Matching 和相关技术

计算广告

Elliott

Updated: 2019 年 9 月 28 日

Julyedu.com

开篇

Matching 技术, 是在计算广告领域中很重要的技术. 我们都知道, 在搜索/推荐/广告等场景中, 我们有召回和精排两个过程. 匹配技术, 是在召回中最为重要, 也是最为常用的技术之一.

召回是一个拉取候选集的过程, 往往就是一个匹配问题, 而且很多匹配特征会是排序阶段的重要依据. 再进一步说, 搜索, 推荐, 广告本身其实就是一个匹配问题.

Overview

1. Unified Overview
2. Matching for Searching
3. Matching for Recommendation

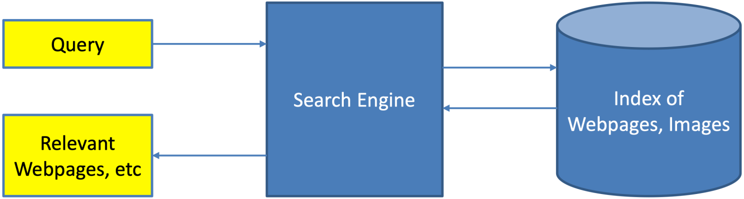
Unified Overview

Search Engine

Information pull: a user pulls information by making a specific request

User intent is explicitly reflected in query:
- Keywords, questions

Content is in
- Webpages, images, …



Key challenge: query-document semantic gap

Query Document Mismatch

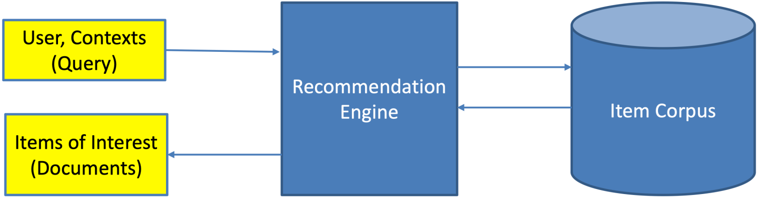
Query	Document	Term matching	Semantic matching
seattle best hotel	seattle best hotels	partial	yes
pool schedule	swimming pool schedule	partial	yes
natural logarithm transformation	logarithm transformation	partial	yes
china kong	china hong kong	partial	no
why are windows so expensive	why are macs so expensive	partial	no

Recommendation Engine

Information push: the system pushes information to a user by guessing the user interest

User Interest is implicitly reflected in:
- Interaction history
- Demographics
- Contexts

Items can be:
- Products, news, movies, videos, friends …

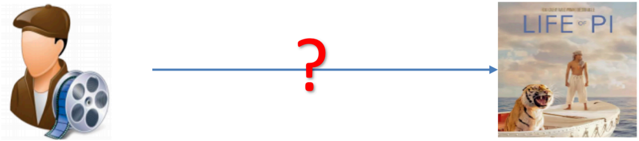


Key challenge: user-item semantic gap

- Even severe than search, since user and item are two **different types of entities** and are represented by different features

Any Overlap?

Movie Recommendation



User Profile (query):
- User ID
- Rating history
- Age, gender
- Income level
- Time of the day

....

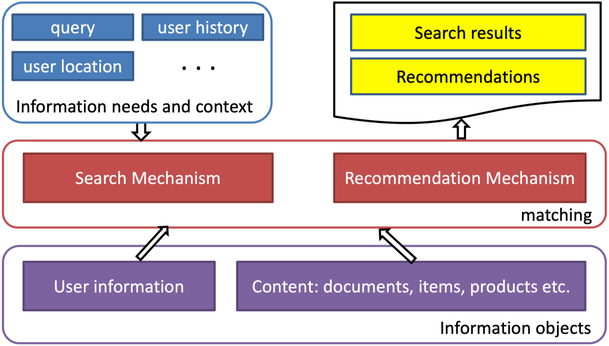
Item Profile (document):
- Item ID
- Description
- Category
- Price
- Image

....

