



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

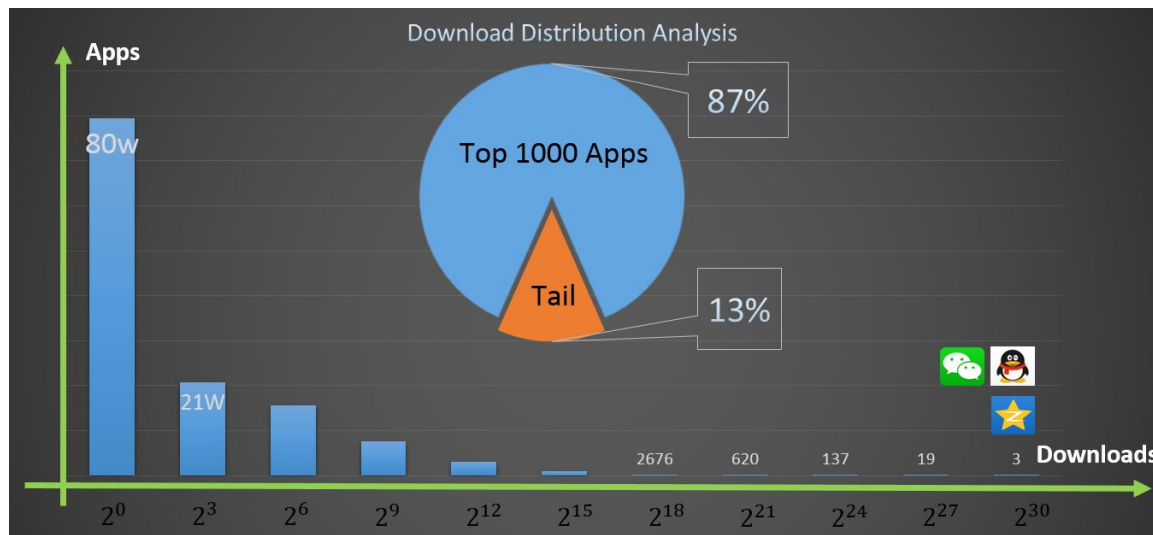
Overview

- ✓ 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

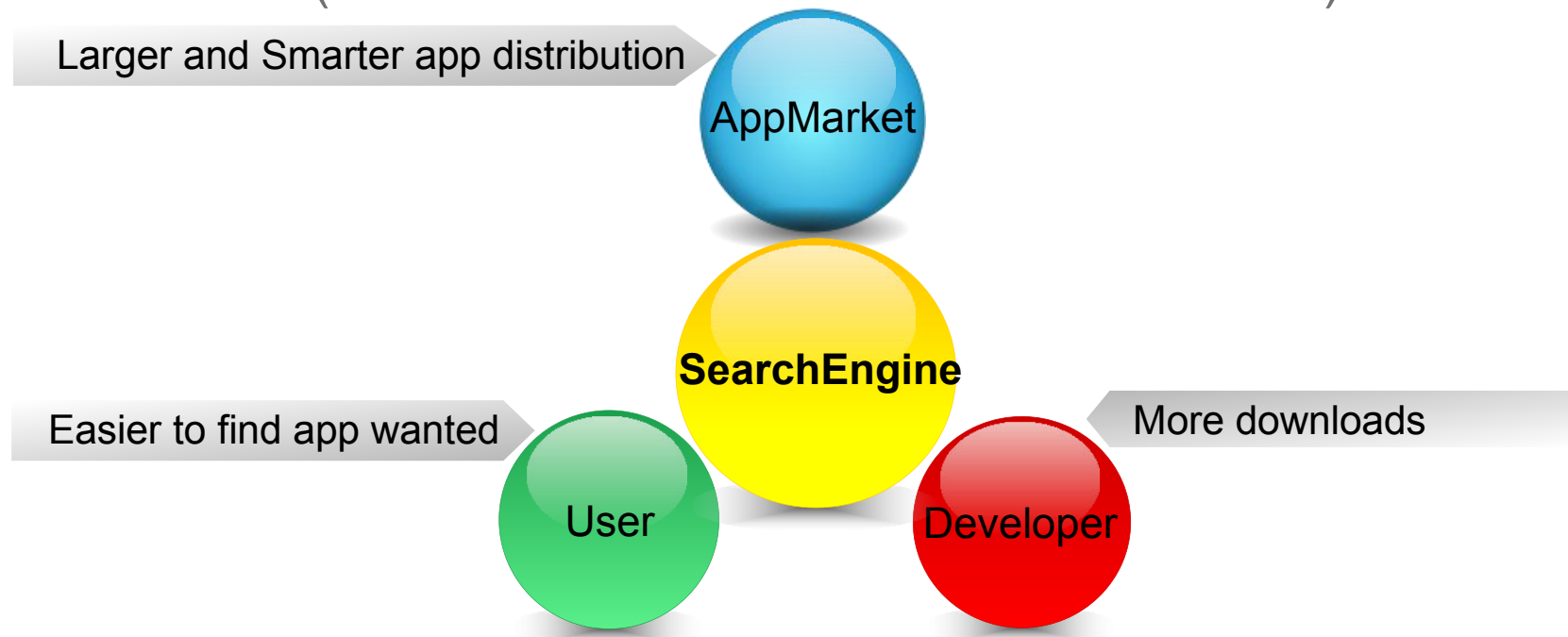
✓ MyApp

- 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)



