Semantic Matching in App Search

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

WSDM Feb 4, 2015

Tencent 腾讯

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.

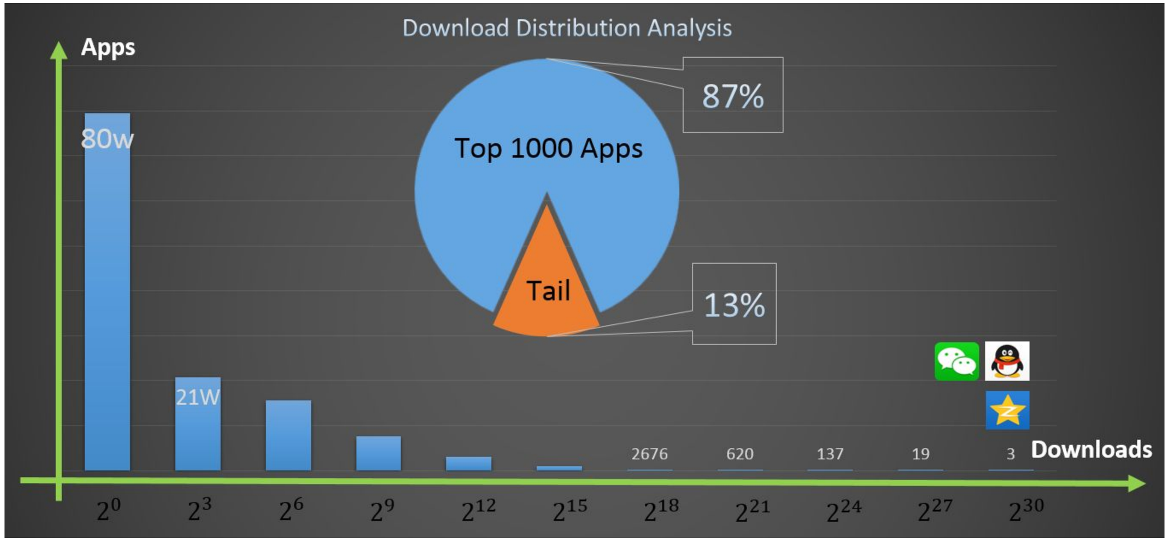
- ✓ In this talk, we will describe a semantic matching platform， which mines topics and tags in big data to enrich query and app representations, and implements learning to rank.

- ✓ 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

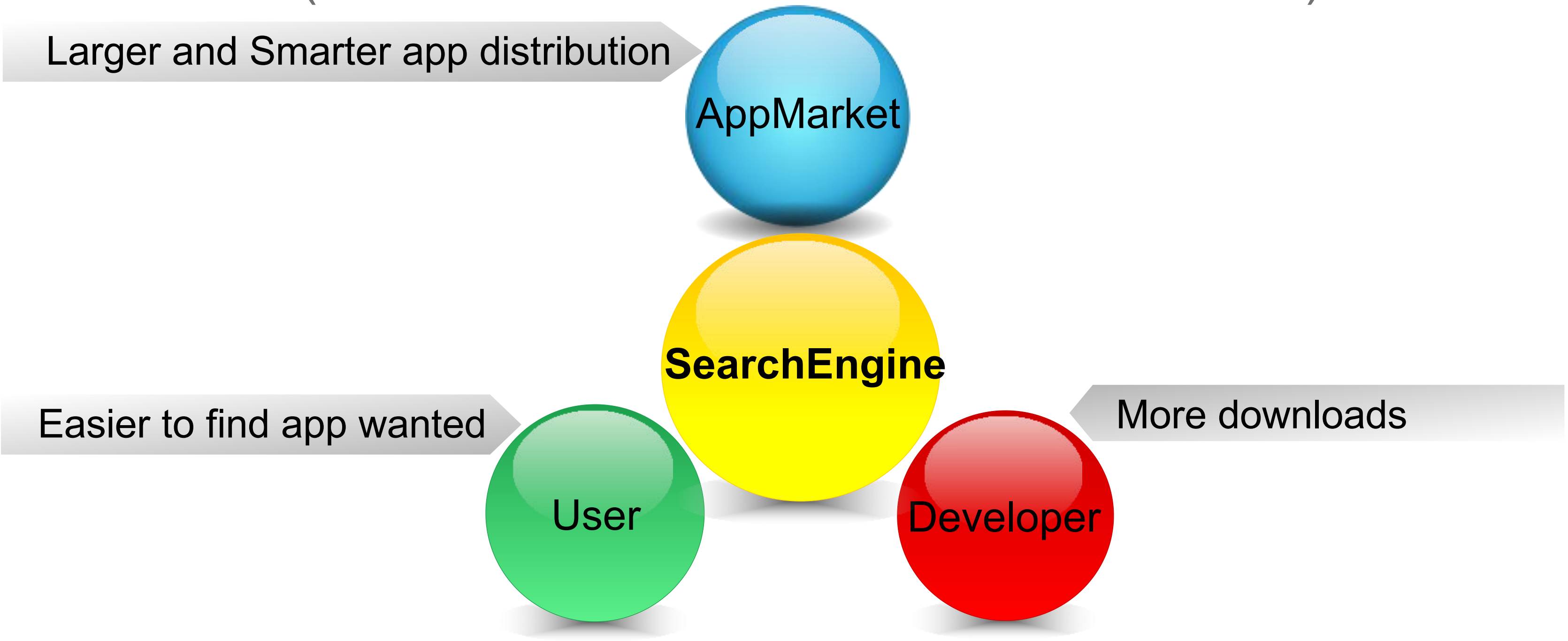
✓ MyApp

- 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
AppMarket
SearchEngine
Easier to find app wanted
More downloads
User
Developer

