УДК 004.89:004.912 DOI: 10.17587/it.26.523-528

А. С. Круглик, магистрант, e-mail: kruglik.a.s@mail.ru,
И. А. Лакман, канд. техн. наук, доц., e-mail: lackmania@mail.ru,
Уфимский государственный авиационный технический университет

Гибридный подход усиленной контентом коллаборативной фильтрации в области рекомендательных систем

Проводится эмпирическое исследование прогнозных свойств гибридной рекомендательной системы, основанной на подходе усиленной контентом коллаборативной фильтрации. Проводится сравнение по четырем различным метрикам качества с подходами рекомендательных систем: случайный прогноз, контентная фильтрация, коллаборативная фильтрация, усреднение прогнозов. Апробация подходов проводилась на данных о фильмах и рейтингах, поставленных пользователями. При применении метода усиленной контентом коллаборативной фильтрации результаты улучшаются на 15...20 % по сравнению с остальными подходами.

Ключевые слова: *рекомендательная система, гибридный подход, контентная фильтрация, коллаборативная фильтрация*

Введение

Быстрое развитие технологий, расширяющее возможности людей и обеспечивающее широкий доступ к почти любой информации, порождает проблемы, связанные с информационным поиском.

Несмотря на наличие мощных поисковых систем быстрый поиск нужной информации, товаров или услуг не всегда является осуществимым. Рекомендательные системы могут решить проблему информационной перегрузки путем фильтрации действительно важных фрагментов информации из огромного количества динамически генерируемых данных и, как следствие, повысить оперативность осуществления поиска.

Рекомендательные системы помогают уменьшить транзакционные издержки поиска информации и выбора товара в интернете, улучшить процесс принятия решений на основе анализа данных. Таким образом, необходимость в эффективных и точных рекомендательных системах, которые обеспечат пользователю надежные, релевантные, персонализированные рекомендации, не может быть переоценена.

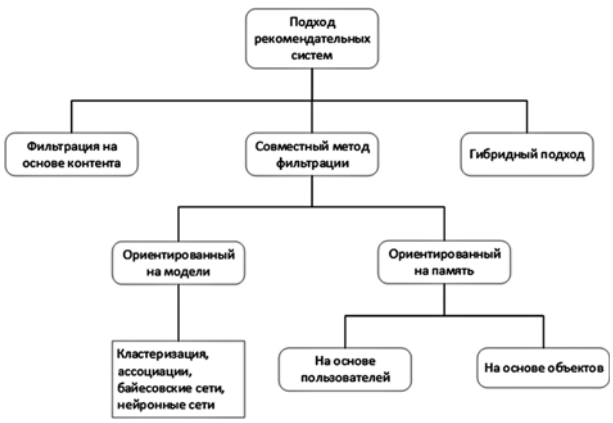

Рис. 1. Классификация подходов рекомендательных систем

Современные рекомендательные системы, как правило, классифицируются в соответствии с их подходом к рейтинговой оценке. Общепринятая формулировка задачи рекомендации впервые была высказана в 1994—1995 гг., и с тех пор эта проблема была тщательно изучена [1]. Условно рекомендательные системы подразделяются на категории (рис. 1) в соответствии с типом генерируемых рекомендаций [2, 3]:

- контентно-ориентированные рекомендации (пользователю рекомендованы объекты, схожие с теми, которые пользователь предпочел ранее);
- рекомендации на основе совместной фильтрации (пользователю рекомендованы объекты, понравившиеся другим пользователям с похожими вкусами и предпочтениями);
- гибридные подходы.

Также существуют следующие подходы к генерированию рекомендаций, основанные на определенных признаках [3, 4]:

- рекомендации на основе сообществ (содружеств) (пользователю рекомендованы объекты, предпочитаемые его друзьями);
- рекомендации на основе социально-демографической информации (пользователю рекомендованы объекты исходя из его социально-демографических признаков);
- рекомендации на основе знаний (пользователю рекомендованы объекты, которые подходят ему лучше с точки зрения знаний предметной области).

Для преодоления недостатков отдельных подходов используют различные гибридные методы, основанные на совмещении"чистых" (негибридных) техник предсказания.

Целью исследования является улучшение качества прогнозирования предпочтений пользователей разработанной рекомендательной системы за счет применения гибридных подходов.

