Рынок недвижимости г. Иркутска В период с 2010 по 2016 год количество сделок только по жилой недвижимости на вторичном рынке выросло примерно на 20%

Ежегодно в Иркутской области вводится в эксплуатацию не менее 30 000 новых квартир.

На сайте Avito.ru выставлены на продажу более 5000 квартир.

Процесс покупки недвижимости связан с некоторыми проблемами:

Невозможность отсеять «неликвидные» объявления, повторяющиеся, «пустые», и прочее. В свою очередь крайне тяжело отследить на таких торговых площадках некорректно составленные объявления.

Если объект "сильно не похож" на объекты из всех полученных классов, то считать информацию об этом объекте неактуальной.

Для риелторов невозможно

1. отследить многократное повторение одного и того-же объявления.
2. курировать собственные сделки и объекты.
3. составлять клиентскую базу непосредственно на сайте и пр.

В связи с вышеизложенным есть необходимость в создании рекомендательной системы, ориентированной непосредственно на рынок недвижимости.

Рекомендательные системы (РС) [1] -- Информационные систем поддержки принятия решений, предназначенные для оценки уровня интереса \emph{пользователя} к определенному продукту или сервису (\emph{объекту}) на основе имеющейся информации о пользователе и/или объекте. Отрасль разработки РС начала активно развиваться при появлении онлайн-сервисов продаж, и в настоящее время РС -- одно из активных направлений развития систем поддержки принятия решений, ориентированное прежде всего на коммерческое использование, а также на решение задач повышения продуктивности поиска релевантной информации.

В коммерции РС позволяют решать задачи установления, что именно представляет ценность для потребителя в виде набора конкретных объектов (например, товаров или услуг), сужение вариантов выбора и предоставление схожих вариантов других объектов, тем самым упрощая выбор. РС позволяют также выявлять новые характеристики объектов, например, при помощи ведения классификаций объектов и анализа набора известных признаков. Использование РС позволяет отделам снабжения коммерческих фирм-поставщиков предоставлять уникальный сервис каждому потребителю, увеличивая его доверие и лояльность к поставщику, увеличивая продажи и конверсию, а также получая и накапливая больше знаний о потребителях.

Рекомендательные системы появились в интернете достаточно давно, около 20 лет назад. Однако настоящий подъем в этой области случился примерно 5-10 лет назад, когда произошло соревнование NetflixPrize. Компания Netflix тогда давала в прокат не цифровые копии, а рассылала VHS-кассеты и DVD. Для них было очень важно повысить качество рекомендаций. Чем лучше Netflix рекомендует своим пользователям фильмы, тем больше фильмов они берут в прокат. Соответственно, растет и прибыль компании. В 2006 году они запустили соревнование NetflixPrize. Они выложили в открытый доступ собранные данные: около 100 миллионов оценок по пятибалльной шкале с указанием ID проставивших их пользователей. Участники соревнования должны были как можно лучше предугадывать, какую оценку поставит определенному фильму тот или иной пользователь. Качество предсказания измерялось при помощи метрики RMSE (средне-квадратичное отклонение). У Netflix уже был алгоритм, который предсказывал оценки пользователей с качеством 0.9514 по метрике RMSE. Задача была улучшить предсказание хотя бы на 10\% — до 0.8563. Победителю был обещан приз в \$1000000. Соревнование длилось примерно три года. За первый год качество улучшили на 7\%, дальше все немного замедлилось. Но в конце две команды с разницей в 20 минут прислали свои решения, каждое из которых проходило порог в 10\%, качество у них было одинаковое с точностью до четвертого знака. В задаче, над которой множество команд билось три года, все решили каких-то двадцать минут. Опоздавшая команда (как и многие другие, участвовавшие в конкурсе) остались ни с чем, однако сам конкурс очень сильно подстегнул развитие в этой области.

Основными проблемами в области РС являются

отсутствие общего базиса оценивания качества систем (по предметным областям), включая объективную информацию о реальных оценках реальных пользователей, нестабильность методов оценивания и высокая их зависимость от "шума";

неиспользованный потенциал научных исследований: новые научные результаты не внедряются в практические приложения (большинство работающих РС базируются на простых методах), данные существующих практических реализаций РС научно не исследуются, нет тесного взаимодействия со снежными областями анализа данных, а так же друг с другом, низкий научный интерес к РС;

в оценке удовлетворенности не учитывается факторы конфиденциальность, безопасности данных, разнообразие, разметка и презентация информации; в значимом количестве РС моделирование пользователя было крайне примитивно - набор ключевых слов, собственная статья или просто фрагмент текста, представляющих научные интересы пользователя.

Среди открытых проектов выделяются MyMediaLite, LensKit, Mahout, Duine, RecLabCore, easyrec и Recommender.