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

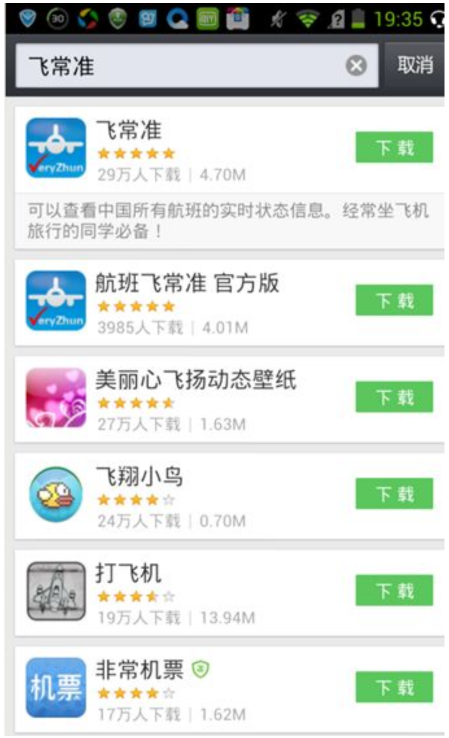
- Non-Name, always colloquial expression
- Content/Category/Function related
- User-habit of web search is brought to app search on mobile
- e.g.

  "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.



飞常准

飞常准
★★★★★
29万人下载 | 4.70M
可以查看中国所有航班的实时状态信息。经常坐飞机旅行的同学必备！

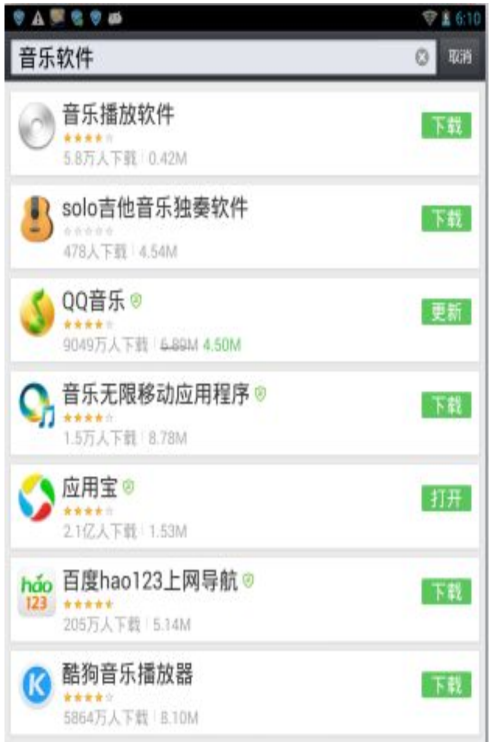
航班飞常准 官方版
★★★★★
3985人下载 | 4.01M

美丽心飞扬动态壁纸
★★★★★
27万人下载 | 1.63M

飞翔小鸟
★★★★☆
24万人下载 | 0.70M

打飞机
★★★★☆
19万人下载 | 13.94M

非常机票
★★★★☆
17万人下载 | 1.62M



音乐软件

音乐播放软件
★★★★☆
5.8万人下载 | 0.42M

solo吉他音乐独奏软件
☆☆☆☆☆
478人下载 | 4.54M

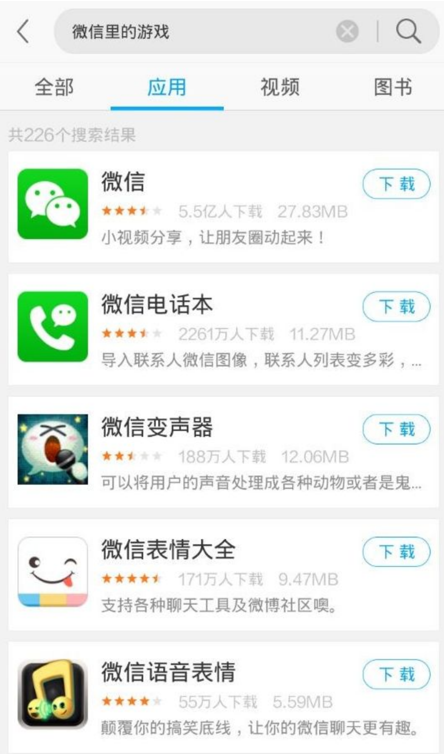
QQ音乐
★★★★☆
9049万人下载 | ~~5.89M~~ 4.50M

音乐无限移动应用程序
★★★★☆
1.5万人下载 | 8.78M

应用宝
★★★★☆
2.1亿人下载 | 1.53M

百度hao123上网导航
★★★★★
205万人下载 | 5.14M

酷狗音乐播放器
★★★★☆
5864万人下载 | 8.10M



微信里的游戏

共226个搜索结果

微信
★★★★☆ 5.5亿人下载 27.83MB
小视频分享，让朋友圈动起来！

微信电话本
★★★★☆ 2261万人下载 11.27MB
导入联系人微信图像，联系人列表变多彩，...

