OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.8

ОБЗОР СИСТЕМ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ

Шереметова Е.И.

Санкт-Петербургский государственный политехнический университет,
г. Санкт-Петербург, Россия

sheremetova.ei@gmail.com

В обзоре представлены различные виды систем коллаборативной фильтрации и их алгоритмы. Основным параметром для объединения алгоритмов в группы является сходство в их математической реализации. Предложенная классификация позволяет облегчить выбор алгоритма для конкретной задачи, имеющей ряд особенностей (количество пользователей и объектов в системе, время вычислений и др.).
Ключевые слова: рекомендующие системы, коллаборативная фильтрация, интеллект-карта.

Введение

Рекомендующие системы широко используются в области информационных технологий. Такие системы работают с информацией, полученной после обработки истории предпочтений пользователей.

В последнее время появилось большое количество коммерческих приложений, основанных на построении предположений о том, какие позиции могут заинтересовать пользователя системы. Именно это сделало актуальным вопрос выбора наиболее подходящего алгоритма для реализации системы.

Ниже рассматриваются получившие распространение на сегодняшний день алгоритмы коллаборативной фильтрации, а также их классификация в зависимости от типа используемых при реализации математических методов.

1. История рекомендующих систем

Первой рекомендующей системой, ставшей прародителем современных систем, является система оценки текста – одного из наиболее сложных материалов для анализа. Рекомендующая система Information Tapestry project от компании Xerox Palo Alto Research Center была разработана в 1992 году и позволяла фильтровать текстовые сообщения.

Впервые термин «collaborative filtering» был введен ее разработчиками в статье «Using collaborative filtering to weave an information tapestry» [Goldberg et al., 1992]. Кроме того, в этой статье была сформулирована идея о том, что в процесс создания рекомендаций должны быть вовлечены люди, дающие оценки изученным ими документам.

Рекомендующие системы получили известность относительно недавно – в середине 1990-х годов. В это же время появилось четкое разделение в понятиях «системы коллаборативной фильтрации» и «рекомендующей системы». Это вызвано тем, что рекомендующие системы могут основываться на разных подходах к решению главной задачи – вычислению предположительной оценки объекта конкретным пользователем. Рассмотрим существующие на сегодняшний день классификации рекомендующих систем.

2. Современные рекомендующие системы

2.1. Классификация рекомендующих систем

Рекомендующие системы в работе А.В. Пономарева были разделены на три вида – контентные, коллаборативной фильтрации и гибридные системы [Пономарев, 2013]. На рисунке 1 представлена интеллект-карта, описывающая виды рекомендующих систем. Рассмотрим каждую из этих систем подробнее.