User habits in App Search

- ✓ Two kinds of queries in app search

	Ratio of Query Number	Ratio of QV Number
Precise Search	88%	75%
Fuzzy Search	12%	25%

- ✓ 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

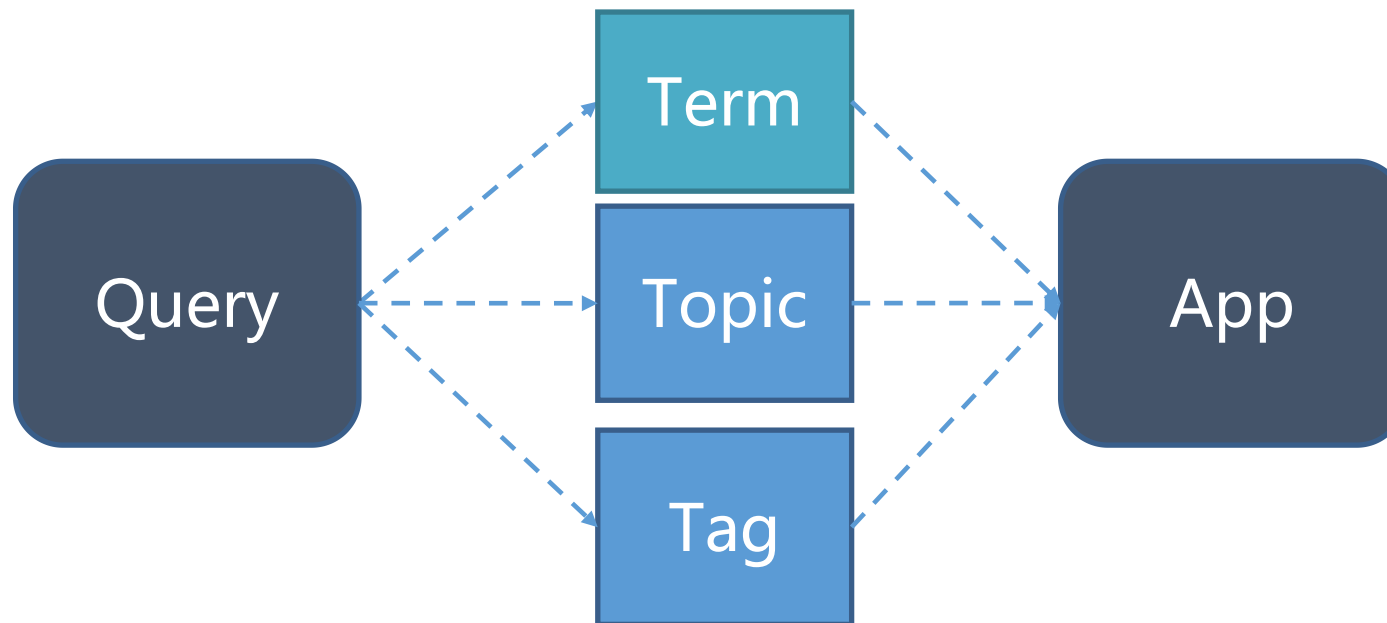
- Long tail challenge is more obvious in app search than that in web search.
- **Data shortage**: app data is much less than web data
- **Text shortage**: app name and desc. is the only text annotation for app
- Miss-match occurs when user and developer use different terms to describe the same semantic, 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: $P(\text{word}|\text{topic})$

405	Topic	4th:
406	音乐	0.132216
407	钢琴	0.028424
408	吉他	0.017029
409	节奏	0.016839
410	流行	0.016373
411	歌曲	0.014719
412	乐器	0.010274
413	摇滚	0.010129
414	美妙	0.009480
415	舞蹈	0.009356
416	享受	0.009028
417	旋律	0.008329
418	乐队	0.008154
419	鸭子	0.007877
420	鳄鱼	0.007746
421	古典	0.007462
422	弹奏	0.007156
423	经典	0.006974
424	模拟	0.006748

1920	Topic	19th:
1921	穿越	0.042110
1922	王爷	0.012890
1923	女子	0.012129
1924	天下	0.011216
1925	时空	0.008626
1926	公主	0.008613
1927	王妃	0.007789
1928	古代	0.007546
1929	皇后	0.007475
1930	后宫	0.006362
1931	美人	0.006362
1932	太子	0.006257
1933	前世	0.005625
1934	千年	0.005508
1935	皇帝	0.005412
1936	美男	0.005068
1937	现代	0.004985
1938	男子	0.004914
1939	桃花	0.004332