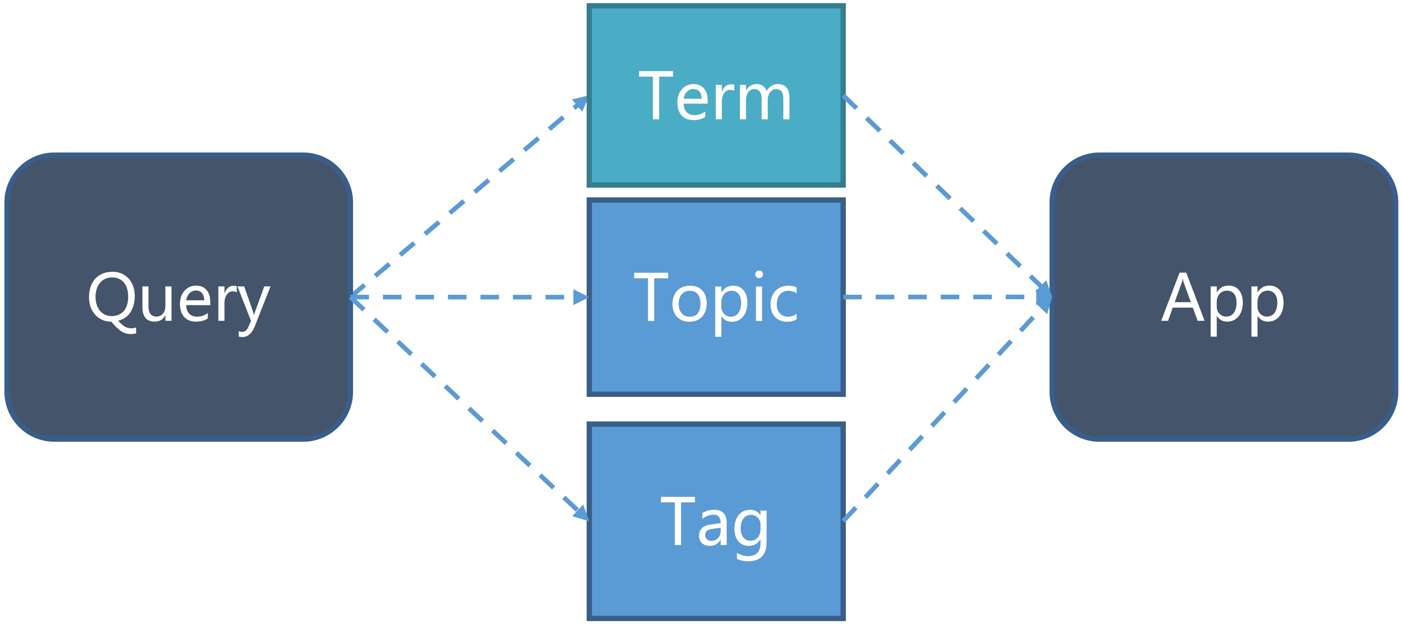
微信变声器
★★★☆☆ 188万人下载 12.06MB
可以将用户的声音处理成各种动物或者是鬼...

微信表情大全
★★★★★ 171万人下载 9.47MB
支持各种聊天工具及微博社区噢。

微信语音表情
★★★★☆ 55万人下载 5.59MB
颠覆你的搞笑底线，让你的微信聊天更有趣。

Semantic Matching Methodology

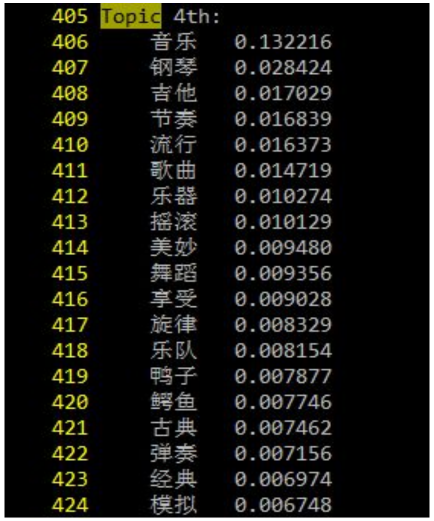
- ✓ How to describe "Semantic "?
  - ● Term + Topic + Tag
- ✓ Enrich query and app representations by topic and tag
- ✓ Perform query-app matching with the representations
- ✓ Hybrid Ranking Model: LTR



