УДК 004.78, 004.048

Е.Е. Пятикоп, канд.техн.наук,
ГВУЗ "Приазовский государственный технический университет",
г. Мариуполь, Украина
pee_pstu@ukr.net

Исследование метода коллаборативной фильтрации на основе сходства элементов

В статье приводится классификация методов коллаборативной фильтрации, их описание. Описаны математические основы метода выдачи рекомендаций на основе сходства элементов (Item-based). Представлен подход нормализации данных с использованием базовых прогнозов. Приведены результаты экспериментов реализации метода.

Ключевые слова: коллаборативная фильтрация, пользователи, оценки, подобие элементов, нормализация данных, среднеквадратичная ошибка.

Введение

Объем информации во всемирной паутине постоянно увеличивается. Каждый день мы сталкиваемся с выбором и множеством вариантов. Какой фильм посмотреть? Какой телефон купить? Какую книгу прочесть? Размеры пространств этих решений зачастую объемные: ресурс Либрусек представляет почти 270 000 книг и каждый месяц более 5000 обновлений [1], а Amazon.com имеет более 410 000 наименований продуктов только в Kindle Store [2]. Поддержка принятия решения в информационных пространствах такого масштаба является серьезной проблемой. Поэтому, чтобы помочь пользователю найти необходимую информацию интенсивно используются рекомендательные системы. Использование таких систем позволит интернет-магазинам ускорить прибыль, любителям музыки открыть новых, неизвестных им ранее артистов, и т.д. Рекомендательные системы полезны не только для информационных ресурсов и порталов электронной коммерции, но и могут также открыть новые возможности в области безопасности, автомобильной промышленности и др. [3-4]. На сегодняшний день одним из подходов разработки рекомендательных систем является использование методов коллаборативной фильтрации (КФ). Коллаборативная фильтрация – класс методов построения рекомендаций (прогнозов) на основе известных предпочтений (оценок) группы пользователей.

Основная идея алгоритмов коллаборативной фильтрации заключается в предложении новых элементов для конкретного пользователя на основе предыдущих предпочтениях пользователя или мнения других единомышленников пользователя. На сегодняшний день исследователи разработали целый ряд алгоритмов КФ [5-8], которые можно разделить на две основные категории:

1. Методы, основанные на анализе имеющихся оценок, – *анамнестические* [1] *методы (Memory-based).* Эти алгоритмы основываются на статистических методах, чтобы найти группу пользователей близких к целевому пользователю. Этот подход еще называют метод ближайших соседей: использование предшествующих оценок, сделанных клиентом, и анализ оценок других пользователей, которые имеют подобные предпочтения. Тогда рекомендации (прогноз) для целевого пользователя формируются на основании вычисления некой меры похожести по всем накопленным данным.

2. Методы, основанные на анализе модели данных, – *модельные методы (Model-based).* В этом случае сначала по совокупности оценок формируется описательная модель предпочтений пользователей, товаров и взаимосвязи между ними, а затем формируются рекомендации на основании полученной модели. Процесс формирования рекомендаций разбит на два этапа: ресурсоемкое обучение модели в отложенном режиме и достаточно простое вычисление рекомендаций на основе существующей модели в реальном времени. Эти алгоритмы могут быть основаны на вероятностном подходе [7-8], кластерном анализе [10], анализе скрытых факторов [11].

3. Методы, основанные на объединении предыдущих алгоритмов, – *гибридные методы*.

Эти подходы в свою очередь могут быть разбиты далее на группы методов, как показано на рисунке 1.

Так, методы на основе соседства (близости) разделяются на анализ:

- сходства пользователей (User-based);
- сходства элементов (Item-based).

[1] АНАМНЕСТИЧЕСКИЙ, АНАМНЕЗ [нэ], -а, м. (спец.). Совокупность медицинских сведений, получаемых путем опроса обследуемого и знающих его лиц.

Целью обоих направлений является выделение схожих объектов в группы на основе матрицы оценок [5-7]. В первом случае определяется сходство пользователей: найти других пользователей, чьи прошлые оценки поведения похожи на те, что и у текущего пользователя, и использовать их оценки других элементов для прогнозирования предпочтения текущего пользователя. Второй подход, на основе сходства элементов, впервые предложен в [12-13], и эта версия используется в Amazon.com в настоящее время [14]. В этом случае вместо того чтобы использовать подобие между поведением пользовательских оценок для прогнозирования предпочтения, используется сходство между оценками моделей элементов. Если два элемента, как правило, имеют одинаковые оценки пользователей, то они похожи, и пользователи должны иметь аналогичные предпочтения для подобных элементов.