3637	Topic	36th:
3638	学习	0.075638
3639	英语	0.073277
3640	单词	0.044185
3641	语言	0.031221
3642	中文	0.025337
3643	英文	0.024916
3644	字母	0.018868
3645	词汇	0.017435
3646	翻译	0.016071
3647	发音	0.012925
3648	词典	0.012054
3649	记忆	0.009291
3650	字典	0.009190
3651	日语	0.009146
3652	口语	0.008899
3653	汉字	0.007143
3654	语音	0.006712
3655	西班牙	0.006567
3656	拼音	0.006306

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



天天酷跑

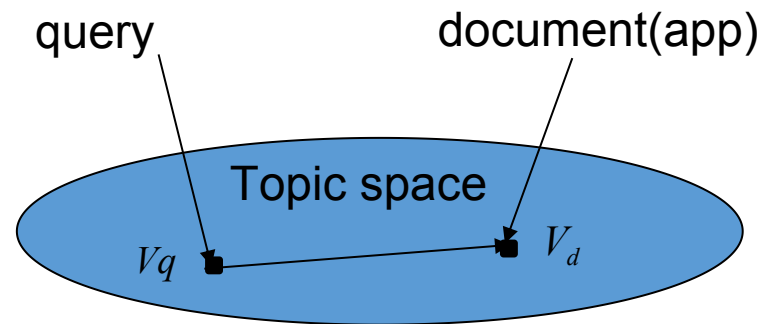
★★★★★ 14个好友在玩

3.3亿人下载 1.0.16.0 腾讯

TopicId	Category	Probability	TopicWords
31	跑酷[游戏]	0.625815	游戏, 跳跃, 障碍, 收集, 跑酷, 躲避, 动作, 不断, 控制, 路上, 奔跑, 道具, 避开, 挑战, 逃亡,
26	关卡[游戏]	0.0649123	游戏, 玩家, 关卡, 挑战, 难度, 益智, 休闲, 玩法, 画面, 道具, 过关, 闯关, 乐趣, 等级, 一共,
70	rpg[游戏]	0.0593985	游戏, 玩家, 系统, 技能, 战斗, 体验, 丰富, 特色, 装备, 角色, 画面, 玩法, 华丽, 全新, 网游,

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.