There may be no overlap between user features and item features
Matching cannot be done on the superficial feature level!

Unified View on Matching in Search and Recommendation

- Common goal: matching a context (may or may not include an explicit query) to a collection of information objects (product descriptions, web pages, etc.)
- Difference for search and recommendation: features used for matching!

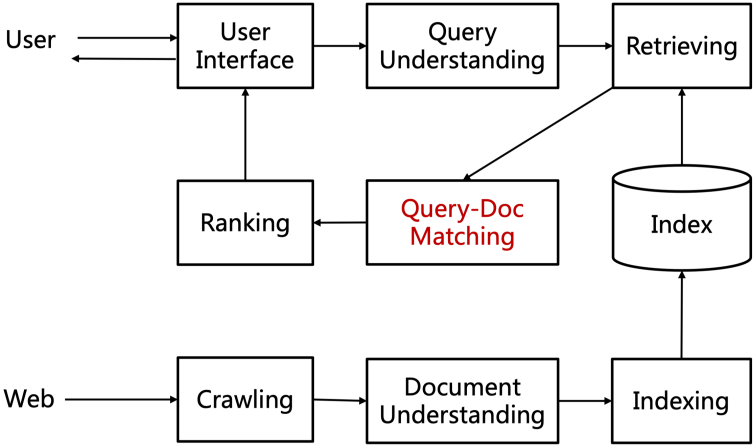


Semantic Gap

- Query-document Mismatch
  - Same intent can be represented by different queries (representations)
  - Search is still mainly based on term level matching
  - Query document mismatch occurs, when searcher and author use different representations
- User-item Semantic Gap
  - Features are used to represent a user and an item may be totally different (e.g., ID feature)
  - Even when they partially overlap in features, it is insensible to conduct direct matching

Matching for Searching

Query-Document Matching



Key Factors

Corpus

- Query: Down the ages noodles and dumplings were famous Chinese food
- Doc: down the ages dumplings and noodles were popular in China

- Semantic Gap: Semantically similar words: famous vs popular, Chinese vs China
- Order of Words: noodles and dumplings, dumplings and noodles

Does Order Really Matter? A survey points out, Over 80% of the potential information in language being in the choice of words without regard to the order in which they appear.

Matching in Term Space

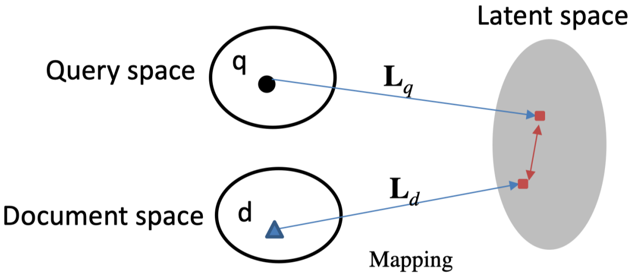
- TF-IDF: word frequency times inverse document frequency
- BM25: some revision of TF-IDF: Not linear

Matching in Latent Space

- Queries and documents have similarities
- Project queries and documents to latent space
- The goal is to make similar items distance to be smaller

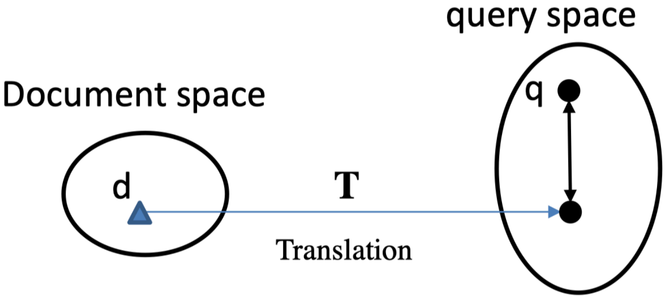
Bridging the Semantic Gap

- Latent space models bridge semantic gap between words through reducing the dimensionality (from term level matching to semantic matching)
- and correlating semantically similar terms