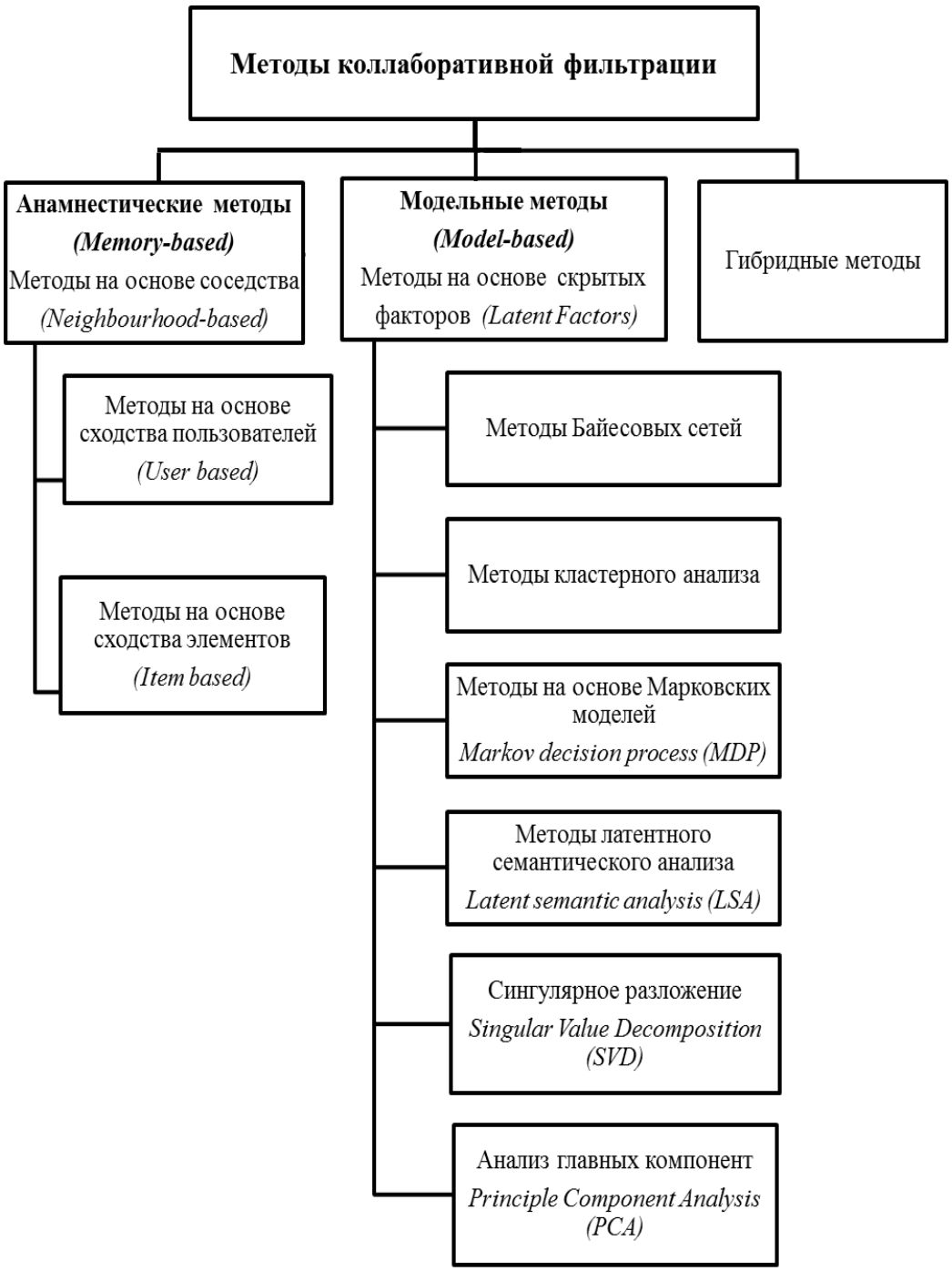

Рисунок 1 – Классификация методов коллаборативной фильтрации.