1. Content-based – контентные системы, базирующиеся на оценке схожести собственных характеристик объектов.

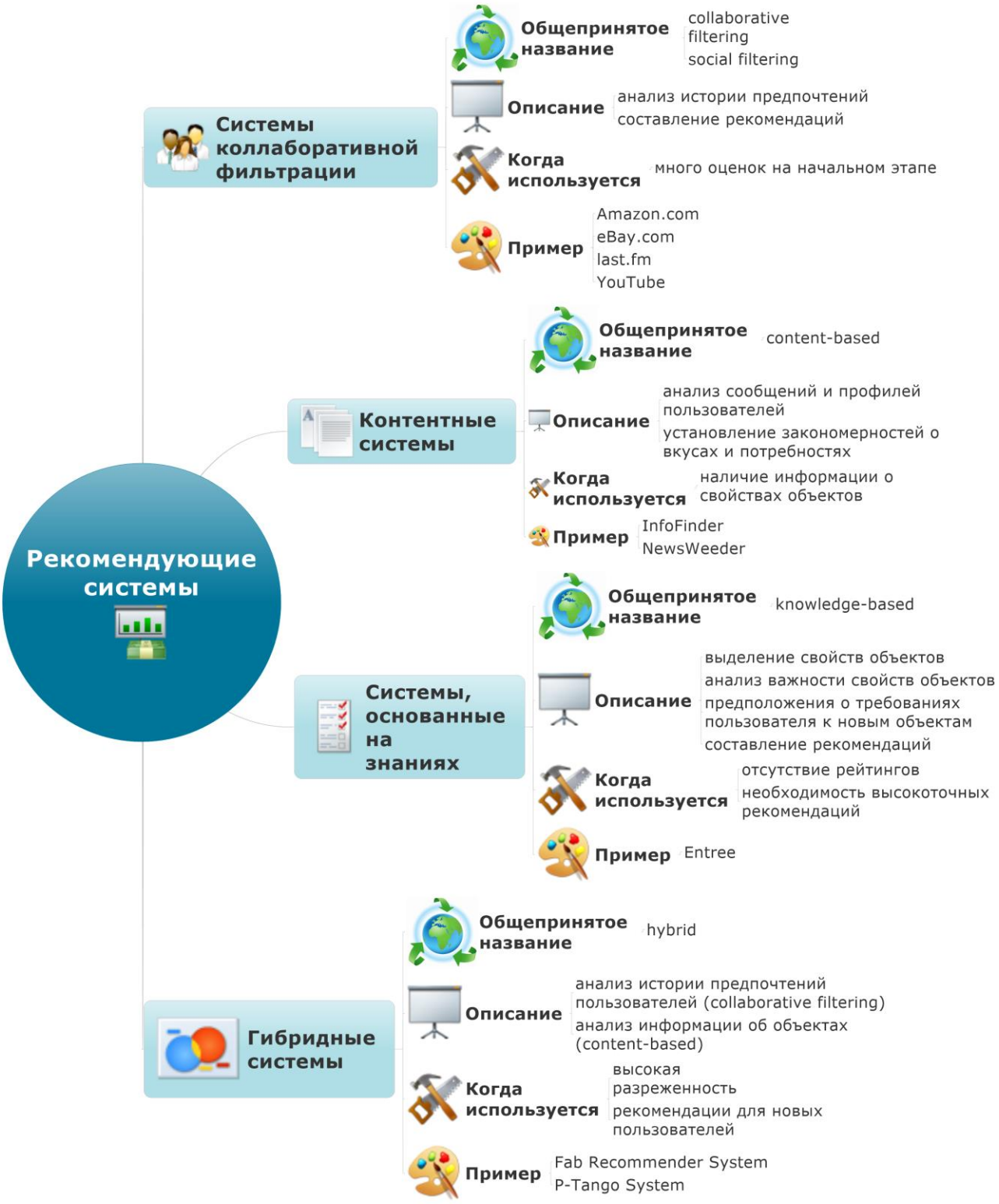

Рисунок 1 – Типы рекомендующих систем

Принцип работы заключается в анализе текстовой информации (документы, URL, сообщения пользователей, профили пользователей и др.) и установлении на ее основе закономерностей о вкусах, предпочтениях и потребностей пользователя [Su&Khoshgoftaar, 2009].

Примером content-based является система NewsDude, рекомендующая новости на основе анализа их содержания.

2. Collaborative filtering – системы коллаборативной фильтрации. Такие системы используются в сервисах Amazon.com, eBay.com, YouTube и др.

Основу их работы составляет анализ истории предпочтений пользователей (совершенные

покупки, просмотренные фильмы и т.д.). Этот подход исключает проблему зависимости работы системы от предметной области (content-based). Однако у таких систем возникают трудности при отсутствии достаточной «истории» предпочтений [Burke, 2002]. Эта проблема решена в гибридных рекомендующих системах, сочетающих в себе несколько различных подходов к построению рекомендующих систем.

3. Hybrid – гибридные системы, которые комбинируют принципы работы систем коллаборативной фильтрации и контентных систем. Преимущество этого типа систем заключается в возможности работы с разреженными матрицами оценок и с новыми пользователями системы.

В работе Robin Burke описывается новый вид гибридной системы – EntreeC, которая включает в себя методы системы коллаборативной фильтрации и методы систем, основанных на знании о предпочтении пользователя не в области конечного объекта, а в области совокупности предъявляемых к нему требований [Burke, 2000].