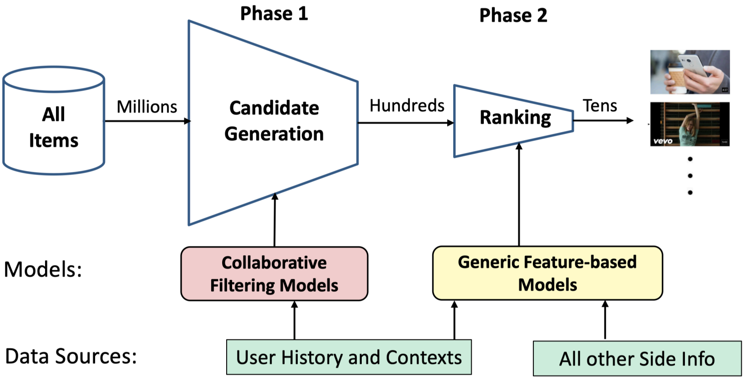
Matching with Translation Model

Given a sentence C in source language, translates it into sentence E in target language



Matching for Recommendation

Recommendation Engine Architecture



Collaborative Filtering

- User-Based CF
- Item-Based CF
- Model-Based CF

Matrix Formulation

User	Movie	Rating
Alice	Titanic	5
Alice	Notting Hill	3
Alice	Star Wars	1
Bob	Star Wars	4
Bob	Star Trek	5
Charlie	Titanic	1
Charlie	Star Wars	5
. . .	. . .	. . .

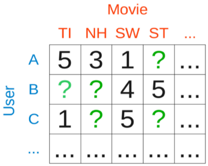
Input Tabular data

Movie

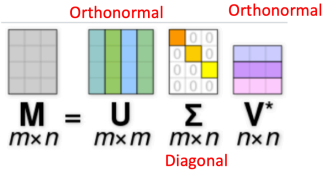
User	TI	NH	SW	ST	...
A	5	3	1	?	...
B	?	?	4	5	...
C	1	?	5	?	...
...	...	...	...	...	...

Rating Matrix
(Interaction Matrix)

Singular Value Decomposition (SVD)



Rating Matrix



Steps to use SVD for CF:

1. Impute missing data to 0 in **Y**
2. Solving the SVD problem
3. Using only *K* dimensions in **U** and **V** to obtain a low rank model to estimate **Y**

Singular Value Decomposition (SVD)

Solve SVD

$$\arg\min(Y - U\Sigma V^T)^2$$

Some Weakness:

- Same Weight for implicit and explicit values
- No regularization is enforced –easy to overfit

Matrix Factorization

- Model: $\hat{y}_{ui} = v_u^T v_i$
- Loss Function:
  $L = \sum_u \sum_i w_{ui}(y_{ui} - \hat{y}_u i)^2 + \lambda(\sum_u ||v_u||^2 + \sum_i ||v_i||^2)$

- ID Embedding
- Inner Product
- Multiple loss function available, and multiple regularization available

Factored Item Similarity Model

Use rated items of the user to represent him.

$$\hat{y}_{ui} = (\sum_{j \in R_u} q_j)^T v_i$$

In fact, it is another form of MF, although it is classified as item-based CF.

SVD++

Combine these two techniques together

$$\hat{y}_{ui} = (v_u + \sum_{j \in R_u} q_j)^T v_i$$

Feature-based Models

Here we use Factorized Machine:

$$y = w_0 + \beta^T x + x^T W x$$

- If we only use ID, it is SVD
- If we use item ID and user historically rated item id, it is FISM
- If we add user id as well, it is SVD++

Some discrepancy

- People only would like to rate item they like
- Better MSE does not mean better ranking result

From Pointwise to Pairwise

- Higher rating > lower rating
- Observed Ones > unseen ones

THANK
YOU