Для определения сходства между пользователями или элементами можно использовать такие подходы:

- расстояние Эвклида, Хемминга;
- корреляция Пирсона;
- ранговая корреляция Спирмена;
- коэффициент Жаккара;
- косинусное подобие.

Коллаборативная фильтрация на основе сходства пользователей (User-based) имеет высокую точность. Однако, недостатком является ресурсоемкость (требование к памяти) и сложность (количество вычислений, требуемое для получения рекомендаций). К тому же вычисление степени близости может производиться только в реальном времени, так как данные о текущей транзакции становятся доступными только в момент выработки рекомендаций. Поэтому данный метод может применяться только к относительно небольшим базам данных.

В алгоритме на основе сходства элементов (Item-based) степень близости анализируемого элемента ко всем остальным может быть вычислена в отложенном режиме по расписанию, так как вектора рейтингов всех элементов доступны до момента формирования рекомендации. Таким образом этот алгоритм оказывается более эффективным с точки зрения времени формирования рекомендаций благодаря возможности проведения отложенной предобработки данных.

Для описанных выше методов есть необходимость в хранении всей матрицы данных, т.е. предпочтений пользователей об элементах. В связи с этим возникают трудности при прогнозе предпочтений для новых пользователей или при появлении новых элементов, т.к. для них еще нет оценок. Также ограничивается возможность методов при обработке больших объемов данных. Во многих случаях хранение всей матрицы предпочтений избыточно: как правило, пользователи и элементы делятся на группы с аналогичными профилями предпочтений. Например, многие научно-фантастические фильмы будут нравиться в аналогичной степени тем же наборам пользователей. Поэтому возникает задача в понижении размерности матрицы оценок. Такие задачи решают методы второй группы (рис. 1).

В этом случае возможен вариант объединения пользователей (элементов) в кластеры (профили) с помощью некоторого индекса сходства. Элементы и оценки, данные пользователями из одного кластера, используются для вычисления рекомендаций. Кластерные модели лучше масштабируются, т.к. сверяют профиль пользователя с относительно небольшим количеством сегментов, а не с целой пользовательской базой. Сложный и емкий кластерный подсчет ведется с в оффлайн режиме. Эта задача может выполнена на основе разных математических подходов [10, 11, 17].

В статье рассматривается использование метода на основе сходства элементов с нормализацией данных.

Постановка задачи

Информационная область для систем КФ состоит из пользователей, которые выразили предпочтения для различных предметов. Предпочтение (оценка) часто представляется в виде триплета (пользователь, предмет, оценка). Эти оценки могут принимать различные формы, в зависимости от рассматриваемой системы. Некоторые системы используют вещественную или целочисленную оценочную шкалу, такую как 0-5 звезд, другие используют бинарные или тройные меры. Множество всех триплетов оценок формирует разреженную матрицу, называемую матрицей оценок. Пары (Пользователь, предмет), в которых пользователи не отдали предпочтение предмету, являются неизвестными значениями этой матрицы (Табл. 1).

Таблица 1 – Пример матрицы оценок

	Элемент 1	**Элемент 2**	**Элемент 3**
Пользователь 1	3	?	2
Пользователь 2	?	4	3
Пользователь 3	5	4	?

При использовании системы КФ необходимо решить две задачи: 1) спрогнозировать оценку или предпочтение, которое пользователь отдаст предмету. Целью прогноза является заполнение в матрице оценок недостающих значений; 2) выдача рекомендации, т.е. формирование ранжированного списка N элементов для данного пользователя.

Определим математические обозначения для привязки различных элементов модели рекомендательных систем. Генеральная совокупность состоит из набора пользователей U и набора элементов I.

I_u - множество элементов, оцененных пользователем u.

U_i - множество пользователей, которые оценили элемент i.

$r_{u,i}$ - оценка пользователя u для элемента i.

r_u - вектор всех оценок пользователя u.

r_i - вектор всех оценок элемента i.

$\bar{r}_u$ и $\bar{r}_i$ - средние значения оценок пользователя u и элемента i соответственно.

Рекомендательный прогноз обозначим как $\hat{r}_{u,i}$.