Существует множество исследований, посвященных применению гибридизации подходов к реализации рекомендательных систем. Применяется уменьшение размерности на группе контентно-ориентированных профилей (латентное семантическое индексирование). Используются фильтр-боты как агенты, анализирующие контент и в то же время являющиеся дополнительными участниками во множестве пользователей совместной фильтрации. Подход "совместной фильтрации через контент" основывается на традиционной совместной фильтрации при поддержке профиля пользователя с информацией о контенте [1]. Некоторые исследователи предложили комбинировать исходные данные, например, аспектная модель Хофманна была расширена добавлением агрегированной информации, состоящей из трех компонентов: пользователи, объекты, контент объектов [2].

Методология исследования

Большинство современных рекомендательных систем используют гибридный подход, сочетающий совместную фильтрацию, контентную фильтрацию и другие подходы. Нет никаких причин, по которым несколько различных методов не могли бы быть гибридизированы. Гибридные подходы могут быть реализованы несколькими способами [1, 5]:

- путем составления прогнозов рекомендаций на основе контента и на основе совместной фильтрации по отдельности с последующим их объединением;
- путем добавления основанных на контенте возможностей к подходу, основанному на совместной фильтрации (и наоборот);
- путем объединения подходов в одну модель.

Существуют исследования, в которых эмпирически сравнивают качество гибридных методов с подходом на основе совместной фильтрации и с методикой фильтрации контента [1, 5, 6]. Они показали, что гибридные методы могут дать более точные рекомендации, чем применение "чистых" подходов. Эти методы могут также использоваться для преодоления некоторых распространенных проблем в рекомендательных системах, таких как "холодный запуск" и "разреженность".

Исследования гибридных систем определили семь вариантов подходов [3]:

- *взвешенный* — численное объединение оценок различных компонентов рекомендаций;
- *переключение* — выбор между компонентами рекомендаций и применение выбранного метода;
- *смешанный* — рекомендации от разных рекомендательных подсистем объединены вместе, для формирования единого списка рекомендаций;
- *комбинация свойств* — характеристики, полученные из разных источников знаний, объединены и учтены в едином алгоритме рекомендаций;
- *расширение характеристик* — формирование новых свойств, являющихся входными данными в следующую подсистему;
- *каскадный подход* — каждая рекомендательная подсистема получает свой приоритет, при этом более низкий приоритет уступает подсистемам с высоким приоритетом;
- *мета-алгоритм* — применяется методика рекомендации в виде модели, используемой в последующей методике.

В настоящем исследовании предлагается методика построения рекомендательной системы, в основе которой лежит гибридный подход усиленной контентом совместной фильтрации (Content-Boosted Collaborative Filtering,CBCF), дающий более хороший результат, чем подход на основе совместной фильтрации (Collaborative Filtering, CF) и системы на основе контента (Content-based, CB) [7, 8].

В усиленной контентом совместной фильтрации сначала создается псевдовектор рейтингов v_u для каждого пользователя $u \in U$. Этот вектор v_u состоит из оценок объектов, предоставленных пользователем u, а вместо отсутствующего рейтинга какого-либо объекта стоит прогноз рейтинга от алгоритма рекомендательной системы на основе контента [7]:

$$v_{ui} = \\ = \begin{cases} r_{ui}, \text{если пользователь } u \text{ оценил объект } i; \\ c_{ui}, \text{иначе.} \end{cases}$$

В приведенном выше уравнении r_{ui} обозначает фактическую оценку, предоставленную пользователем u для объекта i, в то время как c_{ui} является прогнозом рейтинга от алгоритма на основе контента.

Вместе взятые псевдовекторы пользовательских оценок всех пользователей образуют плотную псевдоматрицу V. Далее алгоритм со-

вместной фильтрации использует эту плотную матрицу в качестве входных данных. Сходство между активным пользователем *a* и другим пользователем *u* вычисляется с использованием коэффициента корреляции Пирсона, косинусного подобия или другой меры. Вместо первоначальных оценок от пользователей используются рейтинги, предоставляющие собой псевдовекторы рейтингов v_a и v_u [7].

Для тестирования предлагаемой методики построения рекомендательной системы был выбран набор данных The Movies Dataset. Он представляет собой ансамбль данных, собранных из баз данных TMDb и Movie Lens лабораторией Group Lens. Информация о фильмах, актерах и ключевых словах были получены из открытого API TMDb (открытый API позволяет получить данные без авторизации и регистрации). Несмотря на использование API TMDb набор данных не одобрен или не сертифицирован TMDb. Их API также обеспечивает доступ к данным о многих дополнительных фильмах, актерах и актрисах, членах съемочной группе и телешоу [9].