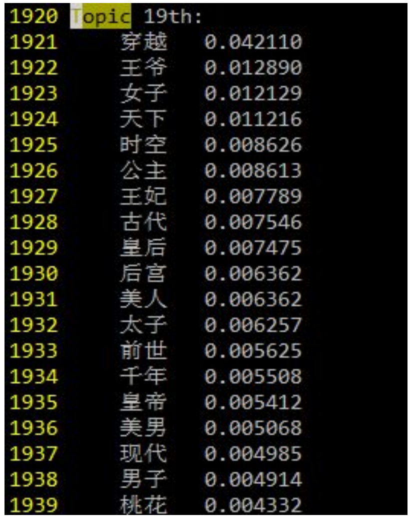
Matching with Topic

✓ Topic Model

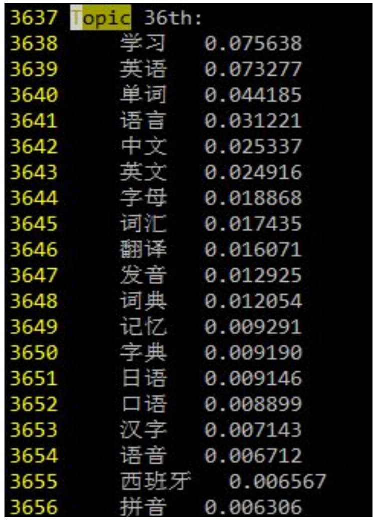
- Using Layered LDA(Latent Dirichlet Allocation) model
- MPI based parallel computing framework
- Topic probability distribution over term space: P(word|topic)



```
Topic 4th:
        音乐    0.132216
        钢琴    0.028424
        吉他    0.017029
        节奏    0.016839
        流行    0.016373
        歌曲    0.014719
        乐器    0.010274
        摇滚    0.010129
        美妙    0.009480
        舞蹈    0.009356
        享受    0.009028
        旋律    0.008329
        乐队    0.008154
        鸭子    0.007877
        鳄鱼    0.007746
        古典    0.007462
        弹奏    0.007156
        经典    0.006974
        模拟    0.006748
```

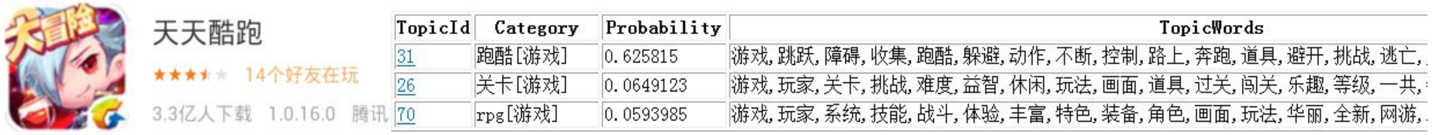


```
Topic 19th:
        穿越    0.042110
        王爷    0.012890
        女子    0.012129
        天下    0.011216
        时空    0.008626
        公主    0.008613
        王妃    0.007789
        古代    0.007546
        皇后    0.007475
        后宫    0.006362
        美人    0.006362
        太子    0.006257
        前世    0.005625
        千年    0.005508
        皇帝    0.005412
        美男    0.005068
        现代    0.004985
        男子    0.004914
        桃花    0.004332
```



```
Topic 36th:
        学习    0.075638
        英语    0.073277
        单词    0.044185
        语言    0.031221
        中文    0.025337
        英文    0.024916
        字母    0.018868
        词汇    0.017435
        翻译    0.016071
        发音    0.012925
        词典    0.012054
        记忆    0.009291
        字典    0.009190
        日语    0.009146
        口语    0.008899
        汉字    0.007143
        语音    0.006712
        西班牙    0.006567
        拼音    0.006306
```

- Assign million apps to 1000+ topics
- Doc probability distribution over topic space: P(topic|doc)



天天酷跑

★★★★ 14个好友在玩

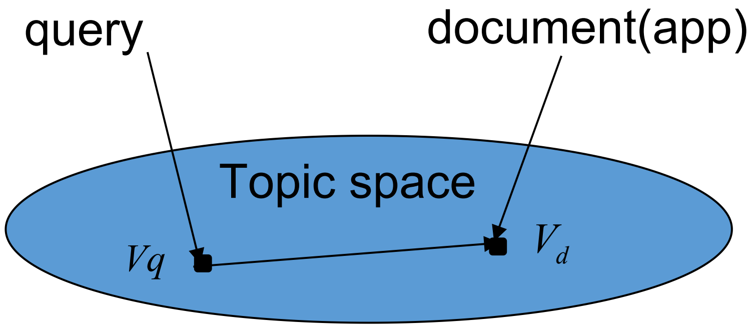
3.3亿人下载 1.0.16.0 腾讯

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 Vq and documents V_d are both represented with probability distributions over topics
  - ● Calculate topic match score between Vq 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.