Метод на основе сходства элементов

Шаг 1: для каждого элемента j вычисляется мера близости к элементу i. Для этого можно использовать один из указанных выше подходов, например, коэффициент Пирсона:

$$s_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j))}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}} \quad (1)$$

где $U \in U_i \cup U_j$ – множество пользователей, которые оценили элементы *i* и *j*.

Шаг 2: выбираем множество элементов *S*, наиболее близких к объекту *i*. В работе [12] Савар определил, что достаточные результаты получаются при *k=30* элементов множества *S*. Но эти данные зависимы от рассматриваемой задачи и разреженности матрицы.

Шаг 3: предсказание рейтинга (оценки) объекта на основе рейтингов близких к нему объектов:

$$\hat{r}_{u,i} = \frac{\sum_{j \in S} s_{i,j} \cdot r_{u,j}}{\sum_{j \in S} |s_{i,j}|} \quad (2)$$

Данный алгоритм отражает теоретическую базу метода, но на практике ряд факторов требует переосмысления расчетов.

Как правило, подавляющее большинство оценок неизвестно, и разреженность матрицы оценок достаточно высока. С другой стороны, данные, которое уже имеются в матрице достаточно субъективны. Некоторые пользователи – оптимисты, и их оценки всегда высоки (среднее 4 из 5), другие пользователи – циники, их оценки всегда занижены (среднее 2,5 из 5). Кроме этого, всегда есть элементы, которые нравятся всем.

В целях борьбы с подгонкой разреженных данных с оценками, проводится регуляризация моделей таким образом, чтобы сократить вероятность появления случайных связей между оценками, которые не отражают действительность. Регуляризация контролируется константами, которые обозначаются как λ_1, λ_2, ... Точные значения этих констант определяются перекрестной проверкой. По мере их роста, регуляризация становится все тяжелее.

Для того чтобы оптимизировать производительность выдачи рекомендаций, важно нормализовать оценки до вычисления матрицы подобия. Это может быть достигнуто путем вычисления базового прогноза, в котором инкапсулируют отклонение пользователя и элемента. Пары пользователь-элемент (u,i) для которых оценки $r_{u,i}$ известны составляют множество *K*. Базовый прогноз для неизвестной оценки $r_{u,i}$ обозначается $b_{u,i}$ и определяется формулой:

$$b_{u,i} = \mu + b_u + b_i \quad (3)$$

где μ - общая средняя оценка; b_u и b_i - параметры, которые показывают наблюдаемое отклонение пользователя *u* и элемента *i* соответственно от среднего значения.

Так как все параметры (3) взаимосвязаны, то рассчитывать их необходимо вместе, решив задачу наименьших квадратов [15-16].

$$min \sum_{(u,i) \in K} (r_{u,i} - \mu - b_u - b_i)^2 + \lambda_1 (\sum_u b_u^2 + \sum_i b_i^2) \quad (4)$$

Здесь первая часть $\sum_{(u,i) \in K} (r_{u,i} - \mu - b_u - b_i)^2$ стремится найти b_u и b_i, которые соответствуют данным оценок. Часть регуляризации $\lambda_1 (\sum_u b_u^2 + \sum_i b_i^2)$ позволяет избежать подгонки данных, штрафом за величину параметров.

Для метода на основе сходства элементов этот подход отразится так. Расчет меры близости основан только на оценках пользователей, которые оценили оба элемента:

$$s_{i,j} = \frac{n}{n + \lambda_2} \cdot p_{i,j} \quad (5)$$

где *n* – количество пользователей, которые оценили оба элемента *i* и *j*; λ_2 - константа регуляризации; $p_{i,j}$ – коэффициент корреляции Пирсона по формуле (1).

Прогнозируемое значение $\hat{r}_{u,i}$ получим как средневзвешенную оценку соседних элементов, в то время как корректировки для пользователей и элементов проводятся через базовые прогнозы:

$$\hat{r}_{u,i} = b_{u,i} + \frac{\sum_{j \in S} s_{i,j} \cdot (r_{u,j} - b_{u,j})}{\sum_{j \in S} |s_{i,j}|} \quad (6)$$

$$min \sum_{(u,i) \in K} \left(r_{u,i} - \mu - b_u - b_i - \frac{\sum_{j \in S} s_{i,j} \cdot (r_{u,j} - b_{u,j})}{\sum_{j \in S} |s_{i,j}|} \right)^2 + \quad (7)$$
$$+ \lambda_3 (\sum_u b_u^2 + \sum_i b_i^2 + \sum_{j \in K} s_{i,j}^2)$$

$$b_u \leftarrow b_u + \gamma_1 \cdot (e_{i,j} - \lambda_1 \cdot b_u) \quad (8)$$

$$b_i \leftarrow b_i + \gamma_1 \cdot (e_{i,j} - \lambda_1 \cdot b_i) \quad (9)$$

$$e_{i,j} = r_{i,j} - \hat{r}_{i,j} \quad (10)$$

где γ_1 – константа регуляризации.

