Оглавление

[Введение 3](#_Toc393088981)

[Описание предметной области 5](#_Toc393088982)

[Обзор подходов к выработке рекомендации 15](#_Toc393088983)

[Алгоритм выработки рекомендаций 26](#_Toc393088984)

[Выработка рекомендаций 28](#_Toc393088985)

[Архитектура приложения 29](#_Toc393088986)

[Заключение 30](#_Toc393088987)

[Список литературы 32](#_Toc393088988)

# Введение

Всемирная паутина представляет собой коллекцию веб-сайтов и ее веб-контента. Данная сеть имеет тенденцию к непрерывному и динамичному росту. В этой связи, эффективный поиск веб-контента становится наиболее актуальной задачей, чем когда-либо, поскольку число пользователей растет совместно с развитием Сети. Места портала с поисковыми системами являются достаточно популярными и в практике используемыми инструментами для поиска веб-контента, хотя некоторые многообещающие усилия применяются для более эффективного и действенного использования Сети, такой как семантическая паутина. Человек, пользуясь интернетом, оставляет за собой массу следов. Значительная часть этой информации может быть использована в коммерческих целях, и соответственно представляет интерес для компаний, ведущих свой бизнес в интернете. На этой почве сложилось прикладное научное направление web mining.

Основными задачами, которые решает web mining являются, в частности, персонализация контента, смысл которой заключается в том, чтобы доводить до человека информацию с учетом уже накопленных о нем данных. Декларированная информация о пользователе, такая как вручную введенные запросы, может проследить динамику интересов пользователя. Таким образом, например, можно повысить эффективность рекламы. Так, у большинства порталов есть свои поисковые системы, используемые для поиска соответствующего веб-контента с целью определения поисковых запросов пользователей.

Сегодня можно наблюдать стремительно совершенствование технологий рекомендательных систем. Первые же разработки берет свое развитие в 90-х гг., хотя толчком, выведшим создание рекомендательных систем на новый уровень, стал конкурс Netflix Prize, организованный в 2006 году компанией Netflix.

Рекомендательные системы способствуют пользователям, споткнувшимся с проблемой огромного выбора, определить, что именно лучше всего отвечает их потребностям. Наиболее развитые страны анализируют предпочтения и интересы каждого пользователя и предлагают персонифицированные рекомендации. Данные системы в процессе функционирования осуществляют сбор данных о пользователях, используя при этом сочетания явных и неявных методов. Стоит также отметить, что рекомендательные системы приобретают все большее распространение в качестве перспективной технологии, поскольку ее применение позволяет ему решить проблему выбора среди большого числа альтернатив.

Выработка рекомендация представляет собой автоматизированную форму персонализации контента. Данная система может основываться на поиске используемых веб-профилей, сопровождаемые системой рекомендаций, которая может отвечать на личные интересы пользователей. Медиа-сервис youtube, «советующий» видео-ролики на основе привязанной к учетной записи истории предыдущих просмотров, - один из многочисленных примеров.

Целью данной дипломной работы состоит в реализации приложения по выработке рекомендаций ссылок. Для достижения поставленной цели необходимо решить следующие задачи:

* разработать алгоритм выработки рекомендаций;
* облечь алгоритм в конкретное приложение.

В данной работе использованы идеи и общая схема рекомендателя ссылок, изложенные в [1].

Приложение реализовано в виде readonly REST-сервиса.

Текст состоит из введения, описания предметной области и обзора подходов к выработке рекомендаций, описания алгоритма, краткого описания архитектуры приложения, заключения и списка использованных источников.

# Описание предметной области

Дословный перевод понятия «web mining» звучит, как «добыча данных в Web». Web Intelligence (Веб-интелект) готов «открыть новую главу» в достаточно скором развитии электронного бизнеса. Возможность определять предпочтения и интересы каждого посетителя, осуществляя наблюдение за его поведением, представляет собой серьезное и критичное преимущество конкурентной борьбы на рынке электронной коммерции.

Системы web mining могут ответить на множество вопросов, включая, какие посетители являются потенциальными клиентами интернет - магазина, какая группа потребителей интернет – магазина приносит наибольшую выручку, каковы интересы конкретных посетителей или групп посетителей.

Технология web mining включает в себя методы, способные на базе данных сайта найти новые, ранее неизвестную информацию и которые в перспективе можно использовать в практике. Иными словами, web mining использует технологию data mining (рис. 1) для анализа неструктурированной, распределенной, неоднородной и значительной по объему информации, которую содержат web-узлы. Под data mining, в свою очередь, понимается мультидисциплинарная область, возникающая и развивающаяся на основе таких наук, как прикладная статистика, распознавания образцов, искусственный интеллект, теория баз данных и т.д.

Data Mining

Другие дисциплины

Статистика

Теория БД

Машинное обучение

Алгоритмизация

Искусственный интеллект

Распознание образов

Визуализация

Рис. 1. Data Mining как мультидисциплинарная область

В соответствии с таксономией web mining [31], выделяют два базовых направления:Web Usage Mining и Web Content Mining.

Web Usage Mining представляет собой обнаружение закономерностей в действиях пользователей web-узла или их групп. В данном случае анализу подлежит следующая информация:

* страницы, просматриваемые пользователем;
* последовательность просмотра страниц.

Помимо всего прочего анализируются группы пользователей, выделяемые из общего числа на основе истории просмотра web-узла.

Web Usage Mining состоит из следующих составляющих:

* предварительная обработка;
* операционная идентификация;
* инструменты обнаружения шаблонов;
* инструменты анализа шаблонов.

Так, при применении web mining перед разработчиками возникает два типа задач. Первый относится к сбору данных, второй – к использованию методов персонификации. В конечном итоге результаты сбора некоторого объема персонифицированных ретроспективных данных о конкретном пользователе, система накапливает определенную информацию о нем и может рекомендовать ему, к примеру, определенный набор товаров или услуг. На базе имеющихся знаний о посетителях сайта web-система в состоянии выявить определенные группы клиентов, а также рекомендовать им товары или же предлагать товары в рассылках.