Проводимую методику исследования можно уложить в следующую схему. На первом этапе разрабатывалась рекомендательная система на основе контентной фильтрации Content-based с последующим прогнозом рекомендации. Для этого предварительно проводилась предобработка данных: были удалены объекты с некорректной или с отсутствующей информацией, дубликаты объектов; проведен стемминг слов (слово заменяется на его основу); были удалены стоп-слова и редко встречающиеся слова для описания контента объектов. В последующем признаки были разделены на две группы на основе природы объектов (фильмы): описание контента (описание фильма и его слоган) и описание признаков (ключевые слова сюжета, актеры, режиссер и жанры). Из всех актеров в фильме были отобраны первые три по важности. Затем данные преобразовывались в вид, пригодный для обработки рекомендательной системой. Для описания контента была сформирована матрица токен—объект, основанная на *TF—IDF* мере(здесь *TF* — мера, равная отношению числа вхождений некоторого слова к общему числу слов документа, а *IDF* — мера, равная отношению общего числа документов к числу документов, в которых встречается это слово). Выбор в пользу данного метода был основан на возможности сузить признаковое пространство, что особенно актуально при обработке текстовых данных в виде разряженных матриц. Для описания признаков была сформирована матрица токен—объект, основанная на подсчете токенов *TF*. По каждой из преобразованных матриц была сформирована матрица схожести с помощью косинусной меры. В конце выполнения данного этапа проводился прогноз рейтингов. Были получены спрогнозированные рейтинги для всех неоцененных объектов у всех пользователей по двум матрицам токен—объект, с последующим их усреднением по двум группам признаков.

На втором этапе разрабатывалась рекомендательная система на основе совместной фильтрации Collaborative Filtering с последующим прогнозом. Сначала была рассчитана матрица схожести пользователей. Для этого на основе матрицы рейтингов с помощью косинусной меры формировалась квадратная матрица схожести всех пользователей. В качестве меры сходства была выбрана косинусная мера, так как использование в качестве меры коэффициента корреляции Пирсона ухудшало метрики качества рекомендательной системы. Далее были спрогнозированы рейтинги, и для каждого пользователя через матрицу схожести находили 20 наиболее схожих пользователей — единомышленников. Значения рейтингов были рассчитаны с использованием рейтингов единомышленников и их нормализованных степеней сходства.

На третьем этапе разрабатывалась наивная гибридная рекомендательная система Naive Hybrid, которая усредняла прогнозы от рекомендательных систем на основе контента и на основе совместной фильтрации.

На четвертом этапе разрабатывалась рекомендательная система Random, прогнозирующая случайные значения рейтингов согласно закону равномерного распределения.

На пятом этапе разрабатывалась рекомендательная система на основе усиленной контентом совместной фильтрации Content-Boosted Collaborative Filtering с последующим прогнозом. Сначала подсистема контентной фильтрации прогнозировала рейтинги. Далее исходные и спрогнозированные рейтинги были объединены в плотную (полностью заполненную) псевдоматрицу. Это один из ключевых моментов данного гибридного подхода. Затем подсистема совместной фильтрации рассчитывала матрицу схожести пользователей, но не на основе исходной матрицы рейтингов, а на основе псевдоматрицы. В этом и состоит основная идея данного гибридного подхода. В результате подсистема совместной фильтрации прогнозировала рейтинги с учетом полученной ранее матрицы схожести.

На шестом этапе для пяти разработанных рекомендательных систем (Random, Content-based, Collaborative Filtering, Naive Hybrid, Content-Boosted Collaborative Filtering) рассчитывали четыре метрики, позволяющие оценить качество полученных прогнозов и провести

сравнение алгоритмов. В отличие от многих исследований [1, 2], в которых используются или описываются от одной до трех метрик, сравнение было проведено по четырем метрикам. Каждая метрика оценивает определенное качество рекомендательной системы: ошибка прогноза рейтинга объекта, разделимость множества объектов на подходящие пользователю и неподходящие, ранжирование объектов по убыванию релевантности. Такой подход позволяет дать обоснование мощности предсказывания выбранной рекомендательной системы.