На основе представленных математических выкладок были проведены исследования по использованию описанного метода. для оценки эффективности использовалась среднеквадратичное отклонение.

$$RMSE = \sqrt{\frac{1}{n} \sum_{u,i} (\hat{r}_{u,i} - r_{u,i})^2} \quad (11)$$

где $r_{u,i}$ – известная оценка пользователя u для элемента i, $\hat{r}_{u,i}$ – спрогнозированная оценка.

Результаты исследований

В качестве исходных данных использовались таблицы базы данных с оценками пользователей объемом – 20000 строк, с данными о пользователях – 200 строк и данными о книгах – 1500 строк.

Метод на основе сходства элементов имеет параметры обучения, такие как размер соседства K, коэффициент скорости обучения базового отклонения оценок γ_1, коэффициент регуляризации базового отклонения оценок λ_1. Последние два параметра исследованы [15, 16] и приняты: коэффициент скорости обучения базового отклонения оценок γ_1=0,001. коэффициент регуляризации базового отклонения оценок λ_1 =0,005.

Приняв за основу эти данные, определи оптимальное количество факторов, при точности ε=0,00001.

Таблица 2 – Результаты эксперимента

№ п/п	Размер соседства	Погрешность при обучении (RMSE)	Погрешность при тестировании (RMSE)
1	25	0.87515	0.94378
2	50	0.87228	0.93456
3	75	0.86981	0.93103
4	100	0.86766	0.92748
5	125	0.86574	0.92657
6	150	0.86269	0.92486
7	175	0.86027	0.92361
8	200	0.85833	0.92276

Значения RMSE, полученные при тестировании метода адекватны для заданной задачи, т.к. все исследователи стремятся, в соответствии с запросами Netfix, уменьшить RMSE с 0.9514 до 0.8563.

На рисунке 2 показано графическое отображение данных таблицы 2. Согласно полученным результатам можно сделать вывод о том, что с увеличением размера соседства, погрешность прогнозов уменьшается. Для исходных данных оптимальным размером соседства является k =200.

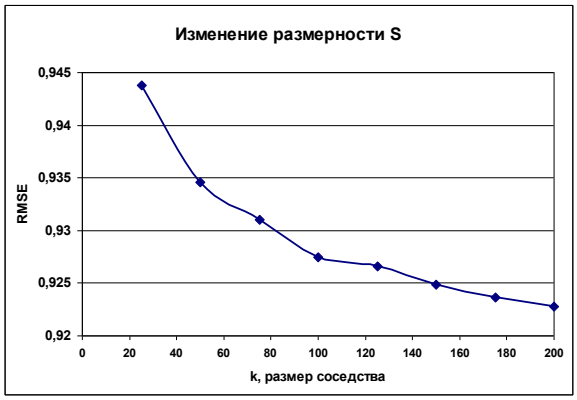

Рисунок 2 – Результаты эксперимента

Немаловажным фактором является время обучения модели. Проведен сравнительный анализ времени обучения модели в зависимости от объема данных (Таблица 3).

Таблица 3– Результаты эксперимента

№ п/п	Объем обучающих данных в строках	Время обучения модели Item-based (чч:мм:сс)
1	2500	0:23:33
2	5000	0:45:12
3	7500	1:03:25
4	10000	1:32:16
5	12500	1:54:47
6	15000	2:19:14
7	17500	2:57:26
8	20000	3:29:21

Очевидно, что с ростом объема обучающих данных, время обучения модели возрастает сильнее, но этап обучения происходит в оффлайн режиме.

Дальнейшие исследования предполагается направить в использовании методов поиска скрытых факторов с сокращение размерности измерений на основе сингулярного разложения матриц.