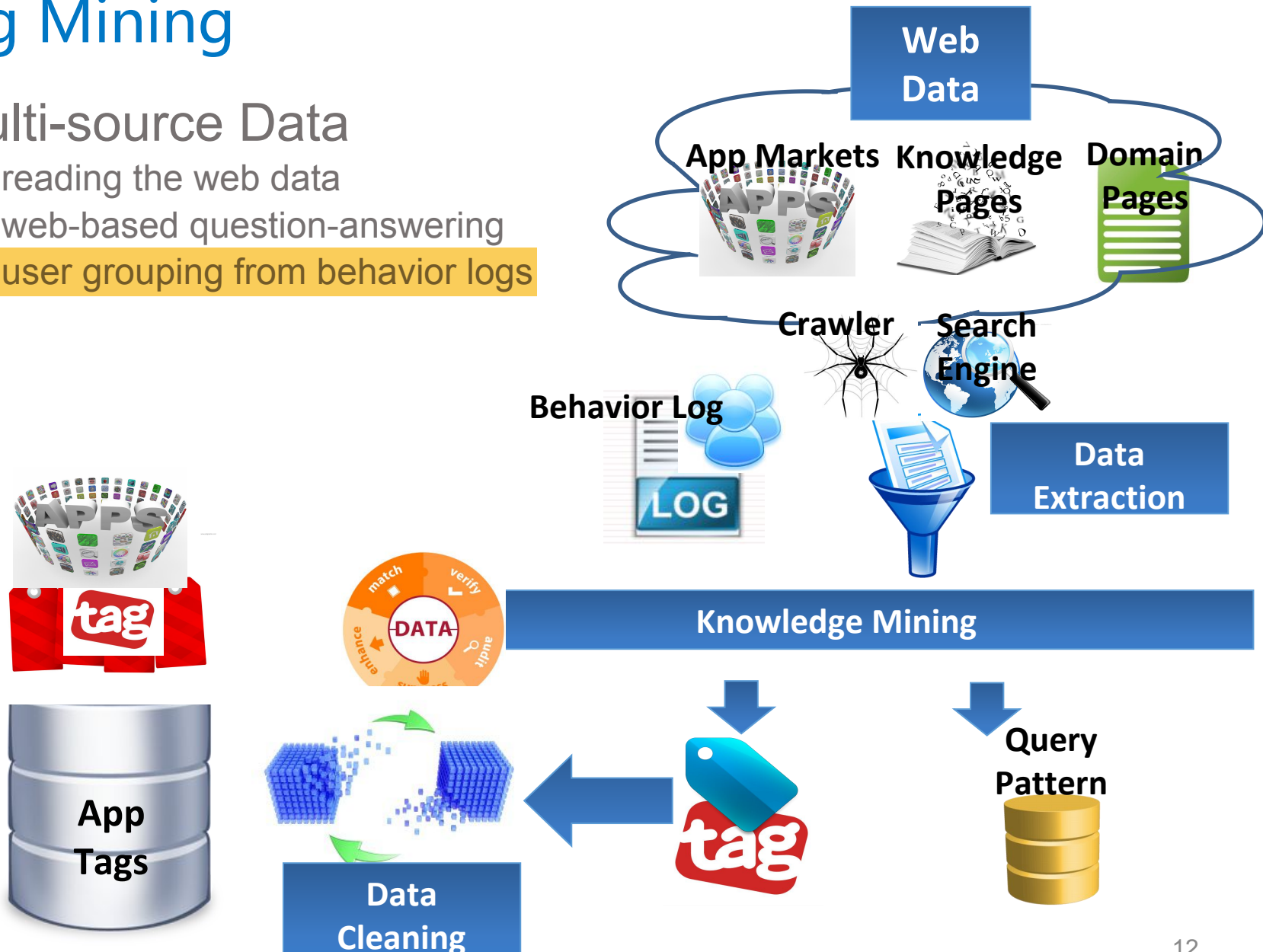
✓ Matching with Tag

- 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



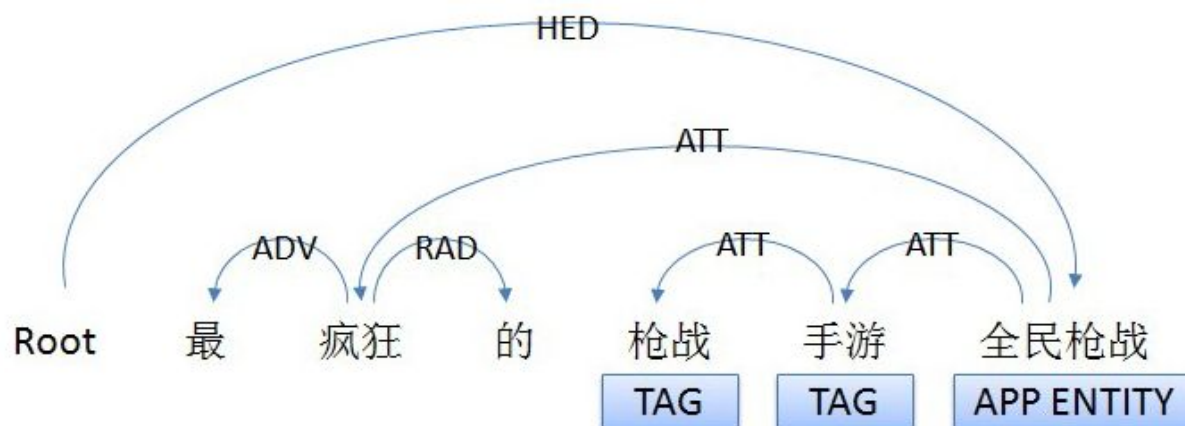
Tag extraction from web data

- ✓ Data from web
 - Structured data

中文名	全民枪战 APP	发行日期	2014年
其他名称	CA	音乐	小旭音乐
游戏类型	射击	内容主题	战争
游戏平台	ios, Android	玩家人数	中型射击手游
发行商	cmge中国手游	游戏下载	4399游戏盒、摸摸、小皮游戏

Diagram illustrating structured data extraction. The word "射击" (Shooting) is circled in blue and has an arrow pointing to a central "TAG" node. The word "战争" (War) is also circled in blue and has an arrow pointing to the same "TAG" node. The word "中型射击手游" (Medium shooting mobile game) is circled in blue and has an arrow pointing to the "TAG" node.

- Unstructured text
 - using template to extraction



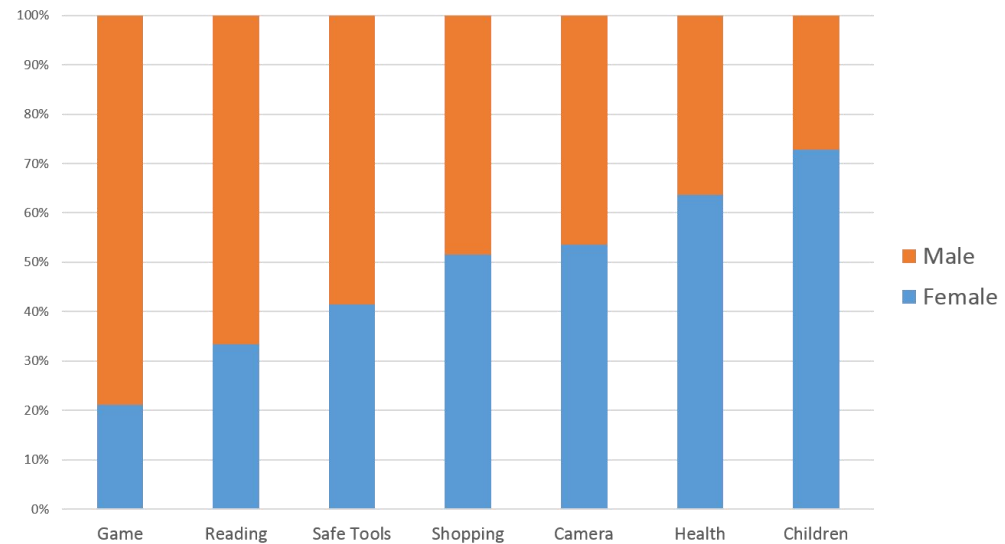
Tag from question-answering

- Using web-based question-answering to perform completion of missing tag-app pairs.