Метрики качества предсказания формировались на основе множества P пар (пользователь, объект), для которых алгоритм предсказал рейтинг p_{ui} объекту u от i-го пользователя, т.е. $P = \{(u,i) | p_{ui} \neq \varnothing,\ u \in U,\ i \in I\}$; r_{ui} — реальный рейтинг. При сравнении алгоритмов использовались следующие метрики качества:

- средняя абсолютная ошибка (Mean Absolute Error, *MAE*)

$$MAE = \frac{1}{|P|} \sum_{(u,i) \in P} |p_{ui} - r_{ui}|;$$

- средняя квадратичная ошибка (Root Mean Square Error, *RMSE*)

$$RMSE = \sqrt{\frac{1}{|P|} \sum_{(u,i) \in P} (p_{ui} - r_{ui})^2};$$

- площадь под ROC-кривой (Area Under the ROC Curve, *AUC ROC*). *AUC* может находится в интервале от 0 до 1: при идеальном алгоритме значение будет равно 1 или близкое к 1, а при случайном поведении рекомендаций (так называемые бесполезные, рандомные рекомендации) — 0,5 [5, 7];
- метрика оценивания ранжирования *MAP@k* (*Mean Average Precision at k*)

$$MAP@k = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{K_j} \sum_{k=1}^{K_j} Precision@k,$$

где N — число пользователей, т.е. $|U|$, или (поисковых) запросов; k — число топ-объектов. Метрика *MAP* (*среднее*) в данном исследовании будет означать усреднение значении *MAP*@1, *MAP*@2, ..., *MAP*@10. *Precision* —точность с учетом случайной природы ошибок:

$$Precision@k = \frac{TP@k}{k},$$

где *TP@k* — число релевантных объектов в топ-k рекомендаций.

Методика в виде алгоритмов была реализована на языке программирования Python.

Вычислительный эксперимент

Для вычислительного эксперимента реализации предложенной методики использовали файлы из набора данных The Movies Dataset, содержащие метаданные для всех 45 000 фильмов, перечисленных в полном наборе данных Movie Lens [9]. Набор данных состоит из фильмов, выпущенных до июля 2017 г. включительно, и содержит информацию о составе актеров, съемочной команде, ключевых словах сюжета, бюджете, доходе, киноплакате, дате выхода, языках, производственных компаниях, странах производства фильмов, подсчете голосов TMDb, средних значениях голосов и др. Также имеются файлы, содержащие 26 миллионов оценок от 270 000 пользователей для всех 45 000 фильмов. Рейтинги находятся в диапазоне от 0,5 до 5,0.

Group Lens предоставляет набор данных The Movies Dataset в виде следующих файлов [9]:

- movies_metadata.csv: основной файл метаданных фильмов. Содержит информацию о 45 000 фильмов, представленных в полном наборе данных Movie Lens. В качестве атрибутов фильмов рассматриваются киноплакат, бюджет, доход, дата выпуска, языки, страны производства, кинокомпании, жанры фильма, название, слоган, словесное описание, длительность, режиссеры и др.;
- keywords.csv: содержит ключевые слова сюжета и описания для фильмов MovieLens. Доступно в виде строкового объекта JSON;
- credits.csv: содержит информацию об актерах, режиссере и съемочной группе для всех фильмов. Доступно в виде строкового объекта JSON;
- links.csv: файл, содержащий идентификаторы TMDb и IMDB для всех фильмов, представленных в полном наборе данных MovieLens;
- ratings.csv: полный набор из 26 миллионов оценок от 270 000 пользователей для всех 45 000 фильмов;
- links_small.csv: содержит идентификаторы TMDb и IMDB для небольшого подмножества (9000 фильмов) полного набора данных;
- rating_small.csv: подмножество 100 000 оценок от 700 пользователей для 9000 фильмов. Апробация алгоритмов проводилась именно на этих оценках.

Для корректности проведения вычислительного эксперимента использовалась стандартная процедура слепой валидации: для каждого пользователя поставленные им оценки фильмам были разделены на два множества — тренировочное (обучающее) и тестовое.

Первое содержит в себе 70 %, а второе — 30 % от всех оценок пользователя.

Последовательно были выполнены шесть этапов разработанной методики исследования на обучающем множестве оценок. На седьмом этапе обученные алгоритмы всех пяти рекомендательных систем были протестированы на тестовом множестве оценок. Таблица содержит сводку всех рассчитанных метрик качества результатов тестирования на подмножестве данных.

Обсуждение результатов

В результате проведенного вычислительного эксперимента лучшие метрики качества прогнозирования были получены (рис. 2) для подхода Content-boosted Collaborative Filtering построения рекомендательной системы (метрики MAE и RMSE имели меньшие значения, а AUCROC и MAP (*среднее*) — большие).

