Кислинский Вадим

Постановка задачи:

На базе данных онлайн-магазина и интернет-журнала построить рекомендательную систему, которая выполняет следующие задачи:

1. подбирает наиболее релевантные товары для статьи в журнале;
2. подбирает наиболее релевантные статьи для пользователя в целом;
3. подбирает наиболее релевантные статьи для пользователя в конкретной ситуации;
4. подбирает наиболее релевантные товары для пользователя.

Имеющиеся данные:

- действия пользователей на сайтах: просмотры, покупки товаров, поисковые запросы, просмотры статей;
- описания, названия, характеристики товаров;
- тексты и метаинформация статей.

Метрики качества:

- Для задачи 1:
  - конверсия(кол-во купленных товаров из рекомендаций / кол-во просмотров статьи);
  - NDCG;
- Для задач 2, 3:
  - кол-во просмотров статей;
  - средний чек клиента;
  - bounce rate(процент от общего количества посещений, в рамках которых состоялось не более одного просмотра страницы);
  - NDCG, precision@k, recall@k.
- Для задачи 4:
  - конверсия(кол-во покупок из рекомендаций / кол-во просмотров рекомендаций);
  - NDCG, recall@k, precision@k.

Бейзлайны:

1. Подбор товаров на основе текстовой похожести: для каждого текста статьи и текстового описания товара строятся tf-idf вектора, после чего для каждой статьи подбирается top-N ближайших товаров в построенном вектором пространстве.
2. Рекомендации наиболее популярных статей.
3. Рекомендации наиболее популярных статей.
4. Матричная факторизация, iALS.

Кислинский Вадим

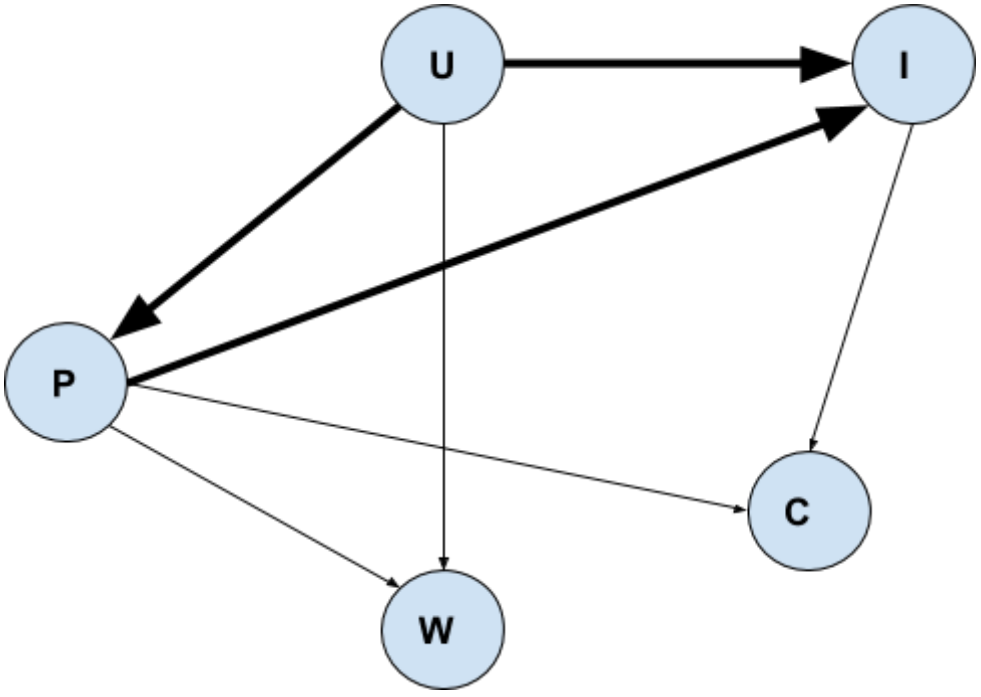
Тематическое моделирование:

С помощью тематического моделирования можно построить систему, которая будет работать, как единое целое. Для решения четырех описанных задач надо научиться предсказывать следующие вероятности: $p(\text{товар} \mid \text{статья})$; $p(\text{статья} \mid \text{пользователь})$; $p(\text{товар} \mid \text{пользователь})$.

Введем обозначения:

- **U** - множество пользователей;
- **I** - множество товаров;
- **P** - множество статей;
- **W** - множество слов(словарь);
- **C** - множество категорий.

Построим граф зависимостей между имеющимися сущностями, связь **A** -> **B** означает, что объекты из сущности **A** могут содержать(получать) объекты из сущности **B**. Жирными выделенны стрелки связей, которые нам наиболее интересны.



Таким образом в терминах тематического моделирования имеется два контейнера документов: пользователи и статьи. “Словами” в этих документах являются товары и слова. Также есть одна модальность - категория.