Semantic Matching in App Search

Juchao Zhuo, Zeqian Huang, Yunfeng Liu, Zhanhui Kang, Xun Cao, Mingzhi Li, Long Jin

WSDM Feb 4, 2015
Outline

✓ Overview
✓ Challenge in App Search
✓ Semantic Matching
  ● Matching with Topic
  ● Matching with Tag
  ● Learning to rank
✓ Applications and Evaluations
✓ Conclusion
Overview

✓ With the rapid growth of smartphones, app market has become a significant mobile internet portal. As an important function in app market, app search gains lots of attentions.

✓ Miss-match is the critical challenge in app search. Semantic matching is a key technology to reduce miss-match.

✓ In this talk, we will describe a semantic matching platform, which mines topics and tags in big data to enrich query and app representations, and implements learning to rank.

✓ The semantic matching platform is used by “Myapp” app market, one of the top three android app markets in China.
App Search Application in Tencent

✓ MyApp

- An android app market with a peak distribution of over **100 millions** in one day of 2014.
- App search engine contributes more than 40% to mobile app new-downloads.
- Rapid growth: available apps from 0.3 million to 1.2 million within one year.
- Long Tail: apps which were downloaded over a million times accounted for less than 0.1%.
Objectives of App Search

- **App search objectives**
  - Facilitate the app market, users and developers

- **App search metrics**
  - Downloads / QV(Query views) / UV(User views)
  - CTR(Click-Through-Rate) / ROP(Rate of Penetration)
  - NDCG (Normalized Discounted Cumulative Gain)

Larger and Smarter app distribution
Easier to find app wanted
More downloads
User habits in App Search

✓ Two kinds of queries in app search

<table>
<thead>
<tr>
<th></th>
<th>Ratio of Query Number</th>
<th>Ratio of QV Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precise Search</td>
<td>88%</td>
<td>75%</td>
</tr>
<tr>
<td>Fuzzy Search</td>
<td>12%</td>
<td>25%</td>
</tr>
</tbody>
</table>

✓ Precise Search
● Search by app name, mostly prompted by the search box

✓ Fuzzy Search
● Non-Name, always colloquial expression
● Content/Category/Function related
● User-habit of web search is brought to app search on mobile
● e.g.
  "微信里的游戏" (game in wechat)
  "音乐软件" (music application)
  "报时间的软件" (application that reminds time)
Challenge in App Search

✓ Miss-match

● Long tail challenge is more obvious in app search than that in web search.
● **Data shortage:** app data is much less than web data
● **Text shortage:** app name and desc. is the only text annotation for app
● Miss-match occurs when user and developer use different terms to describe the same semantic, traditional term matching can not fix it.

● e.g.
Semantic Matching Methodology

✓ How to describe "Semantic "?
  • Term + Topic + Tag
✓ Enrich query and app representations by topic and tag
✓ Perform query-app matching with the representations
✓ Hybrid Ranking Model: LTR
Matching with Topic

✓ Topic Model
  • Using Layered LDA (Latent Dirichlet Allocation) model
  • MPI based parallel computing framework
  • Topic probability distribution over term space: $P(\text{word}|\text{topic})$

• Assign million apps to 1000+ topics
• Doc probability distribution over topic space: $P(\text{topic}|\text{doc})$
Matching with Topic

✓ Query inference with topic
  • Each query is regarded as a document
  • **Challenge**: short text has not enough information to inference
  • **Solution**: Expanding query with collection of click apps

✓ Topic matching
  • Map query and documents in the topic space
  • Query $V_q$ and documents $V_d$ are both represented with probability distributions over topics
  • Calculate topic match score between $V_q$ and $V_d$
Matching with Tag

✓ Limitation of Topic Matching
  ● Text corpora of app documents is not large enough to support large topic number.
  ● Significant difference may still exist between apps in the same topic.
  ● Long tail queries lack statistical click data to expand, even after topic matching, many tail queries are still unknown.

✓ Matching with Tag
  ● More fine-grained semantics can be described by tag.
  ● Most app stores assemble tags on human editorial curation.
  ● Our system can monitor the app ecosystem in real time, and automatically extract tags and assign them to apps from multi-source data.
Tag Mining

✓ Multi-source Data
- reading the web data
- web-based question-answering
- user grouping from behavior logs
Tag extraction from web data

✓ Data from web
  ● Structured data

  ● Unstructured text
    ● using template to extraction
Tag from question-answering

- Using web-based question-answering to perform completion of missing tag-app pairs.
Tag from user behavior logs

- Users Profile
  - gender, age, location, ...
- Users Behavior
  - search, download, install, ...