Для предсказания самих рейтингов были использованы стандартные формулы подхода совместной фильтрации вместо предложенного авторами оригинального метода [1]: после расчета схожести пользователей псевдоматрица была заменена на исходную матрицу рейтингов. Это улучшает результаты прогноза

Сводка результатов тестирования подходов (подмножество данных)

Подход к построению рекомендательной системы	Метрика качества прогнозирования			
	MAE	RMSE	AUC ROC	MAP (среднее)
Random	1,3539	1,6709	0,5042	0,1222
Content-based (CB)	0,7337	0,9457	0,6853	0,1336
Collaborative filtering (CF)	0,7744	0,9993	0,6682	0,1477
Naive Hybrid	0,7274	0,9358	0,6945	0,1507
Content-Boosted Collaborative Filtering (CBCF)	0,5837	0,7809	0,8059	0,1968

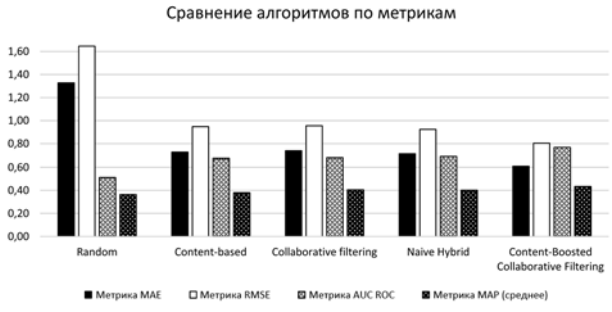

Рис. 2. Сравнение алгоритмов по метрикам (подмножество данных)

рейтингов и ограничивает влияние псевдорейтингов. Также не было применено предложенное авторами оригинального метода гармоническое взвешенное среднее для упрощения расчетов и лучшей интерпретации результатов.

Подход Content-Boosted Collaborative Filteringдля разработки рекомендательных систем хорошо интерпретируется. Обычный метод совместной фильтрации считает величину сходства двух пользователей тогда и только тогда, когда они оценили один или несколько абсолютно одинаковых объектов. В противном случае, если они оценивали разные объекты, эти пользователи будут абсолютно не схожи. Однако данное предложение не учитывает, что если сами объекты, оцененными этими пользователями, похожи, то и эти два пользователя могут быть потенциально схожи. Такой метод нахождения потенциальных схожих пользователей увеличивает число ближайших соседей [7]. При этом ранжирование единомышленников может кардинально измениться, следовательно, изменятся и предсказанные значения по сравнению с обычным методом CF.

Простой метод CF не может предсказать рейтинг от пользователя объекту, который не был оценен ни одним пользователем — это означает, что этот объект никогда не будет рекомендован. Подход усиленной контентом совместной фильтрации CBCF решает эту проблему с помощью совместного применения алгоритма CB, предварительно предсказывая для этого объекта оценки от других пользователей [1, 7].

Заключение

Проведенные эксперименты показали, что рекомендательная система, построенная на основе гибридного метода усиленной контентом совместной фильтрации работает лучше, чем системы, базирующиеся на "чистых" подходах (на основе контента, совместной фильтрации, случайном подходе), а также лучше гибридной наивной (равновзвешенной) системы. Включение информации о контенте в совместную фильтрацию может значительно улучшить качество предсказания рекомендательной системы в среднем на 15...20 % по четырем метрикам тестирования.

Подход Content-Boosted Collaborative Filtering элегантно использует контент в рамках совместной фильтрации. Он преодолевает недостатки обоих методов — совместной фильтрации и контентной фильтрации — путем укрепления контентом совместного метода фильтрации. Кроме того, благодаря модуль-

ной природе данной гибридной системы любые улучшения в совместной фильтрации или контентно-ориентированной подсистеме могут быть легко использованы для создания более мощной рекомендаций системы.

Недостатком реализованной рекомендательной системы является увеличение времени обучения алгоритма и требуемой для этого оперативной памяти. Эти проблемы решаются применением распараллеливания алгоритма и кластеризации пользователей и объектов с последующим отбором наиболее подходящего для обучения алгоритма кластера.