Методы фильтрации содержания

В статье [5] решается задача анализа профиля пользователя в социальной сети ВКонтакте для решения проблемы холодного старта в решении задачи рекомендации жанров и произведений музыки и фильмов. Авторами разработана РС «EZSurf» автоматизирующая процесс веб-сёрфинга и фильтрации контента, используя профиль пользователя в социальной сети "ВКонтакте", а также API сервисов last.fm, TheMovieDB для получения сведений о схожих объектах (музыкальных произведений). Такой подход существенно упрощает хранилище данных РС, поскольку не требует создания собственной системы классификаций и базы объектов.

В статье [9] рассматривается задача выделения N объектов с наивысшими оценками интереса, задача top-N, при применении фильтрации контента. Предлагается математическая модель контентной рекомендательной системы, основанная на нечетких множествах, критерий оценки качества рекомендаций и алгоритм решения задачи. Математическая модель и алгоритм протестированы на данных сайта last.fm.

\subsection{Методы коллаборативной фильтрации}

Подходы, основанный на коллаборативной фильтрации, в настоящее время более популярны, чем подходы на основе фильтрации содержимого, вероятно из-за того, что представляет собой отражение практического опыта: большинство коммерческих РС вынуждены решать проблему недостатка информации, "холодный старт", а также адаптируемости существующих сообществ пользователей к новым объектам.

Математические обозначения элементов модели сравнения состоит из набора пользователей и набора объектов . В виде – множество элементов, оцененных пользователем – множество пользователей, которые оценили объект – оценка пользователя u для объекта - вектор всех оценок пользователя – вектор всех оценок объекта и – средние значения оценок пользователя и объекта соответственно. Сравнительная оценка обозначается . Для задания этой оценки сначала задается мера близости объекта к объекту . Рассмотрим несколько популярных вариантов оценки близости.

Коэффициент Пирсона [3]:

где \(U=U\_i\cup U\_j\) - множество пользователей, которые оценили объекты и .

Косинус угла между двумя векторами и :

Затем производится формирование конечного множества объектов наиболее близких к объекту Вычисление рейтинга объекта o делается по формуле:

Популярный подход к формированию множества рекомендаций - это упорядочивавшие всех объектов по критерию схожести и выборке некоторого фиксированного количества объектов с максимальным рейтингом [Нефедова]. В качестве меры similarity двух объектов выступает cos угла между -мерными векторами.

В [3,10] так же представлен обзор способов использования вышеупомянутых методов вычисления оценок, которые разделены на два класса - анамнестические, т.е. основывающиеся на одновременной обработке всех имеющихся данных, и модельные, где производится предварительная обработка данных, выполняемая, например, раз в сутки. Второй класс позволяет быстрее вычислять оценки интереса, однако не обеспечивает актуальности данных. В классе аналитических способов, как правило используются методы многомерного анализа данных на основе "ближайшего соседства" (Neighbourhood-based), в то время как в модельных методах используется методы анализа скрытых факторов (LatenetFactors). Существуют гибридные методы, объединяющие оба предыдущих класса.

Гибридные методы

В статье [11] рассмотрена задача разработки алгоритмов оценивания лекционного материала. Авторами предложен алгоритм вычисления близости лекций (объектов), где каждая лекция характеризуется подмножеством некоторого набора значений (например, подмножеством авторов лекций относительно множества всех авторов). Базовый алгоритм реализует подход фильтрации содержания. Для алгоритма подобранны коэффициенты, при помощи которых можно объединять оценки различных атрибутов в одну общую оценку лекции. Наиболее значимыми атрибутами оказались «категории», «авторы», «языки», «название» и «описание». Цель - синтезировать набор лекций фиксированной длины, рекомендованных для просмотра заданному пользователю, из фиксированного множества "новых", не использованных в построении профилей пользователя и объекта.

Далее алгоритм дополняется предсказателем последовательностей лекций: заданы примеры последовательностей из трех лекций, требуется для последовательностей из двух предложить третью, четвертую и т.д. Последовательности лекций приобретены системой неявно, т.е. фиксируя просмотренные пользователем лекции. Алгоритм занял первое место в соревновании, причем со значительным отрывом от второго места. Производится внедрение результатов исследований в области анализа сигналов и предсказания последовательностей событий.

Выделены основные проблемы, требующие решения как на этапе разработки РС, так и на этапе их эксплуатации:

пользователи неохотно предоставляют информацию о себе и своих потребностях, либо разработчики РС уделяют мало внимания процессу информационного наполнения профиля пользователя;

в предметных областях, связанных с большой стоимостью объекта или услуги (где принимается серьезные решения по вложения материальных средств), информационные модели объекта и профиля пользователя сложны по своей структуре и связи компонент структуры, что требует явного представления концептуальной модели предметной области во время выполнения РС как своих основных функций, так и функций предсказания значений атрибутов объекта или профиля пользователя на основе прецедентов;

для предыдущего пункта важным является также разработка пользовательского интерфейса, позволяющего в удобной для пользователя форме и достаточно гибко задавать запросы к РС, а также визуализировать результаты, предлагаемые РС.