Под Web Content Mining понимается автоматический поиск и извлечение качественной информации из различных источников Internet, перегруженных «информационным шумом». В данном случае также речь идет о различных средствах кластеризации и аннотировании документов.

Таким образом, основными задачами web mining согласно [1] можно разделить на следующие категории:

* предварительная обработка данных для web mining;
* обнаружение шаблонов и открытие знаний с применением ассоциационных правил, временных последовательностей, кластеризации и классификации;
* анализ полученной информации.

Традиционный процесс data mining включает в себя следующие этапы:

* анализ предметной области;
* постановка задачи;
* подготовка данных;
* построение моделей;
* проверка и оценка моделей;
* выбор модели;
* применение модели;
* коррекция и обновление модели.

Рассмотрим первый этап процесса data mining – анализ предметной области.

Исследования представляет собой процесс познания определенной предметной области, объекта или явления с конкретной целью. Процесс исследования – наблюдение свойств объектов для выявления и оценки важных с позиции субъекта-исследователя закономерных отношений между показателями данных свойств.

В связи с чем, решение любой задачи в сфере разработки программного обеспечения должно начинаться с изучение предметной области.

Предметная область представляет собой бесконечную часть реального мира, содержащую в себе существенные и не значащие данные с позиции проводимого анализа. Описание предметной области всегда осуществляется в какой-то знаковой системе.

Знания о предметной области включают факты, относящиеся к данной области, характерные для нее закономерности, гипотезы о возможных связях между процессами, фактами и явлениями в ней, процедуры для решения типовых задач. Экспериментальные знания, в свою очередь, представляют собой знания, располагаемые специалистов в некоторой предметной области.

Базой успешной оценки алгоритмов рекомендательной системы является необходимым понимание того, какими свойствами должна обладать база данных дл того, чтобы наилучшим образом смоделировать задачи, по решению которых оценивается качество работы рекомендательной системы. Так, можно выделить три категории свойств баз данных:

* свойства предметной области, отражающие характер рекомендуемого контента, а не какую-то конкретную систему;
* внутренние свойства, которые отражают суть конкретной рекомендательной системы, из которой берутся данные (и, возможно, ее методы сбора данных);
* свойства выборки, которые отражают характер распределения данных.

Свойства предметной области состоят из:

* тематики рекомендуемого/ранжируемого контента и связанный с этим контекст, в котором имеет место ранжирование или выработка рекомендаций;
* задач, стоящих перед пользователем, помогающие решать рекомендательной системой;
* потребности в новинках и качестве;
* соотношения затрат и выгод от истинных/ложных позитивных/негативных рекомендаций;
* степени детализации истинных предпочтений потребителей.

Как правило, рекомендательные системы разрабатывались для сфер индустрии развлечений (музыка, кино и др.), однако существуют и текстовые модели для фильтрации массивов документов (к примеру, Usenet news). При изучении предметной области должна быть создана модель. Знания из различных источников должны быть формализированы при помощи каких-либо средств. [[1]](#footnote-2) В данном случае могут быть использовано текстовое описание предметной области или специализированные графические нотации. Так, существует большое количество различных методик описания предметной области, в частности, методика структурного анализа DADT и основанная на нем IDEF0, диаграммы потоков данных Гейна-Сарсона, методика объективно-ориентированного анализа UML и иные. Модель предметной области осуществляет описание процессов, которые происходят в предметной области, и данные, которые в этих процессах применяются.

Так, даже в рамках одной конкретной предметной области существуют несколько контекстов. Рекомендательные системы по кинофильмам могут быть сформированы в on-line режиме, в видеомагазине или в салоне видеопроката, а также как часть системы цифрового TV и ее производных (TiVo, цифровые видеорекордеры).

Одним из самых важных свойств предметной области, по мнению экспертов, заключается в соотношении между желанием узнать что-то новое и желанием получить рекомендацию высшего качества. В некоторых предметных областях основная цель пользователя заключается в нахождении рекомендаций относительно еще неизвестных ему объектов. Так, McNee и другими было оценены рекомендательные системы и обнаружено, что пользователи были уже удовлетворены тогда, когда среди рекомендуемых объектов был хотя бы один единственный полезный объект, ранее им незнакомый.

Рекомендации могут являться точными, однако они не меняют потребительского поведения. Желание получить высококачественные рекомендации противостоит желанию найти что-то новое. Как следствие, это интуитивно отражает желание пользователя полностью довериться рекомендациям в принятии решение о приобретении/ прочтении/ прослушивании какой-либо вещи, а не просто рассматривать полученные рекомендации как один из многих факторов, помогающих в принятии решения.

В качестве иного важного свойства предметной области можно выделить соотношение затрат и получаемых пользователями выгод от предоставленных рекомендаций. В рекомендательной системе, функционирующей в области видео, стоимость ложных положительных рекомендаций (то есть высоко оцененных и активно рекомендуемых системой объектов, но оказывающихся недостаточно интересными для пользователей) для пользователей достаточно низка ($3 и 2-3 часа потраченного на просмотр времени). Стоимость ложной негативной рекомендации (то есть. плохой рекомендации в отношении объекта, интересного потребителю) почти равна 0, а выгода от рекомендаций для пользователя огромна (из года в год огромное число фильмов выходит на экраны, и просмотр всего ассортимента видеомагазина может оказаться очень тяжелым процессом). Данное обстоятельство в большей степени объясняет то, почему рекомендательные системы в области видел достаточно успешны. Другие предметные области, имеющие схожие свойства (к примеру, художественная литературы) имеют массивы данных, схожие с имеют схожие с видео-рекомендательно-системными, и результаты, которые показывают эти же массивы, возможно, по нашему мнению, транслировать и на те другие предметные области с целью более детального обсуждения анализа соотношения между получаемыми пользователями выгодами от использования рекомендательных систем и затратами (денежными и временными) на базе коллаборативной фильтрации.

Иным не менее важным свойством, права неявным, является степень детализации истинных предпочтений потребителей. Стоит заметить, что степень детализации пользовательских предпочтений может быть отличной от диапазона и степени детализации оценок (рейтингов) объектов, что представляет собой внутреннее свойство информационного массива.