Список литературы

1. **Adomavicius G., Tuzhilin A.** Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions // IEEE Transactions on Knowledge and Data Engineering. June 2005. Vol. 17, N. 6.
2. **Encyclopedia** of Machine Learning. Chapter № 00338. 2010.
3. **Burke R.** Hybrid Web Recommender Systems // Wayback Machine, The Adaptive Web. Lecture Notes in Computer Science. Lecture Notes in Computer Science.Springer-Verlag, Berlin, Germany. May 2007. Vol. 4321. P. 377—408,
4. **Sidnooma Christian Kabore.** Design and implementation of a recommender system as a module for Liferay portal // Master's Thesis. Barcelona School of Computing, University Polytechnic of Catalunya. Master in Information Technologies. September 2012.
5. **Isinkaye F., Folajimi F., Ojokoh B.** Review Recommendation systems: Principles, methods and evaluation // Egyptian Informatics Journal. 2015. Vol. 16. P. 261—273.
6. **Ricci F., Rokach, L., Shapira, B.** Recommender System Handbook. New York: Springer, 2011.
7. **Melville P., Mooney R., Nagarajan R.** Content-Boosted Collaborative Filtering for Improved Recommendations // Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002).Edmonton, Canada, July 2002. P. 187—192.
8. **Herzog D., Worndl W.** Extending Content-Boosted Collaborative Filtering for Context-aware, Mobile Event Recommendations // WEBIST 2016 — 12th International Conference on Web Information Systems and Technologies. 2016.
9. **The Movies** Dataset. URL: https://www.kaggle.com/rounakbanik/the-movies-dataset.

A. S. Kruglik, Postgraduate, e-mail: kruglik.a.s@mail.ru,
I. A. Lackman, PhD in Technical Sciences, Associate Professor, e-mail: lackmania@mail.ru,
Ufa State Aviation Technical University

Hybrid Approach Content-Boosted Collaborative Filtering in the Field of Recommendation Systems

In this article, we conducted an empirical study of the predictive properties of hybrid recommendation system based on Content-Boosted Collaborative Filtering approach. The aim of the study is to improve the quality of forecasting the preferences of users of the developed recommendation system using of hybrid approaches. They overcome the disadvantages of individual approaches which predictions user preference. It is compared by four different quality metrics with other approaches of recommendation systems: random prediction, content-based filtering, collaborative filtering, averaging prediction. These approaches were tested on data on films and ratings provided by users. Content-Boosted Collaborative Filtering approach improves the result by 15—20 %, compared to other approaches. The usual method of collaborative filtering calculates the similarity of two users if and only if they rated one or more identical objects. Otherwise, if they rated different objects, then these users will be completely different. However, this proposal does not take into account that if the objects themselves rated by these users are similar, then these two users can be potentially similar. This method of finding potential similar users increases the number of nearest neighbors. In addition, the ranking of like-minded people can change significantly, therefore, the value of the predictions will also change in comparison with the usual CF method.

Keywords: *recommendation system, hybrid approach, content-based filtering, collaborative filtering*

DOI: 10.17587/it.26.523-528

References

1. **Adomavicius G., Tuzhilin A.** Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,*IEEE Transactions on Knowledge and Data Engineering*, June 2005, vol. 17, no. 6.
2. **Encyclopedia** of Machine Learning. Chapter № 00338, 2010.
3. **Burke R.** Hybrid Web Recommender Systems, *Wayback Machine, The Adaptive Web, Lecture Notes in Computer Science,* Springer-Verlag, Berlin, Germany, May 2007, vol. 4321, pp. 377—408.
4. **Sidnooma Christian Kabore**. Design and implementation of a recommender system as a module for Liferay portal, Master's Thesis. Barcelona School of Computing, University Polytechnic of Catalunya. Master in Information Technologies, September 2012.
5. **Isinkaye F., Folajimi F., Ojokoh B.** Review Recommendation systems: Principles, methods and evaluation, *Egyptian Informatics Journal*, 2015,vol. 16, pp. 261—273.
6. **Ricci F., Rokach L., Shapira B.** Recommender System Handbook, New York, Springer, 2011.
7. **Melville P., Mooney R., Nagarajan R.** Content-Boosted Collaborative Filtering for Improved Recommendations, *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002),* Edmonton, Canada, July 2002, pp. 187—192.
8. **Herzog D., Worndl W.** Extending Content-Boosted Collaborative Filtering for Context-aware, Mobile Event Recommendations, *WEBIST 2016 — 12th International Conference on Web Information Systems and Technologies*, 2016.
9. **The Movies** Dataset, kaggle: online community, available at: https://www.kaggle.com/rounakbanik/the-movies-dataset.