Таким образом, автор выделяет четвертый вид рекомендующих систем – knowledge-based recommender systems. Такой подход более сложен, однако позволяет значительно увеличить точность генерируемых рекомендаций и уменьшить количество ошибок.

Стоит принять во внимание, что точность предположений не является единственным критерием выбора алгоритма рекомендующей системы. Помимо точности к таким критериям относятся скорость вычислений, количество пользователей и объектов, разреженность матрицы и др.

2.2. Системы коллаборативной фильтрации

При реализации алгоритмов коллаборативной фильтрации в рекомендующих системах используются исходные данные в виде разреженной матрицы оценок (Пользователи-Объекты) Основной задачей алгоритма является заполнение этой матрицы и, таким образом, предоставление данных о том, какие объекты получили наивысшие предположительные оценки для каждого пользователя.

Появление большого многообразия способов вычисления неизвестных оценок привело к необходимости объединения их в группы по определенному признаку. Большинство современных авторов статей, например [Su&Khoshgoftaar, 2009], [Das et al., 2007], [Burke, 2002], выделяют 2 основных типа алгоритмов: model-based и memory-based.

1. Алгоритмы memory-based получили такое название из-за способа вычислений предположительных оценок – предположения строятся на базе рейтингов других пользователей и весов конкретного пользователя/объекта (в зависимости от метода) [Das et al., 2007]. Таким образом, оценивается показатель сходства, от которого зависит результат. Это обеспечивает значительную простоту реализации.

К методам memory-based авторы статьи Xiaoyuan Su и Taghi M. Khoshgoftaar относят методы Neighbor-based, основанные на «соседстве» (сходстве оценок) пользователей, и Item-/User-based top-N, которые из числа соседей выбирают N наиболее часто оцениваемых соседями элементов [Su&Khoshgoftaar, 2009].

2. Алгоритмы model-based основаны на построении «модели пользователя» в соответствии с историей его предпочтений. Среди них автор статьи «Google News Personalization: Scalable Online Collaborative Filtering» Abhinandan Das выделяет следующие алгоритмы: latent semantic indexing (LSI), singular value decomposition (SVD), Bayesian clustering, probabilistic latent semantic indexing (PLSI), multiple multiplicative Factor Model (MMF), Markov Decision process и Latent Dirichlet Allocation (LDA) [Das et al., 2007].

Описанные типы алгоритмов можно представить в виде интеллект-карты (рисунок 2). Типичные ошибки оформления статей.