В итоге, думается, что было бы ошибочно оценивать алгоритм генерации рекомендаций на основе массива данных со свойствами, абсолютно несоответствующими предметной области. Наиболее важным является тот факт, что для того, чтобы задачи, для решения которых разрабатывается алгоритм, являлись схожими с задачами, которые решает система, у которой берется собранный ее информационный массив. В случае, если данные задачи не совпадают, то, как правило, не совпадают и иные характеристики. К примеру, первоначально система MovieLens поддерживала только решение пользовательской задачи «Найти Хорошие Объекты». В результате потребителям всегда предоставлялись самые выигрышные объекты, что, в конечном счете, для хороших объектов имелось больше рейтингов, чем для плохих (как правило, пользователь должен был сделать прямой запрос о ранжировании плохого объекта). Таким образом данные MovieLens, скорей всего, не имеют достаточно рейтингов по менее популярным объектам. Возможно, являлось неуместным пользоваться такими данными для оценки нового алгоритма, который преследовал своей целью выполнение задач по аннотированию сообщений в их контексте. Безусловно, в случае, если алгоритм нацелен на решение общих задач, то лучше всего для его тестирования выбрать массивы данных, которые охватывали бы большое множество сфер и тем.

Состав внутренних свойств информационных массивов имеет следующий вид:

* рейтинги (оценки) объектов - явные, неявные или и те и другие;
* шкала, по которой ранжируются объекты;
* многомерность оценки объектов;
* наличие или отсутствие временных отметок у рейтингов (т.е. указание времени, когда был проранжирован объект).

Явные рейтинги предоставляются напрямую самими пользователями (к примеру, «Пожалуйста, проранжируйте этот объект по шкале от 1 до 5»), в то время как неявные рейтинги выводятся на базе анализа поведения конкретного пользователя.

Шкала ранжирования представляет собой диапазон и степень детализации рейтингов. Простейшая шкала – одинарная, в частности, понравившиеся объекты помечаются, а иные – нет. Унарные шкалы, как правило, встречаются применительно к коммерции, где все, что известно, так это то, приобрел пользователь товар или нет. Данные именуют унитарными, а не бинарными, поскольку непокупка товара Х является не обязательным условием того, что пользователю он не понравился.

Бинарные данные, в свою очередь, включают в себя отдельные обозначения ненравищихся пользователям товаров. Системы, оперирующие с явными рейтингами, как правило, поддерживают 5-ти, 7-ми, или 10-ти балльные шкалы.

Большая часть рекомендательных систем имеют оценку по одному критерию, в то время как в научных и коммерческих системах пользователи могут оценить объект по нескольким критериям. Так, ресторанные гиды Zagat традиционно оценивают рестораны по качеству кухни, обслуживанию и внутреннему декору. Рекомендательные системы по кинофильмам могут разделять оценку фильма на 3три составляющие: сюжет, игра актеров и специальные эффекты. На сегодняшний день до сих пор является сложным найти массивы данных с многомерными оценками объектов.

Наличие временных меток является свойством сбора данных, что особенно необходимо в тех сферах, где ожидается перемена в потребительских интересах, или где реакция пользователей на объекты зависит от их истории взаимодействия с другими объектами.

В качестве иных внутренних свойств информационных массивов можно выделить те, которые касаются практики сбора данных:

* записываются ли предоставляемые пользователю рекомендации;
* наличие демографической информации о пользователе или информации о содержании объекта. Стоит отметить, что на сегодняшний день имеется малое количество рекомендательных систем, записывающих рекомендации, которые предоставляются пользователю, что затрудняет отделение прошлых рейтингов от перспективных, а также смещает оценки. Способность получение демографических данных варьируется в зависимости от системы к системе и собираемых данных. Так, базы данных MovieLens и EachMovie собирают ограниченную демографию, хотя эксперты отмечают, что большая доля демографической информации может представляться ложной по причине подозрения пользователя в маркетинговых вопросах. Можно было ожидать большую достоверность демографических данных в случае, если пользователи поверят в то, что это преследует цели по построению рекомендательной системы.
* смещения, которые связаны со сбором данных. Большое количество информационных массивов имеют смещения, которые основаны на механизме, при помощи которого пользователи имеют возможность ранжировать объекты.

Свойства выборки включают в себя большое количество статических характеристик, обычно рассматриваемых при оценке данных, а именно:

1. плотность рейтингов – устанавливаются в целом, иногда измеряется как средний процент объектов, проранжированных пользователем. Поскольку большинство баз данных имеют неравномерное распределение популярности, то плотность оценок может подвергнуться искусственному манипулированию при помощи включения/исключения тех или иных объектов;
2. число или плотность рейтингов, которые выставляются пользователями, для которых вырабатываются рекомендации - представляет собой опыт пользователя по использованию системы на момент выдачи рекомендаций. Рейтинги от пользователей, имеющие значительный опыт работы с системой, могут быть скрываться для симуляции условий, как будто это новые пользователи;
3. общий размер и характеристики распределения массива данных - некоторые массивы данных включают в себя больше объектов, чем пользователей, одновременно как большинство информационных массивов содержат больше пользователей, чем объектов.

Каждое из вышеуказанных свойств выборки может качественным образом влиять на успех работы различных алгоритмов, а также может отражать определенные правила работы конкретной рекомендательной системы.

Безусловно, что два конкретных свойства распределения данных информационного массива имеют наибольшее значение. Взаимная связь между числом объектов и числом пользователей в базе может способствовать определению, насколько легко будет выстроена корреляция между числом объектов и числом пользователей – данный выбор может привести к различного рода результатам работы среди алгоритмов.

Распределение рейтингов пользователей и объектов также может способствовать выбору параметра и алгоритма, по которым проводится оценка качества работы рекомендательной системы. В данных системах можно наблюдать экспоненциальную кривую популярности, поскольку определенные объекты имеют намного больше оценок по сравнению с иными.

