Рекомендательные системы и поиск закономерностей в данных

Рекомендательные системы

1. Неперсонализованные
   - Используя агрегированные данные (например, средний рейтинг)
2. Content-based
   - Пользователю рекомендуются объекты, похожие на те, с которыми пользователь уже взаимодействовал
   - Похожие оцениваются по признакам содержимого объектов
   - Сильная зависимость от предметной области, полезность рекомендаций ограничена
3. Коллаборативная фильтрация
   - Для рекомендаций используется история оценок как самого пользователя, так и других пользователей
   - Более универсальный подход, часто дает лучший результат
   - Есть свои проблемы (например, холодный старт)

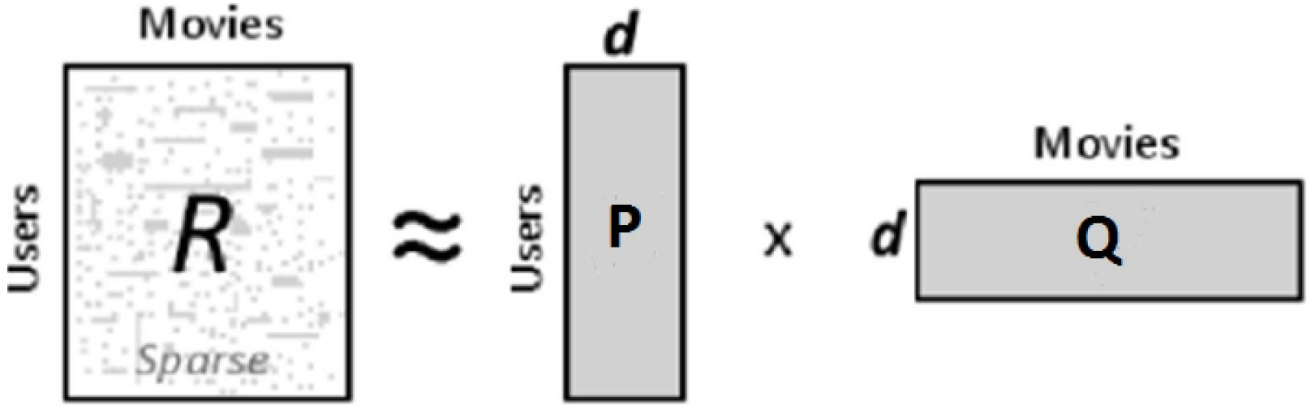
Item-based, User-based

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{j \in I_u} sim(i,j)(r_{uj} - \bar{r}_j)}{\sum_{j \in I_u} sim(i,j)}$$

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in U_i} sim(u,v)(r_{vi} - \bar{r}_v)}{\sum_{v \in U_i} sim(u,v)}$$

Matrix Factorization

ALS



$$\hat{r}_{ui} = \langle p_u, q_i \rangle$$

Метрики Общие

- Если модель предсказывает рейтинги
  - MSE, RMSE, MAE …
- Предсказание события (клик, просмотр)
  - Precision, Recall, F-мера, ROC-AUC
- Обычно пользователь видит только k товаров
  - hitrate@k, precision@k, recall@k

Метрики Специфичные для Рекомендаций

- Покрытие товаров
- Покрытие пользователей
- Новизна
- Прозорливость(serendipity)
- Разнообразие

Покрытие

Какая доля товаров в принципе рекомендуется пользователям.

- Какая доля товаров хоть раз попала в рекомендации
- Можно оценивать разнообразие рекомендаций, используя энтропию

$H(p) = -\sum_{i \in I} p(i) \log p(i)$, где $p(i)$ - доля показа товара i среди всех показов для данной рек. системы

Покрытие пользователей

Имеет смысл следить за долей пользователей, для которых не рекомендуется ни один товар (это может произойти из-за каких-то ограничений модели)

Новизна

Доля новых для пользователя товаров среди рекомендованных.