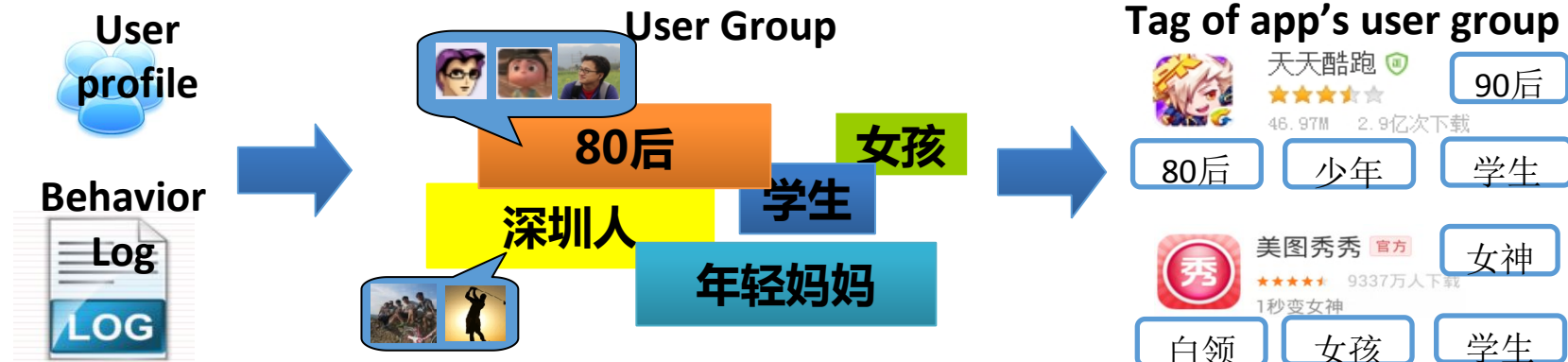
Tag from user behavior logs

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

Download-ratio by gender and category

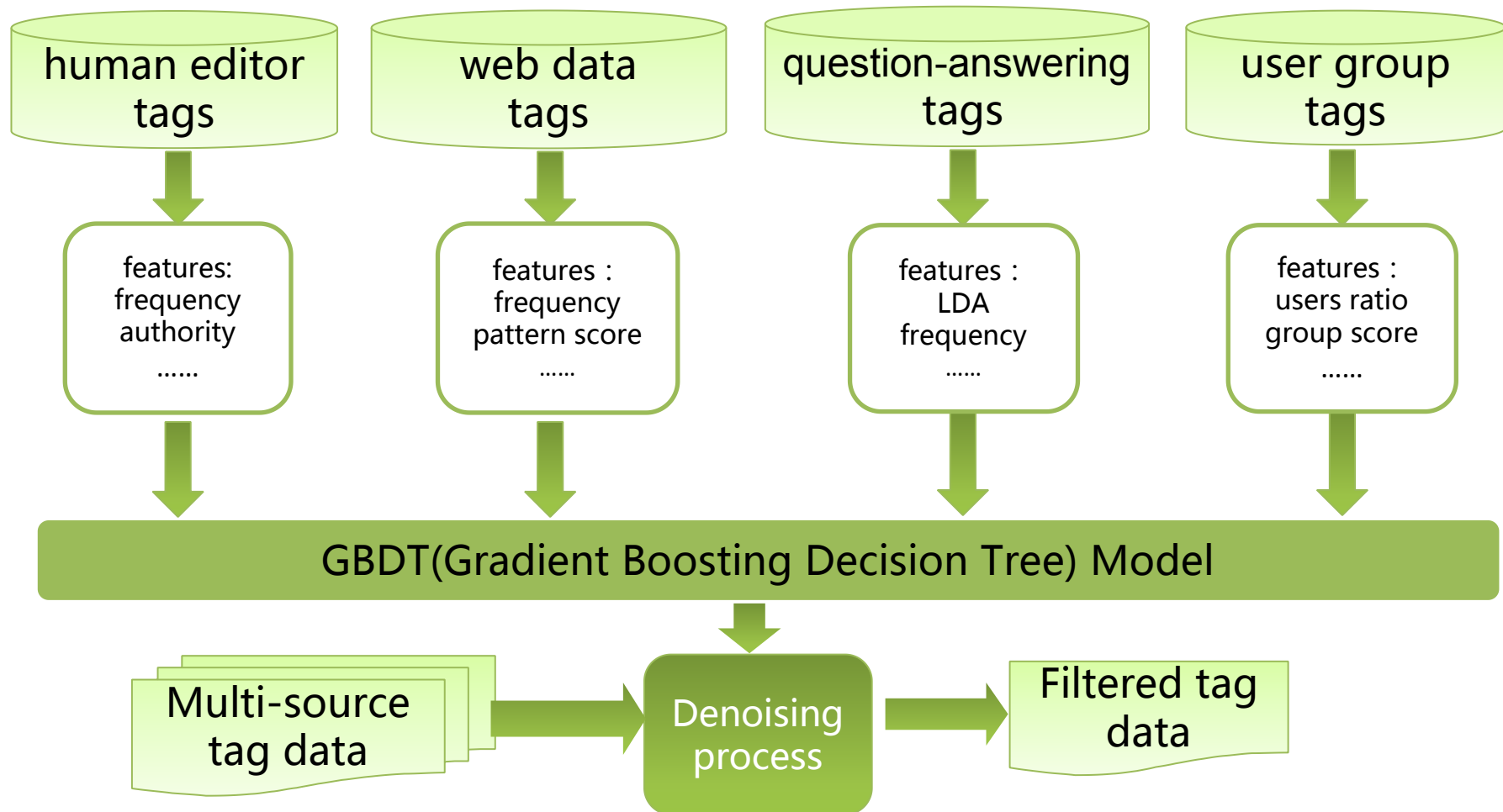


- Based on user-behavior data and natural representation by tag



Tag Denoising

- Using machine learning technology to calculate confidence



Tag statistic

✓ Denoising data

- Over 97% pairs filtered: Treasures are hidden among the sands

Denoising	
Original tag-app pairs	38515130
Filtered tag-app pairs	37677471 (97.83%)
Valid tag-app pairs	837659 (2.17%)