Таким образом, можно отметить, что между категориями свойств предметной области, внутренних характеристик и свойств выработки можно наблюдать значительное взаимодействие.

Большинство научных публикаций, посвященных алгоритмам рекомендательных систем, основанных на коллаборативной фильтрации, базируются на данных одного из этих трех информационных массивов. Отсутствие достаточно широкого диапазона общедоступных баз данных систем коллаборативной фильтрации (особенно со значительным числом рейтингов) остается одной из самых больших проблем в этой сфере.

Многие исследования не обладают ресурсами для создания систем, пригодных для производственного применения и способных собрать достаточно информации для проверки исследовательских гипотез, и таким образом они часто вынуждены ограничивать свое исследование проверкой гипотез, которую можно произвести с использованием небольшого числа существующих баз данных.

Таким образом, первые этап процесса data mining заключает в том, что от того насколько верно смоделирована предметная область, зависит успех дальнейшей разработки приложения data mining.

# 

# Обзор подходов к выработке рекомендации

Рекомендательная система представляет собой систему, дающую рекомендации пользователю.[[2]](#footnote-3) Данное достаточно обобщенное определение понятия наиболее часто встречается в научной литературе, что говорит о том, что под категорию рекомендательных систем подпадает достаточное большое количество различных технологий.

Рекомендательная система позволяет пользователю обращать внимание на свои предпочтения и предоставляет результаты, полезные для него, базируясь на оценках иных пользователей и предположениях самой системы. Стоит заметить, что основное отличие рекомендательной системы от поисковых систем заключается в том, что она не требует четкого запроса. Пользователю предлагается оценить некоторые объекты из коллекции, и на основании его оценок и сравнения их с оценками других пользователей строятся предположения и возвращаются наиболее близкие к ним результаты. Более подробное описание и спецификацию рекомендательной системы указывают при описании конкретного способа ее реализации.

Каждая рекомендательная система предполагает то, что пользователю предоставляется большое количество альтернатив, среди которых он осуществляет выбор. Пользователю может быть недостаточно опыта и знаний для самостоятельного отброса альтернативы, нес ответствующей его потребностям. Так, он в определенной форме (имплицитно или эксплицитно) предоставляет системе информацию относительно своих предпочтений, одновременно о некоторых альтернативах он может даже и не знать. Таким образом, рекомендательная система представляет собой систему, используемую определенный алгоритм фильтрации и существующие сведения о потребностях пользователя, для рекомендации ему набора альтернатив, которые считает наиболее полезными дня него.

Рекомендации предоставляются на основе уже совершенных действий, посещений веб-сайтов, а также приема обратной связи (выставления рейтингов, анкет). В качестве примеров можно выделить следующие: Amazon. Com, Музыка на Yahoo!, TV Recommender, Cinemax.com и т.д. Так, в случае, если ресурс Amazon. Com не может идентифицировать иди еще «не знает» пользователя, тот данный пользователь получит самые общие рекомендации. В процессе совершения покупок осуществляется сбор информации о конкретном пользователе, что улучшает качество рекомендаций, которые учитывают индивидуальные характеристики клиента. Применяя алгоритм совместной фильтрации, базирующейся на товарах, матрица пользователей сходства товаров (косинусы углов между векторами товаров) формируется в отложенном режиме.

В рекомендательных системах используются явный и неявный сбор данных. При использовании явного сбора данных пользователю необходимо заполнить опросные анкеты с целью выявления предпочтений, а при неявном – с целью выявления предпочтений и составления рейтингов происходит автоматическое протоколирование его действий. Самый очевидный способ неявного сбора информации характерен для электронной коммерции, в которых рейтинг товара у пользователя оценивается в зависимости и от количества заказанных единиц, которые пользователь включил в свой заказ.

При разработке рекомендательных систем использую два типа фильтрации:

* коллаборативную фильтрацию;
* контентно-основанные рекомендации.

Помимо всего прочего можно выделить отдельный тип, объединяющий черты вышеуказанных типов, - гибридная фильтрация.[[3]](#footnote-4)

Рассмотрим каждый из типов рекомендательных систем более подробно.

При фильтрации содержимого могут использоваться профили пользователей и объекты. Для простоты описания рекомендательных систем термин «объект» обозначим, как то, что рекомендует система. Так, профили пользователей могут включать:

* демографическую информацию либо ответы на определенный набор вопросов;
* названия жанров, имена актеров, имена исполнителей и т.п. – в зависимости от типа объекта.

Рекомендательные системы, разработанные на основе демографической фильтрации, используют описание пользователей для установки связи между определенными объектами и категориями пользователей, которые заинтересованы в них.

Система с демографической фильтрацией требует наличие информации относительно интересов пользователей, однако дает не персонализированные рекомендации, а лишь обобщенные. Данный подход не подразумевает никакого механизма адаптации пользователя к динамике его интересов. При регистрации пользователя относят к конкретной предметной категории с аналогичными интересами, при последующих изменениях предоставляемые рекомендации не подлежат изменению. С целью решения данной проблемы необходимо внедрить дополнительное редактирование регистрационной формы, иначе пользователю необходимо регистрироваться заново.

В качестве преимущества данного метода можно выделить относительную простоту реализации. Зачастую демографическая фильтрация применяется в комплексе с иными видами фильтрации, что приводит к совершенствованию предоставленных рекомендаций.

На основе стереотипной классификации создаются профили группы пользователей. При использовании данного метода фильтрации пользователю необходимо заполнить анкету при регистрации, данные из которой будут применяться для отнесения его к определенной группе. Система с данным типом фильтрации базируется на обобщении интересов пользователей одной группы и рекомендует им одинаковые объекты.

Неперсонализированные системы рекомендуют своим пользователями определенные объекты, применяя общий рейтинг данного объекта, который определяется при помощи голосования всех остальных пользователей данной системы. Рекомендации такого характера не зависят от пользователя, в ней нуждающихся. Результатом применения данной системы являются одинаковые рекомендации, которые выдаются на один и тот же запрос. Преимуществом данной системы является минимальное взаимодействие системы и ее пользователя.