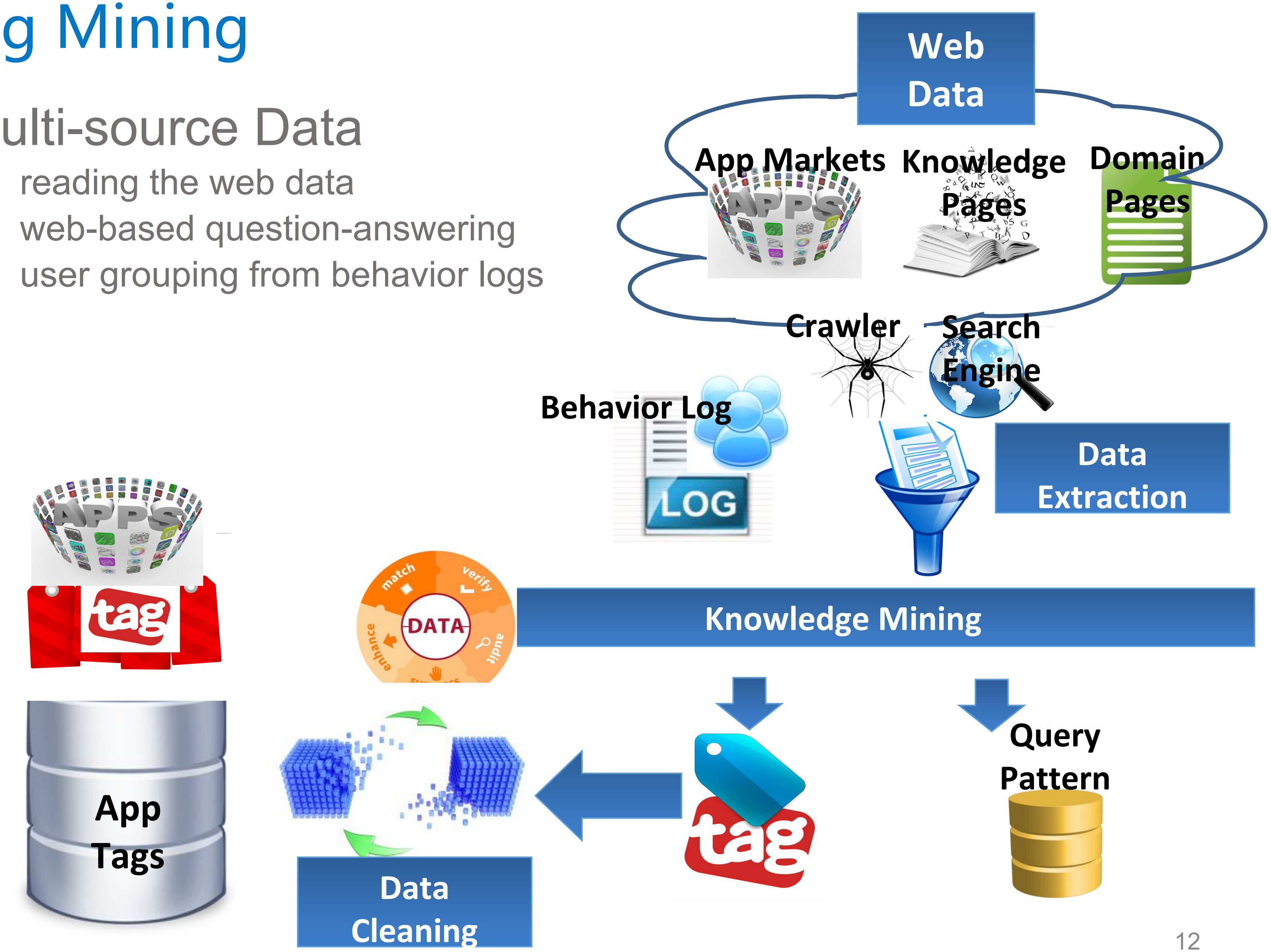
✓ Matching with Tag

- 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

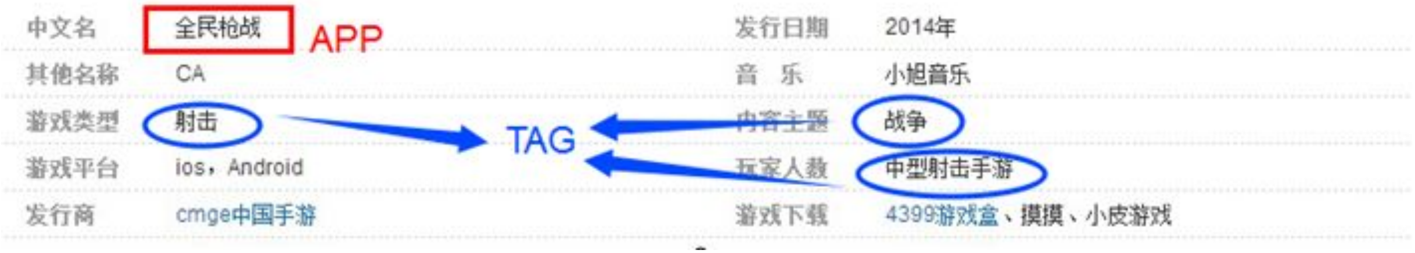
- reading the web data
- web-based question-answering
- user grouping from behavior logs