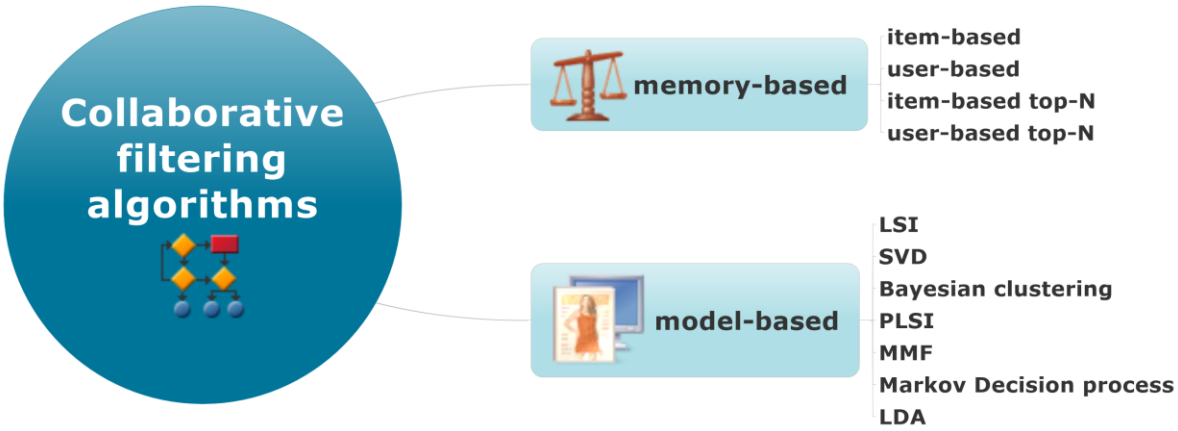

Рисунок 2 – Типы алгоритмов коллаборативной фильтрации. Классический подход

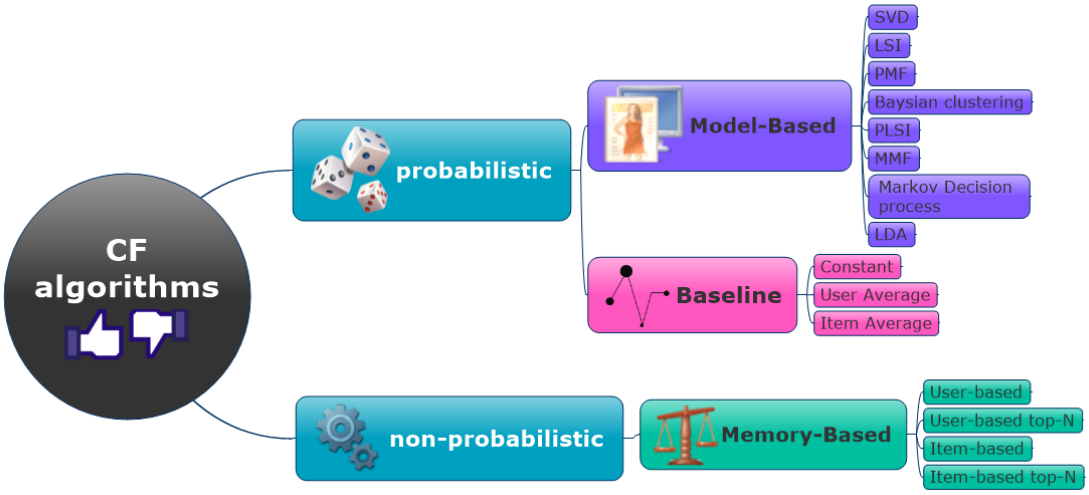

Рисунок 3 – Типы алгоритмов коллаборативной фильтрации. Расширенная классификация

3. Обзор современных классификаций алгоритмов коллаборативной фильтрации

Некоторые исследователи наряду с общепринятой классификацией, приведенной в предыдущей главе, выделяют также еще несколько типов алгоритмов. В статье [Ekstrand et al., 2011] описывается иной подход к классификации. Авторы выделяют отдельную группу вероятностных методов которые основывается на построении вероятностной модели поведения пользователя и использовании ее для построения предположений о его будущем поведении.

К вероятностным методам коллаборативной фильтрации авторы относят LDI, PLSI, SVD, Bayesian networks, Markov Decision process и Bayesian clustering (которые в предыдущей главе рассматривались как методы model-based).

Кроме того, авторы указанной статьи также выделяют еще одну группу методов Baseline Predictors и описывают ее на примере метода User Average.

Однако некоторые исследователи относят к этой группе методов также Constant и Item Average [Su&Khoshgoftaar, 2009]. Их также можно отнести к вероятностным методам, так как они основываются на вычислении вероятностных оценок.

Методы memory-based нельзя классифицировать как вероятностные, поскольку в этих алгоритмах отсутствует этап построения профиля (вероятностной модели) пользователя. Они основываются исключительно на информации о прошлых оценках пользователей и соседстве (пользователей или объектов). Таким образом, эти алгоритмы следует отнести к детерминированному подходу.

Объединив рассмотренные выше современные подходы к классификации, можно построить интеллект-карту методов коллаборативной фильтрации (рисунок 3).

4. Тенденции к развитию систем коллаборативной фильтрации

Дальнейшее развитие систем коллаборативной фильтрации неразрывно связано с решением тех проблем, которые существенно влияют на качество их работы. Ряд исследователей [Sindhwani&Melville, 2008], [Su&Khoshgoftaar, 2009] выделяют три основных вида проблем. Они представлены на интеллект-карте (рис. 4).

4.1. Проблема «холодного старта»

Согласно статье [Schein et al., 2002], такая ситуация возникает при появлении в системе нового пользователя. Многие алгоритмы не предусматривают возможность составления рекомендаций для пользователей, которые еще не оценили ни одной позиции.