В рекомендательных системах, в которых применяется фильтр за содержанием, пользователи представляют собой независимых от иных участников системы. С целью генерирования системой рекомендаций, ей необходим профиль пользователя, который будет отражать его интересы. В профиле в определенной форме, которая, как правило, зависит от выбранного способа предоставления и алгоритмов, хранится информация об объектах (выбранных алгоритмов и свойствах), интересующих пользователей. [[4]](#footnote-5) Помимо всего прочего система содержит в себе информацию обо всех предметах, ею рекомендуемых. Данная система базируется на описании предметов по профилю пользователя и осуществляет поиск в собственной базе данных сходные предметы, после чего рекомендует их ему.

Подход с применением данного типа фильтрации является достаточно уместных в случаях, когда у пользователя есть четко определенные, специфические интересы, и он ищет схожие рекомендации.

Основное преимущество фильтрации за содержанием заключается в том, что первоначально для предоставления рекомендаций не требуется большое количество зарегистрированных пользователей, иными словами, рекомендации не зависят от иных пользователей системы. Важным ограничением данного метода является неспособность системы данной фильтрации рекомендовать новые объекты, не следующие из интересов пользователя.

С целью наличия высокого качества предоставления рекомендаций системой необходимо, чтобы она постоянно имела обратную связь с пользователем и узнавала релевантность предоставленных рекомендаций. [[5]](#footnote-6) Помимо всего прочего, данные сведения направляются на обновление профиля пользователя. Однако возникает иная проблема – пользователь всегда хочет свести к минимуму свое воздействие с системой и достаточно неохотно делиться информацией для обратной связи. Данная проблема является типичной для всех типов рекомендательных систем.

А поскольку рекомендации целиком базируются на информации из профиля пользователя, то чем меньше сведений предоставляет пользователь, тем менее полезным будет набор рекомендаций, которые ему предложит система. Чтобы решить эту проблему используются различные техники автоматического сбора информации для поддержки профиля, что затрудняет реализацию такой системы и не всегда является эффективным. Рекомендационные системы с фильтрацией за содержанием чаще всего используются в системах электронной коммерции. Этот метод применяется также в сочетании с коллаборативной фильтрацией.

Рекомендательные системы с использованием коллаборативной фильтрации отыскивают пользователей с общими интересами и рекомендуют те объекты, которые были высоко оценены такими пользователями. Подобный тип фильтрации похож на то, как мы спрашиваем совет у коллег и друзей, поэтому его иногда называют социальным фильтрованием. Интерес каждого пользователя сохраняется в виде системы рейтингов, в которых видно насколько пользователю нравится или наоборот, не нравится тот или иной предмет. Такая система не обязательно должна иметь подробное описание предмета, ей вполне хватает лишь оценки пользователей о нем. Преимущество такой системы, в отличие от фильтрования за содержанием, заключается в том, что система не генерирует рекомендации исходя лишь из интересов одного пользователя.

Рекомендации в системе с коллаборативной фильтрацией в значительной степени зависят от формулировки «сходства» пользователей. Эта мера «сходства» должна быть установлена определенным образом. Для этого, рекомендательные системы используют различные метрики - векторное сходство, коэффициенты Пирсона, байесовскую вероятностную классификацию.[[6]](#footnote-7) Одним из наиболее используемых алгоритмов для определения сходства является алгоритм «ближайшего соседа». [[7]](#footnote-8)

Методы коллаборативной фильтрации могут быть подразделены на основанные на окружении и основанные на модели подходы. Большинство данных подходов можно обобщить под единый алгоритм, полученные в итог в следующих шагах:

* присвоение веса всем пользователям относительно степени активности участия;
* выбрать k-пользователей, у которых есть самая высокая степень схожести с активным пользователем – обычно называемым окружением;
* вычислить прогноз из взвешенной комбинации оценок выбранных соседей.

Несколько расширений, основанных на окружении коллаборативной фильтрации, которые привели к улучшению производительности, рассмотрены ниже.

Item-based коллаборативная фильтрация подразумевает применение ее к миллионам пользователей и элементам, когда стандартные алгоритмы коллаборативной фильтрации, основанные на окружении, не могут хорошо масштабироваться из-за вычислительной сложности подобных пользователей. В качестве альтернативы Linden, Smith and York было предложена коллаборативная фильтрация item-to-item, где вместо того, чтобы соответствовать подобным пользователям, они должны соответствовать расчетным элементам пользователя подобным элементам. На практике данный подход приводит к более быстрым онлайновым системам, и зачастую результаты проявляются в улучшенных рекомендациях.

Для коллаборативной фильтрации item-based также можно использовать альтернативные метрики, такие как скорректированное подобие косинуса. Хорошее эмпирическое сравнение изменений item-based методики может быть найдено в Sarwar и др.

User-based коллаборативная фильтрация заключается в предложении товаров, которые приобретаются похожими пользователями. Рекомендации осуществляются посредством усреднения рейтинга товаров, предоставляемого другим пользователям, с весами по степени похожести пользователей.

Таким образом, основной принцип рекомендательной системы в данном случае представляется следующим образом: похожим пользователем рекомендуются похожие товары. Похожесть, в свою очередь, определяется одним из способов: content – based и transaction-based.

Transaction-based подразумевает определение схожести товаров, как правило, через оформление их в одну транзакцию, а схожесть пользователей – через совершение схожих покупок.

При применении content – based схожесть определяется на основе характеристик пользователей и товаров. Так, для товаров применяются следующие описания: сюжет, киношкола, режиссер; общее музыкальное направления; функциональное назначение, ценовая группа, категория. Для пользователей – демографические данные, интересы, вкусы, предпочтения из заполненных клиентами анкет. В данном случае рассмотрению подлежат все пользователи и пункты анкет, после чего рекомендации делаются без отношения к специфическим особенностям отдельных пользователей и пунктов. Однако можно выполнить лучше персонализированную рекомендацию, зная больше о конкретном пользователе, в частности, демографической информации или пункте о жанре любимых фильмов и др.

