Обоснование магистерской работы

В настоящее время существует много алгоритмов, которые связывают между собой объекты и пользователей и ищут для пользователей объекты, наиболее их интересующие.

Среди них можно выделить два класса:

- **Явные оценки (Коллаборативная фильтрация).** Такие алгоритмы хранят оценки пользователей об объектах и строят на этой основе предположения о похожих объектах и/или пользователях.
- **Неявные оценки.** Алгоритмы этого класса строят модели поведения, обучаясь на тестовых данных, а затем работая на реальных (data mining, machine learning).

Алгоритмы из обеих групп обеих групп обдают рядом проблем.

Проблемы алгоритмов коллаборативной фильтрации:

- **Первый рейтинг.** Невозможно или бессмысленно делать прогноз об объектах, которым ни один пользователь не поставил оценки.
- **Избирательность внимания.** Плотность оценок неравномерно распределена, причем эта неравномерность усиливается со временем из-за того, что хорошо оцененные объекты чаще попадаются пользователям.
- **Холодный старт.** Чем больше объектов в системе, тем больше информации требуется от каждого нового пользователя, чтобы выдать первую рекомендацию
- **Изменчивость профиля.** Профиль пользователя составляется из оценок, со временем объем профиля растет, а вкусы меняются. Обновлять оценки для всех старых объектов затруднительно.
- **Пересчет «соседей».** При больших размерах матрицы оценок пользователей для объектов сравнение объектов, а так же пересчет множеств похожих пользователей («соседей») занимает много времени.

Проблемы алгоритмов, не учитывающих оценки:

- **Специфика объектов.** Алгоритмы использующие содержание объектов непереносимы на другие объекты. Например, ключевые слова новостей не применимы к рекомендациям фильмов.

Для решения большинства из этих проблем, предлагается выделить из объектов единицу смысла и использовать новое представление данных, а вместе с ним и новый алгоритм для рекомендаций.

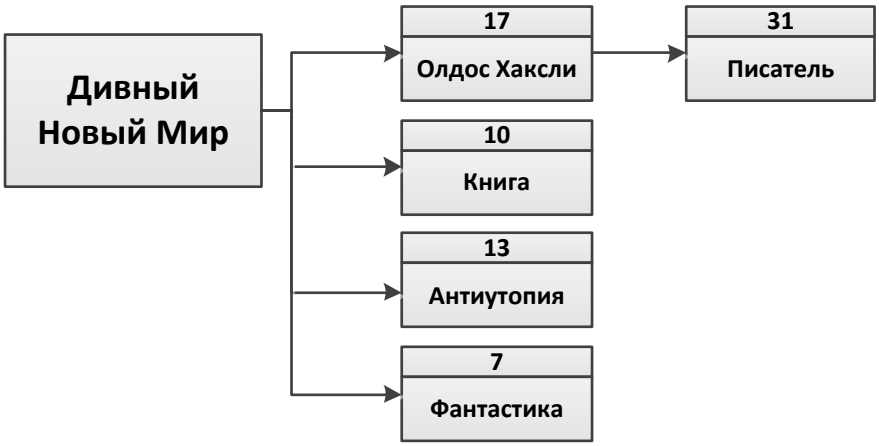
Предлагается отказаться от понятия рейтингов или оценок, и заменить их семантическими связями между объектами *«состоит из»* (в смысле сущности объектов, а не физического устройства вещей).

Каждый объект в системе рекомендаций предлагается представлять как цельную *единицу информации*.

Единицы информации

Каждую *единицу информации* предлагается представлять в виде уникального ключа (id), и множества взвешенных ссылок на другие единицы информации.

Например, *единица информации*, соответствующая книге Олдоса Хаксли «Дивный новый мир», будет содержать ссылки на «Олдос Хаксли», «Книга», «Фантастика» и «Антиутопия». В свою очередь, *единица информации* «Олдос Хаксли» будет ссылаться на «Писатель».



Сравнение объектов осуществляется путем сравнения доли весов общих ссылок от общего веса ссылок. Профиль пользователя предлагается представлять так же *единицей информации*.

При взаимодействии пользователя с объектом (добавление объектов в свой профиль или их оценка) предлагается:

- изменять веса ссылок на одинаковые *единицы информации* у пользователя и у объекта при помощи функции *выравнивания*
- добавлять новые ссылки, как пользователю, так и объекту при помощи *диффузии*

Подобное сравнение и обновления содержания объектов позволяет решить **проблему избыточного внимания** и **проблему долгих пересчетов**, потому как для сравнения двух объектов не требуется обращения к третьим.

Автоматический сбор единиц информации и *диффузия* решает **проблемы холодного старта** и **первого рейтинга**.

Дав возможность пользователю управлять содержимым своего профиля можно решить **проблему изменчивости**, при этом не требуется ни больших усилий от пользователя, ни дополнительных расчетов для сравнения нового профиля и других объектов.

Используя один формат данных для разных рекомендательных систем, можно объединять результаты их работы. Таким образом, профиль пользователя в одной системе позволит рекомендовать ему объекты другой рекомендательной системы. Это решает **проблему специфики объектов**.

Задачи

- Найти оптимальные функции *диффузии* и *выравнивания*
- Опробовать метод на открытой базе фильмов IMDB (http://imdb.com/interface)
- Предложить алгоритм быстрого поиска «похожих» объектов
- Сравнение с существующим алгоритмом Slope One по принципу конкурса Netflix

Научная Новизна

Научная новизна заключается в создании нового представления данных для рекомендательных систем, которое позволяет решить ключевые проблемы существующих рекомендательных систем.

Практический смысл

Такое представления данных позволяет объединить результаты многих систем, что позволит пользователю использовать один профиль во многих системах.

Примечания

Обозначения

Большими латинскими далее обозначаются объекты, маленькими коэффициенты и метрики.

Вес связи от объекта O_1 к объекту O_2 будем обозначать $w(O_1, O_2)$.

Функции выравнивания

Допустим, имеются два объекта A и B причем:

A связан с множеством объектов $S_A = \{O_{a_1}, O_{a_2}, \dots, O_{a_n}\}$

B связан с множеством объектов $S_B = \{O_{b_1}, O_{b_2}, \dots, O_{b_p}\}$

Тогда метрика «похожести» будет выглядеть следующим образом:

$$m(A,B) = l(A,B)\frac{\sum_i^{O_i \in S_a,\ O_i \in S_b}(k(A)W(A,O_i) + k(B)W(B,O_i))}{\sum_i^{O_i \in S_a}(W(A,O_i))\ \sum_i^{O_i \in S_a}(W(B,O_i))}$$

Где, $k(O)$, $l(A,B)$ — функции выравнивания.

Функции диффузии

Допустим, имеются два объекта A и B причем:

A связан с множеством объектов $S_A = \{O_{a_1}, O_{a_2}, \dots, O_{a_n}\}$

B связан с множеством объектов $S_B = \{O_{b_1}, O_{b_2}, \dots, O_{b_n}\}$

При добавлении связи $A \to B$, к A добавляются связи к объектам O_j так, что:

1. $O_j \notin S_a,\ O_j \in S_b$
2. $w(A, O_j) = m(A,B)w(G, O_j)d(A,B,O_j)$

Где $d(A,B,O_j)$ — функция диффузии.