✓ Statistic data

- Over 90K tags mined
- Covers 83% of apps
- Top 100K apps have 8.53 tags in average

Sources	Tags
Web Data	81626
Question-answering Data	8620
User Group	1218
Total	91464

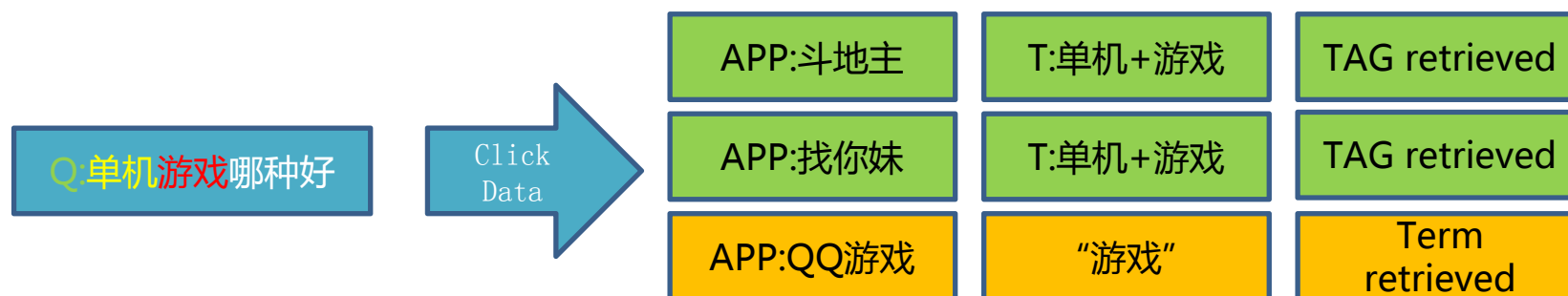
Matching query with tag

- ✓ Using template to map query to tags



- ✓ Using click data to calculate confidence of template

- $P(\text{Template}) = 2/3$



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

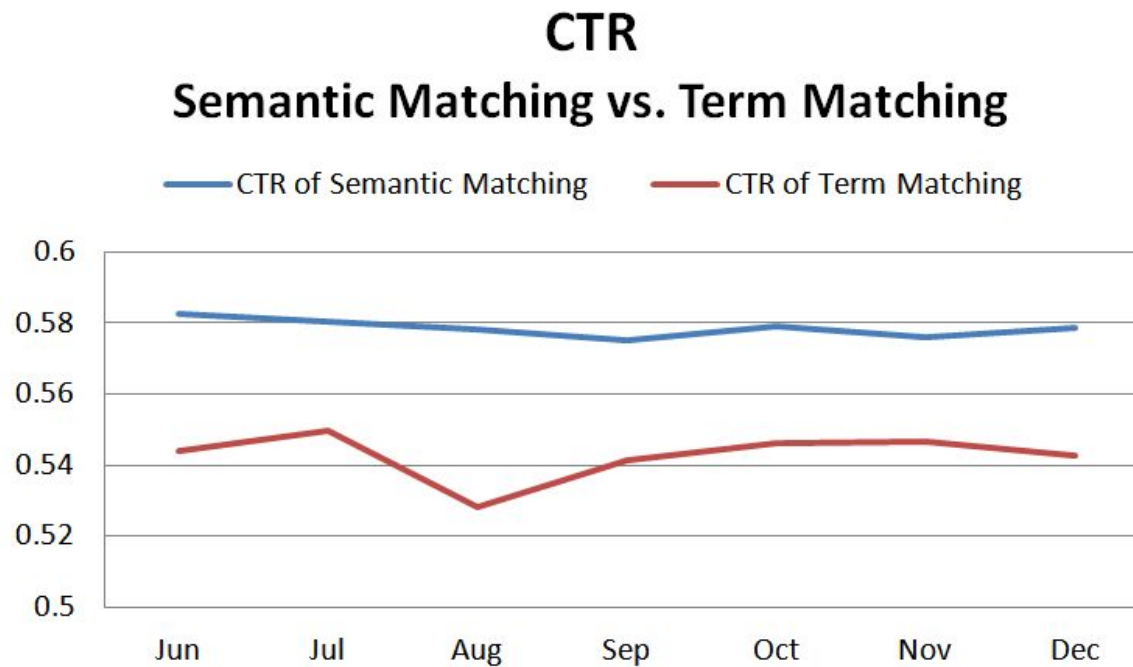
NDCG of baseline	NDCG of LambdaMART	Improvement vs baseline
0.8733	0.9553	9.4%