Много исследований в данной области сосредоточилось на рекомендациях пунктов, связанных с текстовым содержанием, таким как веб-страницы, книги, фильмы. Также несколько подходов рассматривают проблему в отношении задач информационного поиска (IR), где содержание, связанное с предпочтениями пользователей, рассматривается как вопрос. Например, в NewsWeeder документу присваивается рейтинг в tf-idt векторы слова, а затем усредняется, чтоыб получить вектор прототипа каждой категории для пользователя. Чтобы классифицировать новый документ, сравнивается каждым вектором прототипа и дается предсказание рейтинга, базирующегося на подобии косинуса каждой категории.

Взвешенные значения: данный метод характерен для активного пользователя с целью высокой корреляции соседей, основанной на прерывающихся элементах. Данные соседи на основе небольшого количества прерывающихся элементов имеют тенденцию к плохому прогнозированию.

Голосование по умолчанию: альтернативный подход к контакту с корреляциями на базе немногих элементов должен принять значение по умолчанию для оценки элементов, которые не явно оценены. Таким образом, в данном случае является возможным вычисление корреляции, используя объединение элементов, оцененных пользователями, соответствующими в противоположность перекрестку. Такая избирательная стратегия значений по умолчанию, как показывает практика, улучила коллаборативную фильтрацию по Бризу и др.

Обратная пользовательская частота: при изменении подобия между пользователями элементы, которые оценивали все, не так полезны, как менее обобщенные элементов. Составить этот Бриз и другие представленные понятия обратной пользовательской частоты, которая вычисляется, как fi = log n/ni , где ni – число пользователей, которые оценивали элемент i из общего количества n пользователей. Основное предположение данного подхода заключается в том, что элементы, которые универсально любимы или нелюбимы, оцениваются чаще, чем другие.

Рекомендации, основанные на модели коллаборативной фильтрации (Model-based): представляет собой рекомендации, которые совершены при оценке параметров статистических моделей для пользовательских рейтингов. Например, Billsus и Pazzani описывают ранний подход с целью сопоставления коллаборативной фильтрации с проблемой классификаторов для каждого активного пользователя, представляющего пункты особенности как показатели и доступных рейтингов как этикетки, что возможно вместе с методами сокращения размерности. Иные прогнозирующие методы моделирования были также использованы с похожими способами.

Позже скрытый фактор матричные модели факторизации появились в качестве современной методологии в этом классе методов. В отличие от базируемых методов, которые производят рекомендации, основанные на статистических понятиях подобия между пользователями или между пунктами, скрытые факторные модели предполагают, что подобие между пользователями и пунктами одновременно вызвано некоторой скрытой более низко-размерной структурой в данных. Например, рейтинг, который пользователь дает кино, как могло бы предполагаться, зависел бы от немногих неявных факторов, таких как пользовательский вкус через различные жанры кино.

Таким образом, первоначально формируется описательная модель предпочтений пользователя, товаров и взаимосвязи между ними, затем формируются рекомендации на основании полученной модели.

В качестве преимуществ Model-based можно выделить: наличие модели, дающей большее понимание формируемых рекомендаций, а также то, что 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. Недостатки: не поддерживают инкрементального обучения, меньшая точность прогноза, чем Memory-based.

При использовании Memory-based первоначально формируется описательная модель товаров и предпочтений пользователей, а также взаимосвязи между ними, а затем подлежат формированию рекомендации на базе полученной модели. Преимуществами Memory-based являются:

* наличие модели, которая дает большее понимание формируемых рекомендаций;
* процесс формирования рекомендация проходит в два этапа: 1) ресурсоемкое обучение модели в отлаженном режиме; 2) простое вычисление рекомендаций на базе имеющейся модели в реальном времени.

Недостатками модели является не поддержание ее инкрементального обучения и меньшая точность прогноза.

Главная проблема, которая возникает при разработке рекомендательной системы с коллаборативной фильтрацией - это проблема «холодного начала».[[8]](#footnote-9) Система не будет способна эффективно генерировать рекомендации, если большинство из пользователей не укажут свои интересы в профиле. Другим важным моментом, который необходимо учитывать при разработке подобных систем является то, что она должна быть масштабируемой.

Гибридные системы обычно комбинируют коллаборативную фильтрацию и фильтрацию за содержанием, что позволяет решить ряд проблем, имеющих место при применении этих методов по отдельности. Так, недостатком систем с фильтрацией по содержанию, как уже было отмечено, является их ретроспективный характер, иными словами, они неспособы порекомендовать новые предметы. Данную проблему способна решить коллаборативная фильтрация. Рекомендации, которые генерируются при помощи коллаборативной фильтрации, создаются на базе профилей остальных пользователей, в связи с чем, не всегда похожи на объекты, указанных активными пользователями в своих профилях.

В то же время, достаточно весомый недостаток системы с коллаборативной фильтрацией заключается в неспособность давать рекомендации пользователям с нетипичными, оригинальными интересами, вкусами, учитывая нехватку пользователей с аналогически вкусами, которые представляют собой основу генерации рекомендаций. В таких случаях является эффективнее фильтрация по содержанию. Также данный тип фильтрации решает проблему «холодного начала». Пока система не наберет достаточного количества пользователей, применяется фильтрация по содержанию, как только достаточное количество пользователей достигнуто - система начинает работать при помощи методов коллаборативной фильтрации.

Гибридная система информации об интересах и вкусах пользователей предоставляет в профиле в двух видах, в частности, как[[9]](#footnote-10):

* набор атрибутов определенного предмета;
* оценка пользователем.

Данная характеристика представляет одновременно преимуществом и недостатком системы. Преимущество заключается в большом количестве сведений, которые предоставляют возможность применять эффективные алгоритмы фильтрации и генерировать полезные рекомендации. Недостаток же заключается в том, что пользователю необходимо ввести большое количество информации, что делается, как правило, неохотно либо вообще не делается. К сожалению, сочетание коллаборативной фильтрации и фильтрации по содержанию не представляет собой решение проблем маштабируемости системы и защиты приватности.

