительной техники и программирования, НТУ «ХПИ», Харьков, УКРАИНА*

[2] *доцент кафедры вычислительной техники и программирования, канд. техн. наук, НТУ «ХПИ», Харьков, УКРАИНА*

[*] *email: alexibalenko@gmail.com*

На данный момент развитие систем электронной коммерции, оказания услуг и распространения контента достигло той точки, когда просмотр всего объема доступной информации становится практически невыполнимой задачей для рядового пользователя. В связи с этим возникает необходимость предоставить каждому отдельному пользователю рекомендации, относительно тех товаров, услуг или контента, в которых он может быть наиболее заинтересован.

Целью данной работы является построение рекомендательной системы-модуля интернет-сервиса, позволяющего подобрать наиболее подходящие, с точки зрения применяемых методик, для пользователя предложения.

В качестве инструмента подбора выбран метод коллаборативной фильтрации основанный на соседстве. Считается, что система знает каждого пользователя и имеет возможность связать его текущий сеанс со всеми его предыдущими сеансами, что позволяет получать данные о его интересах и оценках.

Выбранный алгоритм предполагает сбор данных о продуктах, которые были просмотрены, приобретены или оценены пользователем для дальнейшего построения вектора предпочтений и подбора продуктов на основании сравнения этого вектора с векторами предпочтений других пользователей.

Для выполнения поиска продуктов выполняется построение матрицы оценок товаров пользователями, состоящей из векторов предпочтений каждого пользователя.

После построения матрицы выполняем поиск наиболее похожих на текущего пользователя пользователей, для этого, на текущем этапе, была выбрана косинусная мера (*sim*):

$$sim(x,y) = \frac{\sum_{i \in P_{u_x,u_y}} r_{u_x,i} r_{u_y,i}}{\sqrt{\sum_{i \in P_{u_x,u_y}} r_{u_x,i}^2} * \sqrt{\sum_{i \in P_{u_x,u_y}} r_{u_y,i}^2}},$$

где P_{u_x,u_y} – подмножество продуктов $i \in I$, которые оценили оба пользователя, $r_{u_x,i}$ и $r_{u_y,i}$ – оценки пользователей x и y для продукта i. Также к векторам пользователей u_x и u_y применяется Эвклидова нормализация, проектирующая их на единичный круг. Похожесть пользователей вычисляется при помощи скалярного произведения – косинуса угла между точками обозначенными векторами. Так как оценки пользователей положительные – результат ограничен [0, 1].

После расчёта схожести каждого $u_x, u_y \in U$ выполняется предсказание оценок для каждого продукта, который не был оценён пользователем $u \in U$. Для расчёта предсказаний предлагается использовать подход взвешенных сумм, который рассчитывается по формуле:

$$r_{u_x,i} = \overline{r_{u_x}} + \frac{\sum_{u_y \in R_{u_x,i}} (R_{u_y,i} - \overline{r_{u_x}}) sim(u_x, u_y)}{\sqrt{\sum_{u_y \in R_{u_x,i}} sim(u_x, u_y)}},$$

где $R_{u_x,i}$ – подмножество пользователей $u_y \in U$, отличных от u_x, оценивших продукт i, а $\overline{r_{u_x}}$ – средняя оценка продукта пользователем u_x. Подход взвешенных сумм принимает средние оценки соседей активного пользователя и взвешивает каждую из них в соответствии со схожестью соседа и активного пользователя.

В итоге выбирается n элементов $i \in I$ с наибольшей предсказанной оценкой $R_{u_x,i}$. Так как предсказанные оценки показывает соответствие релевантности продукта для активного пользователя, мы выбираем первые n наиболее высоко оценённых продуктов из результата расчёта взвешенных сумм.

Для реализации вышеприведенного алгоритма был выбран язык программирования Python. Для выполнения распределённых вычислений была выбрана библиотека Apache Spark. Для создания кластера и выполнения распределенных вычислений на нем выбран сервис Amazon EMR.