- Based on user-behavior data and natural representation by tag
Tag Denoising

- Using machine learning technology to calculate confidence

- Multi-source tag data
  - Denoising process
  - Filtered tag data

- GBDT (Gradient Boosting Decision Tree) Model

- Features:
  - frequency
  - authority
  - pattern score

- Features:
  - LDA
  - frequency

- Features:
  - users ratio
  - group score
Tag statistic

☑ Denoising data
  ● Over 97% pairs filtered: Treasures are hidden among the sands

☑ Statistic data
  ● Over 90K tags mined
  ● Covers 83% of apps
  ● Top 100K apps have 8.53 tags in average

<table>
<thead>
<tr>
<th>Sources</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Data</td>
<td>81626</td>
</tr>
<tr>
<td>Question-answering Data</td>
<td>8620</td>
</tr>
<tr>
<td>User Group</td>
<td>1218</td>
</tr>
<tr>
<td>Total</td>
<td>91464</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Denoising</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original tag-app pairs</td>
</tr>
<tr>
<td>Filtered tag-app pairs</td>
</tr>
<tr>
<td>Valid tag-app pairs</td>
</tr>
</tbody>
</table>
Matching query with tag

- Using template to map query to tags
  - \( \text{query} \)
  - Template: \( \text{Tag} \cdot \text{Category} \cdot + \text{suffix} \)
  - Template Set
  - \( \text{T:单机+游戏} \)

- Using click data to calculate confidence of template
  - \( P(\text{Template}) = \frac{2}{3} \)

- \( \text{Q:单机游戏哪种好} \)
  - Click Data
  - APP:斗地主
  - APP:找你妹
  - APP:QQ游戏
  - “游戏”
  - TAG retrieved
  - T:单机+游戏
  - TAG retrieved
  - Term retrieved
Learning to Rank

✓ Challenge of Ranking
  ● Relevance calculated by different matching model are incomparable
  ● The example data is imbalanced (e.g. colloquial query less than normal)
  ● Most of the features are nonlinear

✓ Using LambdaMart to rank
  ● LambdaMART combines MART and LambdaRank to solve the supervised learning problem
  ● Mart(Multiple Additive Regression Trees) is a gradient boosting tree model
  ● Label training data partitioned by query
  ● Maximizing NDCG by learning relevance score through MART
Learning to Rank Application

✓ LambdaMART of Combine Ranking
  ● 50+ different features
  ● 300,000+ pairwise training data
  ● 3000+ test samples

✓ Offline Experiment measurement

<table>
<thead>
<tr>
<th>NDCG of baseline</th>
<th>NDCG of LambdaMART</th>
<th>Improvement vs baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8733</td>
<td>0.9553</td>
<td>9.4%</td>
</tr>
</tbody>
</table>

✓ Online A/B Test measurement
  ● CTR promoted by 6%↑
Semantic Matching Metrics

✓ Online A/B Test measurement
  ● CTR diff. on query samples
  ● 9.7% queries & 26.9% query views
  ● CTR promoted by 6%~8%↑

[Graph showing CTR comparison between Semantic Matching and Term Matching]

CTR
Semantic Matching vs. Term Matching

CTR of Semantic Matching
CTR of Term Matching

Jun Jul Aug Sep Oct Nov Dec
App Search with Semantic Matching

CTR

Search UV

Download Apps
App Indexing

✓ Besides apps, all digital content inside apps can be offered
✓ Using LambdaMart to rank with different digital content
✓ ROP (Rate of Penetration) of App indexing version speed-up **32.3%↑**
Semantic Matching Application Example

✓ Deeper understanding colloquial form of query
Conclusion

✓ From **Term matching** to **Semantic matching**
  ● Richer representation of semantic

✓ Methodology of Enriching information
  ● Use the web search technology to detect the relationship of app data

✓ What is the next direction of mobile search
  ● More input mode: voice, photograph, two-dimensional code
  ● Search engine should become more intelligent

✓ Stay tuned for 2015!
Acknowledgement

✓ Joint work with many brilliant colleagues from several departments of Tencent

✓ Many constructive inputs are from Yue Wu and Xing Yao