# Алгоритм выработки рекомендаций

Алгоритм выработки рекомендаций состоит из двух частей. Первая часть парсирует access.log веб-сервера и формирует user profiles. Вторая часть на основе построенных user profiles считает рекомендацию.

Предварительная часть алгоритма.

Исходным источником информации является access.log веб-сервера, логирующий обращения к сайту в строках следующего вида (nginx, combined format):

'192.33.5.1 - - [21/Apr/2013:23:16:10 +0400] "GET /some\_page.html HTTP/1.1" 200 498 "-" "-"'.

Лог должен быть распарсен до получения пар <ip, url> для каждого удачного (status code 200) GET-запроса. Каждая такая пара соответствует факту запроса ресурса с определенного ip.

Включать POST-запросы в выборку вряд ли стоит: хотя доступ к некоторым ресурсам может требовать авторизации, например, собственно пост пользовательских данных будет идти отдельно от запроса ресурса, и выражен тот будет через GET. Возможно, для большей реалистичности стоит включать запросы типа 403 Forbidden или 401 Not Authorized - это также характеризует интересы пользователей.

При запросе какой-либо страницы браузер как правило тянет массу статики: css, js, картинки, флеш и т.п. Это «шум», он ничего не говорит о предпочтениях пользователей, и его нужно отфильтровывать.

Access.log веб-сервера Apache мне не знаком и не рассматривается, но и пользоваться им на сегодняшний день перестали.

Отметим, что nginx используется для проксирования запросов. Выполняются они в контейнерах приложений, которые имеют собственные гораздо более информативные логи. Они нас не интересуют.

Итого, на первом шаге для каждого ip мы должны получить user session - булевский вектор запрошенных ресурсов.

Далее, полученное множество user sessions необходимо кластеризовать. Каждый кластер должен соответствовать некоторому «паттерну поведения» пользователей. «Паттерн поведения» в нашем случае будет группировать ресурсы сервера и позволит строить простые суждения типа: «Если человек посетил страничку X.html, то он возможно посетит и страничку Y.html, ведь так сделали остальные посетители странички X.html». Алгоритмов кластеризации существует довольно много, и для простоты был выбран k-means с евклидовым расстоянием в качестве меры близости векторов. Более подробное описание данного алгоритма можно найти в работе Вонга и Хартингана.

Алгоритм k-means строится на основе k-кластеров, которые располагаются на возможно больших расстояниях друг от друга. Основным типом задач, решающего алгоритмом k-means, является наличие предположений (гипотез) относительно числа кластеров, одновременно они должны быть различными настолько, насколько это является возможным. Выбор числа k-кластеров может основываться, как на результатах предшествующих исследований, теоретических соображениях или интуиции.

Следовательно, на втором шаге user sessions кластеризуются.

Каждому полученному кластеру следует сопоставить user profile - определенный на схеме вектор частот появления каждого ресурса в кластере. User profile позволяет судить об относительной заинтересованности в ресурсах для человека, чье поведение удовлетворяет этому паттерну.

Посчитанные user profiles составляют результат предварительной части алгоритма выработки рекомендаций, и должны быть сохранены в базе данных для последующего использования.

# Выработка рекомендаций

Рекомендация - множество пар, где каждому ресурсу поставлено в соответствие действительное число от нуля до единицы, соответствующее предполагаемой степени «интересности» ресурса для данного пользователя. Это словарь оценок с именами ресурсов в качестве ключей.

Рекомендация вычисляется для конкретной user session. Т.е., алгоритм принимает булевский вектор и возвращает словарь степеней принадлежности.

Множество user profiles на этом этапе интерпретируется как нечеткое отношение на схеме user profile X url и соответствует базе правил вида:

"Если user session близка к user profile (i), то ей подходит/интересен url (j)".

В качестве меры близости использован косинусный коэффициент (мера Охаи).

Далее делается нечеткий вывод по алгоритму Мамдани, т.е. для каждого ресурса в каждом user profile вычисляется минимум близости user session к user profile и значения импликации profile => url, а полученные значения консеквента агрегируются с помощью операции max.

Таким образом, итерацией по ресурсам мы получаем словарь степеней принадлежности.

# Архитектура приложения

Приложение реализовано как readonly REST-сервис. Сервис принимает POST-запросы с user session в json формате и возвращает json со словарем оценок.

Приложение написано на python 2.7. В качестве сервера приложений использован uWSGI. Данные хранятся и берутся при обработке запроса из Redis.

Часть приложения, что производит кластеризацию и формирует user profiles, реализована в виде демона, периодически перечитывающего access.log и обновляющего содержимое Redis.

# Заключение

Мы живем на рубеже перехода к информационной эпохе, перехода к внедрению эффективных методов управления сложными и быстропротекающими процессами, изменения методов управления не только технологическими процессами, но также методов управления хозяйственной и социальной деятельностью, что стало возможным благодаря бурному развитию информационных технологий. В результате повсеместной информатизации многие функции управления передаются под контроль сложноорганизованных информационных систем, использующих биологические и компьютерные технологии обработки информации. Обработка информации в подобных системах стала самостоятельным научно-техническим направлением. Для данных систем характерно наличие технологических участков с автоматическим, автоматизированным и интеллектуальным управлением. Построение информационной системы происходит по определенным закономерностям, однако закономерности не всегда могут быть заданы в четком виде, иными словами система обладает поведением, которое определяется закономерностями. Закономерность целесообразности проявляется при любом состоянии системы, благодаря чему могут возникать новые свойства, которые возможно вывести как сумму свойств элементов. Возникает синергетические свойства. Одновременно объединение элементов ведет к возникновению новых свойств, ранее не существовавших, и утрате определенных свойств элементов, то есть происходят сложные качественные изменения, не всегда представляемые и объяснимые.

Поток информации в веб-системе персонализации может быть подвержен существенному количеству ошибок и неуверенностей. Данная неуверенность проникает во все стадии, начиная от веб-образцов навигации пользователя до заключительных рекомендаций. Рекомендательные системы, в свою очередь, представляют собой технологию, стремительно развивающуюся, эффективно использующуюся в различных сферах.