Попытки решения этой проблемы были предприняты в алгоритмах model-based, в которых оцениваются скрытые характеристики пользователей, что позволяет обоснованно составлять рекомендации на этапе первого обращения пользователя к системе. Однако такие рекомендации все же имеют низкую точность.

4.2. Проблема индивидуального подхода

Среди пользователей системы, как правило, находится человек, который по своим предпочтениям значительно отличается от большинства людей. Решение этой проблемы в большинстве современных систем ограничивается использованием алгоритмов, которые не позволяют оценкам таких пользователей вносить значительные изменения в какие-либо общие результаты (таким образом, проблема просто игнорируется).

Рисунок 4 – Проблемы в коллаборативной фильтрации

Однако в статье [Ghazanfar&Prugel-Bennett, 2011] было предложено решение, которое предполагает использование SVM-регрессии для составления рекомендаций пользователям-индивидуалистам и методов кластерного анализа для обычных пользователей.

Следует обратить внимание на то, что такая система использует также и подходы, относящиеся к content-based рекомендующим системам, что обязывает отнести рассмотренный пример к гибридным рекомендующим системам (рисунок 1). Таким образом, вопрос поиска подхода на основе алгоритмов коллаборативной фильтрации все еще остается нерешенным.

4.3. Псевдоклиентские атаки

Данная проблема возникает, когда пользователи намеренно ставят высокие оценки одним товарам и низкие – товарам-конкурентам, пытаясь, таким образом, в коммерческих целях искусственно повлиять на рейтинг товаров. Такие атаки оказывают значительное воздействие на точность рекомендаций, поэтому поиск решения является наиболее актуальной темой из всех рассмотренных в этой главе.

Различные варианты решения описываются в многих современных статьях [Mehta &Nejdl, 2008], [Zou&Fekri, 2013], [Zhan&Kulkarni, 2014]: авторы предлагают использовать комбинации нескольких алгоритмов, среди которых SVD и PSA, спектральная кластеризация и user-based, Belief Propagation (model-based) и PCA – все они относятся к разным группам алгоритмов и дают разные результаты в борьбе с атаками. В данный момент продолжаются поиски решения проблемы с целью создания более точных и универсальных подходов.

Заключение

В обзоре рассмотрены основные алгоритмы коллаборативной фильтрации. В соответствии с принятой классификацией, отраженной на рисунке 3, эти методы были разделены на группы:

- вероятностные методы, которые основываются на построении вероятностных моделей пользователей;
- детерминированные методы, работающие непосредственно с матрицей оценок без создания каких-либо дополнительных профилей пользователей или объектов.

Детерминированные методы исторически были разработаны раньше и являются простыми в реализации. Они по-прежнему применяются в рекомендующих системах, в особенности в сочетании с другими алгоритмами.

Преимущество вероятностного подхода заключается в получении более точных результатов. Эти методы широко используются для решения различных проблем в рекомендующих системах, описанных в главе 4, что указывает на наличие значительных перспектив развития вероятностных алгоритмов коллаборативной фильтрации.

Библиографический список

[Пономарев, 2013] Пономарев А. В. Обзор методов учета контекста в системах коллаборативной фильтрации / А.В. Пономарев //Труды СПИИРАН. – 2013. – Т. 7. – №. 30. – С. 169-188.

[Burke, 2002] Burke R. Hybrid recommender systems: Survey and experiments / R. Burke //User modeling and user-adapted interaction. – 2002. – Т. 12. – №. 4. – С. 331-370.

[Burke, 2000] Burke R. Knowledge-based recommender systems / R. Burke //Encyclopedia of library and information systems. – 2000. – Т. 69. – №. Supplement 32. – С. 175-186..

[Das et al., 2007] Das A. S. et al. Google news personalization: scalable online collaborative filtering / A. S. Das et al. //Proceedings of the 16th international conference on World Wide Web. – ACM, 2007. – С. 271-280.