Tag extraction from web data

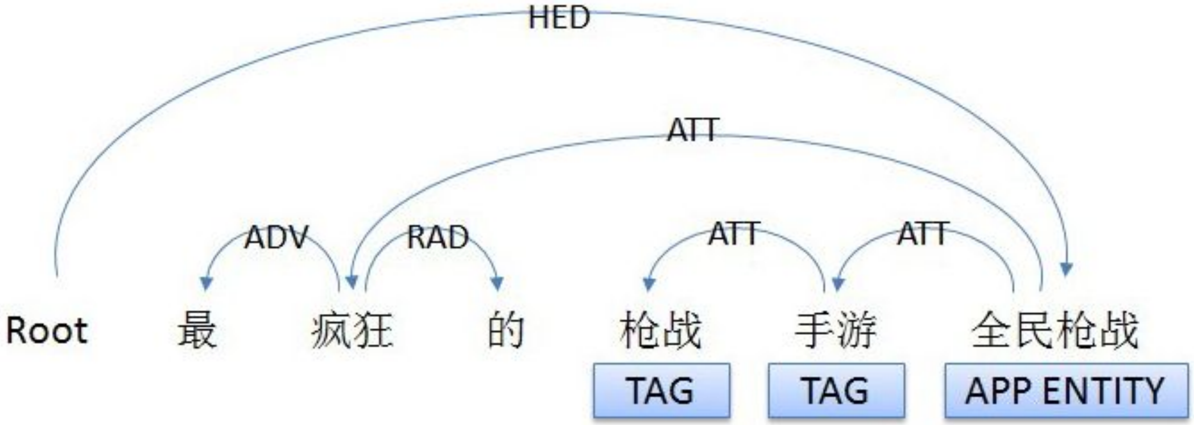
✓Data from web

- Structured data



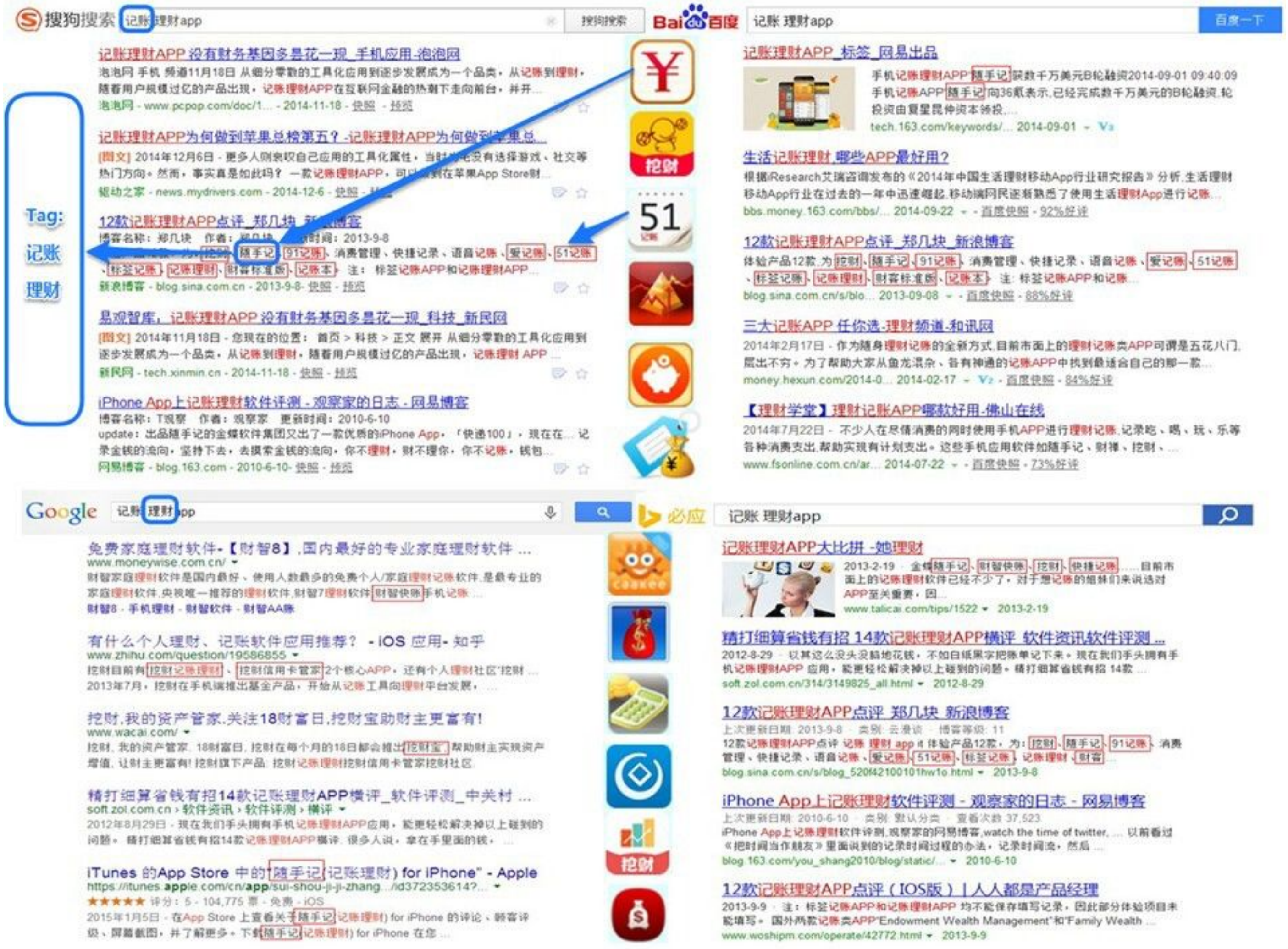
中文名	全民枪战	发行日期	2014年
其他名称	CA	音　乐	小旭音乐
游戏类型	射击	内容主题	战争
游戏平台	ios，Android	玩家人数	中型射击手游
发行商	cmge中国手游	游戏下载	4399游戏盒、摸摸、小皮游戏