Новые - те, которые пользователь видит впервые глобально.

- Добавить в интерфейс возможность сообщать о том, что этот товар пользователь уже видел
- Удалить из обучающей выборки часть товаров пользователя (как будто пользователь видел их на других сайтах)

Прозорливость (serendipity)

Способность системы предлагать товары отличающиеся от всех купленных пользователем ранее.

Например, если пользователь читал только книги конкретного автора, то рекомендацию хорошей с точки зрения пользователя книги, но от другого автора мы будем называть прозорливой.

$p(i)$

Разнообразие

Степень сходства товаров внутри одной пачки рекомендаций.

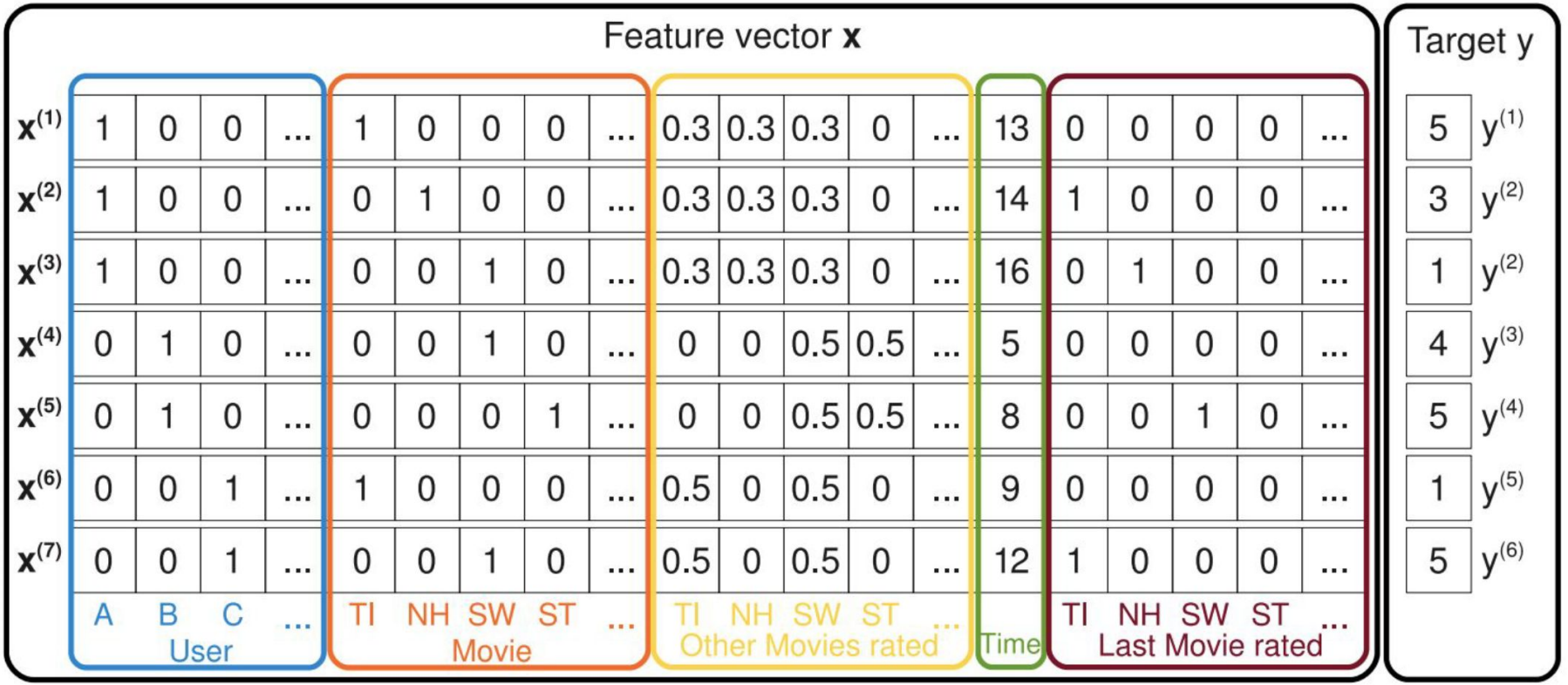
- Например, как среднее попарное расстояние между товарами в одной пачке.