Список литературы

1. Либрусек -Статистика http://lib.rus.ec/stat
2. Amazon.com, "Q4 2009 Financial Results," Earnings Report Q4-2009, January 2010.
3. Рекомендательные системы http://www.numberscompany.ru/products/recommenders
4. M. van Setten, S. Pokraev, and J. Koolwaaij, "Context-aware recommendations in the mobile tourist application compass," Heidelberg, 2004, vol. 3137, pp. 515–548
5. J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence, July 1998.

6. Xiaoyuan Su and Taghi M. Khoshgoftaar "A Survey of Collaborative Filtering Techniques A Survey of Collaborative Filtering Techniques" // Hindawi Publishing Corporation, Advances in Artificial Intelligence archive, USA : 2009. — С. 1-19.
7. G. Adomavicius На пути к новому поколению рекомендационных систем: обзор имеющихся систем и возможные инновации. IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, Июнь 2005 Электронный ресурс: http://artpragmatica.ru/rs/in/pic/58-870-20061024072441-Toward_the_next_generation_of_recommender_systems.doc
8. Гомзин А. Г., Коршунов А. В. Системы рекомендаций: обзор современных подходов // Труды ИСП РАН. 2012. №. Электронный ресурс: http://cyberleninka.ru/article/n/sistemy-rekomendatsiy-obzor-sovremennyh-podhodov.
9. Mustansar Ali Ghazanfar "Building Switching Hybrid Recommender System Using Machine Learning Classifiers and Collaborative Filtering" // International Journal of Computer Science, Электронный ресурс: http://www.iaeng.org/IJCS/issues_v37/issue_3/IJCS_37_3_09.pdf
10. Савчук Т.О., Застосування кластерного аналізу для колаборативної фільтрації / Т.О. Савчук, А.В.Сакалюк // Вісник Хмельницького національного університету. –2011 – №1– С. 186-192
11. Лексин В.А., Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // Математические методы распознавания образов-13. – М. МАКС Пресс, 2007. – С. 488-491
12. Sarwar B. M. Item-based collaborative filtering recommendation algorithms / B. M. Sarwar, G. Karypis, J. A. Konstan // Proceedings of ACM WWW '01, pp. 285–295, ACM, 2001.
13. Karypis G. Evaluation of item-based top-N recommendation algorithms / G. Karypis // Proceedings of ACM CIKM '01, pp. 247–254, ACM, 2001.
14. Linden G. Amazon.com recommendations: Item-to-item collaborative filltering / G. Linden, B. Smith, J. York // IEEE Internet Computing, vol. 7, no. 1, pp. 76–80, 2003.
15. Hu Y., Koren Y., Volinsky C.: Collaborative filtering for implicit feedback datasets. In ICDM- 08, 8th IEEE Int. Conf. on Data Mining, pages 263–272, Pisa, Italy, 2008.
16. Koren Y., Ave P., Park F.: Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model. In: KDD '08 Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (2008)
17. Kurucz M. Methods for large scale SVD with missing values / M. Kurucz, A. A. Benczur, K. Csalogany // Proceedings of KDD Cup and Workshop 2007, 2007.

О.Є. П'ЯТИКОП
ДВНЗ "Приазовський державний технічний університет"

ДОСЛІДЖЕННЯ МЕТОДУ КОЛЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ НА ОСНОВІ БЛИЗЬКОСТІ ЕЛЕМЕНТІВ

У статті наводиться класифікація методів коллаборативної фільтрації, їх опис. Описано математичні основи методу видачі рекомендацій на основі подібності елементів (*Item-based*). Представлено підхід нормалізації даних з використанням базових прогнозів. Наведені результати експериментів реалізації методу.

Ключові слова: коллаборативна фільтрація, користувачі, оцінки, близькість елементів, нормалізація даних, середньоквадратична помилка.

E.E. PYATIKOP
Pryazovskyi State Technical University

THE RESEARCH OF COLLABORATIVE FILTERING METHOD BASED ON NEIGHBORHOOD ELEMENTS

The article is report a classification of methods collaborative filtering, their description. The method Item-Based Collaborative Filtering is described by mathematical formulas. Normalization of the data is shown by the baseline estimates. The paper presents experiments of the method.

Keywords: collaborative filtering, users, rating neighbor Item based, normalizing data, RMSE