✓ Online A/B Test measurement

- CTR promoted by 6%↑

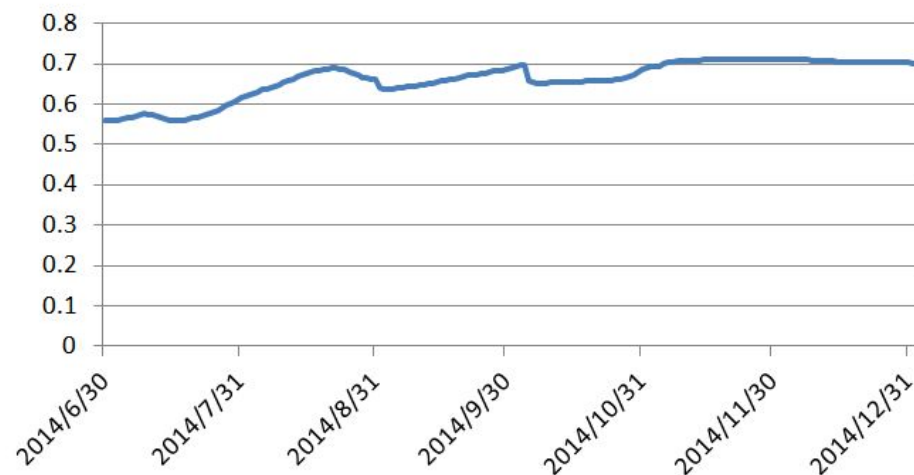
Semantic Matching Metrics

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

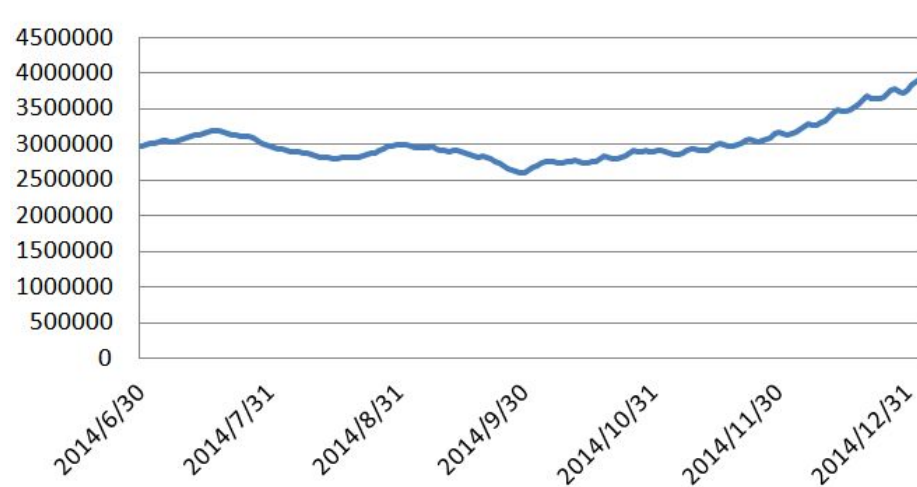


App Search with Semantic Matching