Холодный старт

Используем схожесть контента по его фичам, а не по совместному смотрению

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{j \in I_u} sim(i,j)(r_{uj} - \bar{r}_j)}{\sum_{j \in I_u} sim(i,j)}$$, где I_u - топ N айтемов, наиболее похожих на айтем i

Гибридный подход. Factorization Machines.



Frequent Itemsets and Association Rule Mining

Задача:

Найти ассоциативные правила в наборах множеств. (Например, покупки в супермаркете)

Алгоритмы:

Ищем частые множества. Из частых множеств получение ассоциативных правил уже тривиальный шаг. (Apriori, FP-Growth,)

Sequence Mining

Задача:

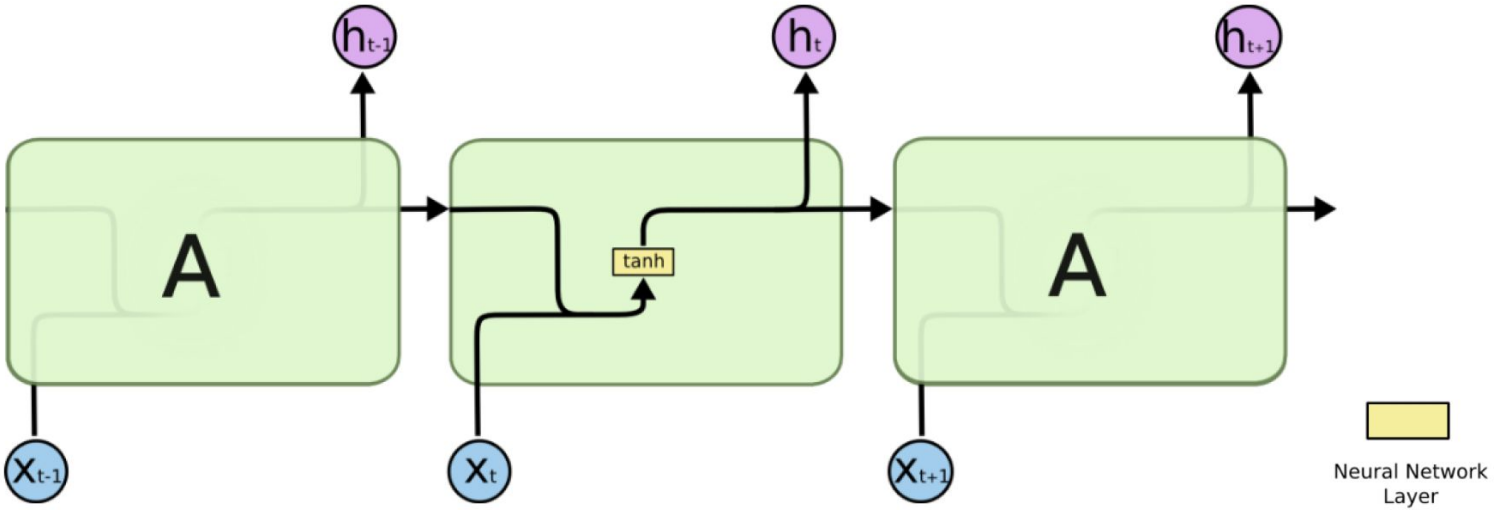
Найти ассоциативные правила в последовательных данных. (Например, последовательность кликов на сайте)

Алгоритмы:

Ищем частые множества. Из частых множеств получаем ассоциативные правила. (GSP, SPADE, PrefixSpan)

RNN

Например, для классификации последовательностей



The repeating module in a standard RNN contains a single layer.