[Ekstrand et al., 2011] Ekstrand M. D., Riedl J. T., Konstan J. A. Collaborative filtering recommender systems / M. D. Ekstrand, J. T. Riedl, J. A. Konstan //Foundations and Trends in Human-Computer Interaction. – 2011. – Т. 4. – №. 2. – С. 81-173.

[Ghazanfar&Prugel-Bennett, 2011] Ghazanfar M., Prugel-Bennett A. Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution / M. Ghazanfar, A. Prugel-Bennett; – 2011.

[Goldberg et al., 1992] Goldberg D. et al. Using collaborative filtering to weave an information tapestry / D. Goldberg et al. // Communications of the ACM. – 1992. – Т. 35. – №. 12. – С. 61-70.

[Mehta &Nejdl, 2008] Mehta B., Nejdl W. Attack resistant collaborative filtering / B. Mehta, W. Nejdl // Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2008. – С. 75-82.

[Sindhwani&Melville, 2008] Sindhwani V., Melville P. Document-word co-regularization for semi-supervised sentiment analysis / V. Sindhwani, P. Melville //Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. – IEEE, 2008. – С. 1025-1030.

[Schein et al., 2002] Schein A. I. et al. Methods and metrics for cold-start recommendations / A. I. Schein et al. //Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2002. – С. 253-260.

[Su&Khoshgoftaar, 2009] Su X., Khoshgoftaar T. M. A survey of collaborative filtering techniques / X. Su, T. M. Khoshgoftaar //Advances in artificial intelligence. – 2009. – Т. 2009. – С. 4.

[Zhan&Kulkarni, 2014] Zhang Z., Kulkarni S. R. Detection of shilling attacks in recommender systems via spectral clustering / Z. Zhang, S. R. Kulkarni //Information Fusion (FUSION), 2014 17th International Conference on. – IEEE, 2014. – С. 1-8.

[Zou&Fekri, 2013] Zou J., Fekri F. A belief propagation approach for detecting shilling attacks in collaborative filtering / J. Zou, F. Fekri //Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. – ACM, 2013. – С. 1837-1840.

COLLABORATIVE RECOMMENDER SYSTEMS

Sheremetova E.I.

St. Petersburg State Polytechnic University, St. Petersburg, Russia

sheremetova.ei@gmail.com

This paper presents different types of recommender systems, especially collaborative filtering systems and their methods. To classify collaborative filtering techniques I researched similarities in their mathematical implementation. The proposed classification can facilitate further assumptions of using algorithms for specific tasks depending on several criteria (number of users, number of items, computation time etc.).

Keywords: recommender systems; collaborative filtering; mind map.

Introduction

Recommender systems are widely used in information technologies.

Nowadays such systems became an important tool in commercial applications based on large information and product spaces. It puts the problem of choosing the most suitable algorithm for implementing in the system.

This paper discusses common collaborative filtering algorithms and their classifications depending on used mathematical methods.

Main Part

The first recommender system Information Tapestry project was developed in 1992 by Xerox Palo Alto Research Center and allowed filtering text messages.

For today there are 4 main types of recommender systems:

1. content-based system use similarity between items a given user has liked in the past.
2. collaborative filtering (CF) identifies users that are similar in their tastes to given user and recommends items they have liked.
3. hybrid systems combine CF algorithms and content-based approach.
4. knowledge-based systems use a knowledge of how an item meet a user's needs.

Each of these approaches has its strengths and weaknesses. Now the most widely implemented technology is collaborative filtering.

CF algorithms use a matrix of preferences for items by users to predict a product a new user might wish to purchase or examine.

Most researchers are describing 2 groups of CF algorithms: memory-based and model-based. But there are several approaches allocating a number of different types of CF techniques which allows developing a new classification proposed in the third section of the paper.

Conclusion

The overview shows an approach to classifying CF techniques which includes 2 main types – determinate algorithms and probabilistic algorithms.

Determinate algorithms were developed earlier. There are a variety of determinate CF algorithm implementations because it is easy to implement them but low accuracy is often the main weakness.

Probabilistic methods, on the contrary, are more complicated but on the other hand they can make more accurate predictions. It indicates to the presence of great prospects for further development.