- Unsructured text
  - using template to extraction



Tag from question-answering

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



搜狗搜索 记账理财app

记账理财APP 没有财务基因多是花一现_手机应用-泡泡网
泡泡网 手机 频道11月18日 从细分零散的工具化应用到逐步发展成为一个品类，从记账到理财，随着用户规模过亿的产品出现，记账理财APP在互联网金融的热潮下走向前台，并开...
泡泡网 - www.pcpop.com/doc/1... - 2014-11-18 - 快照 - 预览

记账理财APP为何做到苹果总榜第五？-记账理财APP为何做到苹果总...
[图文] 2014年12月6日 - 更多人则亲叹自己应用的工具化属性，当时[illegible]没有选择游戏、社交等热门方向。然而，事实真是如此吗？一款记账理财APP，可以[illegible]在苹果App Store财...
驱动之家 - news.mydrivers.com - 2014-12-6 - 快照 - [illegible]

12款记账理财APP点评_郑几块_新浪博客
博客名称：郑几块 作者：[illegible] 更新时间：2013-9-8
[illegible]挖财、随手记、91记账、消费管理、快捷记录、语音记账、爱记账、51记账、标签记账、记账理财、财客标准版、记账本 注：标签记账APP和记账理财APP...
新浪博客 - blog.sina.com.cn - 2013-9-8- 快照 - 预览

易观智库：记账理财APP 没有财务基因多是花一现_科技_新民网
[图文] 2014年11月18日 - 您现在的位置： 首页 > 科技 > 正文 展开 从细分零散的工具化应用到逐步发展成为一个品类，从记账到理财，随着用户规模过亿的产品出现，记账理财 APP ...
新民网 - tech.xinmin.cn - 2014-11-18 - 快照 - 预览

iPhone App上记账理财软件评测 - 观察家的日志 - 网易博客
博客名称：T观察 作者：观察家 更新时间：2010-6-10
update：出品随手记的金蝶软件集团又出了一款优质的iPhone App，「快递100」，现在在... 记录金钱的流向，坚持下去，去摸索金钱的流向，你不理财，财不理你，你不记账，钱包...
网易博客 - blog.163.com - 2010-6-10- 快照 - 预览

Baidu百度 记账 理财app 百度一下

记账理财APP_标签_网易出品
手机记账理财APP"随手记"获数千万美元B轮融资2014-09-01 09:40:09 手机记账APP"随手记"向36氪表示，已经完成数千万美元的B轮融资，轮投资由夏星昆仲资本领投,...
tech.163.com/keywords/... 2014-09-01 - V3

生活记账理财,哪些APP最好用?
根据iResearch艾瑞咨询发布的《2014年中国生活理财移动App行业研究报告》分析,生活理财移动App行业在过去的一年中迅速崛起,移动端网民逐渐熟悉了使用生活理财App进行记账...
bbs.money.163.com/bbs/... 2014-09-22 - - 百度快照 - 92%好评

12款记账理财APP点评_郑几块_新浪博客
体验产品12款,为挖财、随手记、91记账、消费管理、快捷记录、语音记账、爱记账、51记账、标签记账、记账理财、财客标准版、记账本 注: 标签记账APP和记账...
blog.sina.com.cn/s/blo... 2013-09-08 - - 百度快照 - 88%好评

三大记账APP 任你选-理财频道-和讯网
2014年2月17日 - 作为随身理财记账的全新方式,目前市面上的理财记账类APP可谓是五花八门,层出不穷。为了帮助大家从鱼龙混杂、各有神通的记账APP中找到最适合自己的那一款...
money.hexun.com/2014-0... 2014-02-17 - V2 - 百度快照 - 84%好评

【理财学堂】理财记账APP哪款好用-佛山在线
2014年7月22日 - 不少人在尽情消费的同时使用手机APP进行理财记账,记录吃、喝、玩、乐等各种消费支出,帮助实现有计划支出。这些手机应用软件如随手记、财禅、挖财、...
www.fsonline.com.cn/ar... 2014-07-22 - - 百度快照 - 73%好评

Google 记账 理财app

免费家庭理财软件-【财智8】,国内最好的专业家庭理财软件 ...
www.moneywise.com.cn/
财智家庭理财软件是国内最好、使用人数最多的免费个人/家庭理财记账软件,是最专业的家庭理财软件,央视唯一推荐的理财软件,财智7理财软件,财智快账手机记账 ...
财智8 - 手机理财 - 财智软件 - 财智AA账

有什么个人理财、记账软件应用推荐？ - iOS 应用- 知乎
www.zhihu.com/question/19586855
挖财目前有'挖财记账理财'、'挖财信用卡管家'2个核心APP，还有个人理财社区'挖财 ... 2013年7月，挖财在手机端推出基金产品，开始从记账工具向理财平台发展， ...