В рамках данной работы рассмотрено описание предметной области и исследованы разного рода рекомендательные системы, определены их особенности, а также предоставлена их классификация в соответствии с алгоритмов фильтрации. Проанализированы алгоритмы фильтрации и сфер их применения.

Практическая часть работы содержит предложение создание конкретного алгоритма выработки рекомендаций. Данное приложение реализуется как readonly REST-сервис. Сервис принимает POST-запросы с user session в json формате и возвращает json со словарем оценок.

# 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, 17 (6), 2005. – Pp. 734-749.
2. Balabanovic M., & Shoham, Y. Fab: Content-based, collaborative recommendation. - Communications of the Association for Computing Machinery, 40 (3), 1997. – Рр. 66-72.
3. Basu C., Hirsh, H., & Cohen W. Recommendation as classification: Using social and content-based information in recommendation. In Proceedings of the fifteenth national conference on artificial intelligence (AAAI-98), Madison, Wisconsin, 1998. - Рp. 714-720.
4. Bell R., Koren Y., & Volinsky C. Matrix factorization techniques for recommender systems. IEEE Computer 42 (8), 2009. – Рр. 30-37.
5. Berendt В. Web usage mining, site semantics, and the support of navigation.
6. Burke R., Mobasher B., Bhaumik R., & Williams C. Segment-based injection attacks against collaborative filtering recommender systems. - In ICDM ’05: Proceedings of the fifth IEEE international conference on data mining, 2005. - Рp. 557 – 580.
7. Cooley R. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.
8. Fleder D., Hosanagar K. [Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=955984) (англ.) // Management Science, Vol. 55, No. 5, May 2009, pp. 697-712 : журнал. — 2009. — Pp. 1 - 49.
9. Harpale A. S., & Yang Y. Personalized active learning for collaborative filtering. In SIGIR ’99: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, Singapore , 2008. – Рр. 439-446.
10. Haveliwala, T. Topic-Sensitive PageRank: A Context- Sensitive Ranking Algorithm for Web Search, IEEE TKDE. 15(4), pp.784-796, 2003.
11. Melville P.,Mooney R., Nagarajan R. [Content-Boosted Collaborative Filtering for Improved Recommendations](http://www.aaai.org/Papers/AAAI/2002/AAAI02-029.pdf) //University of Texas, USA : Материалы конф. / AAAI-02, Austin, TX, USA, 2002. — 2002. — Pp. 187-192.
12. Melville P.,Mooney R., Nagarajan R. [Content-Boosted Collaborative Filtering for Improved Recommendations](http://www.aaai.org/Papers/AAAI/2002/AAAI02-029.pdf) //University of Texas, USA : Материалы конф. / AAAI-02, Austin, TX, USA, 2002. — 2002. — Pp. 187-192.
13. Linden G., Smith B., and York J. [Item-to-Item Collaborative Filtering](http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf) // IEEE Internet Computing, Los Alamitos, CA USA, 2003. — Pp. 76 - 80.
14. Pan R., & Scholz M. Mind the gaps:Weighting the unknown in large-scale one-class collaborative filtering. In 15th ACM SIGKDD conference on knowledge discovery and data mining (KDD), Paris, France.
15. Rennie J., & Srebro N. Fast maximum margin matrix factorization for collaborative prediction. In International conference on machine learning, Bonn, Germany, 2005.
16. Sammut C., Webb J. (Eds.) [Encyclopedia of Machine Learning](http://www.prem-melville.com/publications/recommender-systems-eml2010.pdf). — NY, USA: IBM T. J.Watson Research Center, 2010. — Т. 1. — 1031 с.
17. Su X., & Khoshgoftaar T.M. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009. – Рр. 1-20
18. Su X., Khoshgoftaar T. M., Zhu X., & Greiner R. Imputation-boosted collaborative filtering using machine learning classifiers. In SAC ’08: Proceedings of the ACM symposium on applied computing. - New York: ACM, 2008. - Рр. 949-950.

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, 17 (6), 2005. – Pp. 739. [↑](#footnote-ref-2)
2. Pan R., & Scholz M. Mind the gaps:Weighting the unknown in large-scale one-class collaborative filtering. In 15th ACM SIGKDD conference on knowledge discovery and data mining (KDD), Paris, France. – Р. 213. [↑](#footnote-ref-3)
3. Melville P.,Mooney R., Nagarajan R. [Content-Boosted Collaborative Filtering for Improved Recommendations](http://www.aaai.org/Papers/AAAI/2002/AAAI02-029.pdf) //University of Texas, USA : Материалы конф. / AAAI-02, Austin, TX, USA, 2002. — 2002. — Pp. 187-192. [↑](#footnote-ref-4)
4. Su X., Khoshgoftaar T. M., Zhu X., & Greiner R. Imputation-boosted collaborative filtering using machine learning classifiers. In SAC ’08: Proceedings of the ACM symposium on applied computing. - New York: ACM, 2008. - Р. 949-950. [↑](#footnote-ref-5)
5. Balabanovic M., & Shoham, Y. Fab: Content-based, collaborative recommendation. - Communications of the Association for Computing Machinery, 40 (3), 1997. – Р. 70. [↑](#footnote-ref-6)
6. Cooley R. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000. – Р. 231. [↑](#footnote-ref-7)
7. Bell R., Koren Y., & Volinsky C. Matrix factorization techniques for recommender systems. IEEE Computer 42 (8), 2009. – Р. 34. [↑](#footnote-ref-8)
8. Bell R., Koren Y., & Volinsky C. Matrix factorization techniques for recommender systems. IEEE Computer 42 (8), 2009. – Р. 35. [↑](#footnote-ref-9)
9. Melville P.,Mooney R., Nagarajan R. [Content-Boosted Collaborative Filtering for Improved Recommendations](http://www.aaai.org/Papers/AAAI/2002/AAAI02-029.pdf) //University of Texas, USA : Материалы конф. / AAAI-02, Austin, TX, USA, 2002. — 2002. — Pp. 187-192. [↑](#footnote-ref-10)