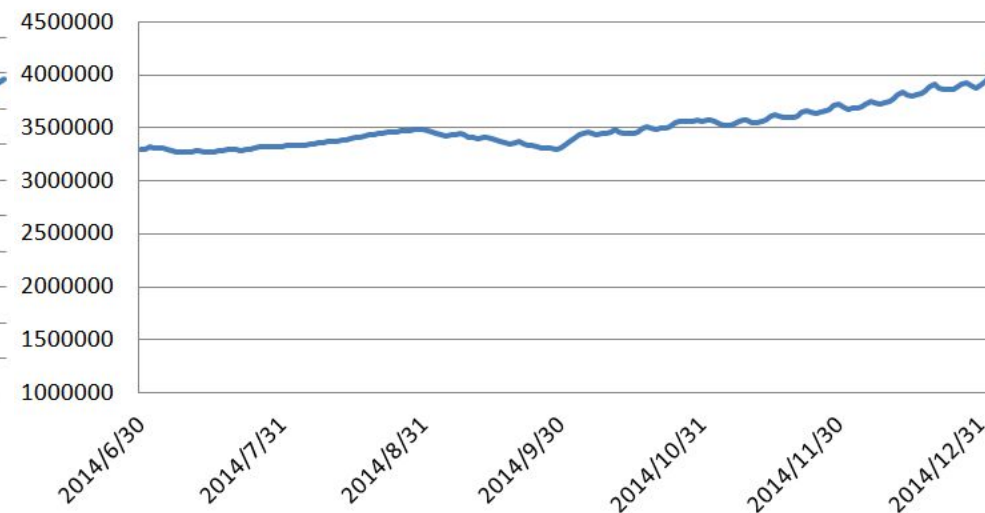
CTR



Search UV

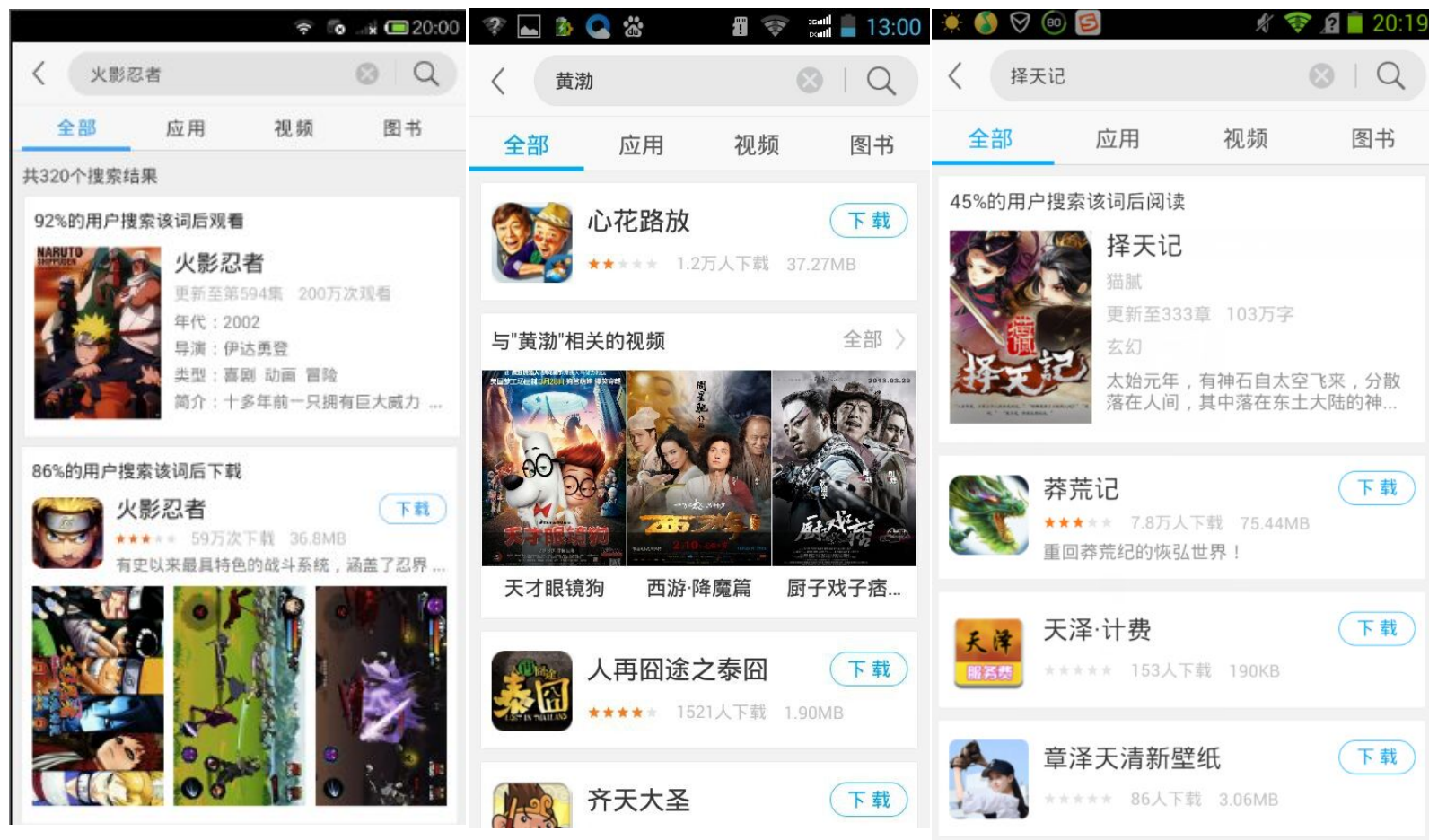


Download Apps



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-dimension 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



内部搜索平台部
Internal Search Platform Dept



腾讯应用宝
就要玩在一起