挖财,我的资产管家,关注18财富日,挖财宝助财主更富有!
www.wacai.com/
挖财、我的资产管家. 18财富日. 挖财在每个月的18日都会推出'挖财宝', 帮助财主实现资产增值, 让财主更富有! 挖财旗下产品: 挖财记账理财挖财信用卡管家挖财社区

精打细算省钱有招14款记账理财APP横评_软件评测_中关村 ...
soft.zol.com.cn › 软件资讯 › 软件评测 › 横评
2012年8月29日 - 现在我们手头拥有手机记账理财APP应用，能更轻松解决掉以上碰到的问题。 精打细算省钱有招14款记账理财APP横评. 很多人说，拿在手里面的钱， ...

iTunes 的App Store 中的"随手记(记账理财) for iPhone" - Apple
https://itunes.apple.com/cn/app/sui-shou-ji-ji-zhang.../id372353614?...
★★★★★ 评分：5 - 104,775 票 - 免费 - iOS
2015年1月5日 - 在App Store 上查看关于随手记(记账理财) for iPhone 的评论、顾客评级、屏幕截图，并了解更多。下载随手记(记账理财) for iPhone 在您 ...

必应 记账 理财app

记账理财APP大比拼 -她理财
2013-2-19 · 金蝶随手记、财智快账、挖财、快捷记账......目前市面上的记账理财软件已经不少了，对于想记账的姐妹们来说选对APP至关重要，因...
www.talicai.com/tips/1522 2013-2-19

精打细算省钱有招 14款记账理财APP横评 软件资讯软件评测 ...
2012-8-29 · 以其这么没头没脑地花钱，不如白纸黑字把账单记下来。现在我们手头拥有手机记账理财APP 应用，能更轻松解决掉以上碰到的问题。精打细算省钱有招 14款 ...
soft.zol.com.cn/314/3149825_all.html 2012-8-29

12款记账理财APP点评 郑几块 新浪博客
上次更新日期 2013-9-8 · 类别 云漫谈 · 博客等级 11
12款记账理财APP点评 记账 理财 app it 体验产品12款，为：挖财、随手记、91记账、消费管理、快捷记录、语音记账、爱记账、51记账、标签记账、记账理财、财客 ...
blog.sina.com.cn/s/blog_520f42100101hw1o.html 2013-9-8

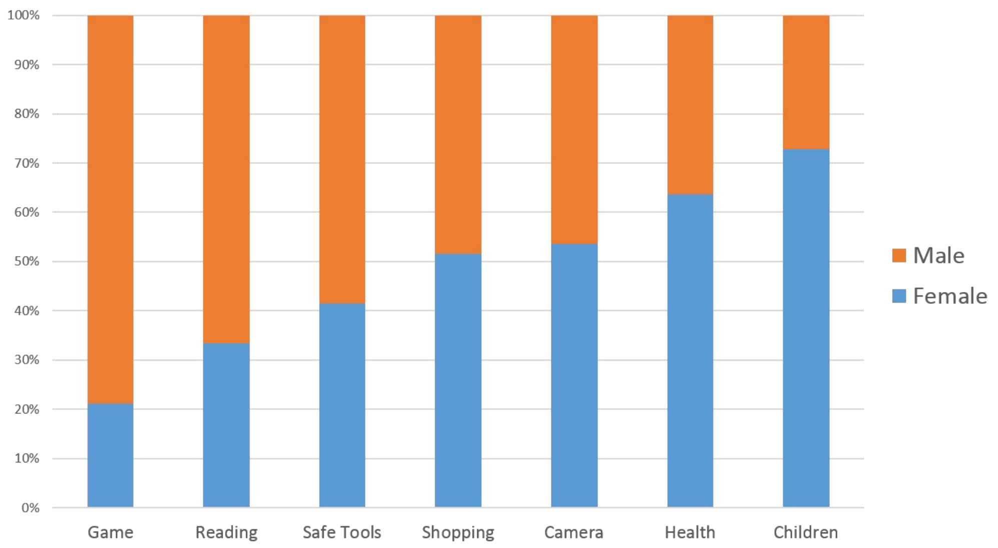
iPhone App上记账理财软件评测 - 观察家的日志 - 网易博客
上次更新日期 2010-6-10 · 类别 默认分类 · 查看次数 37,523
iPhone App上记账理财软件评测,观察家的网易博客,watch the time of twitter, ... 以前看过《把时间当作朋友》里面说到的记录时间过程的办法，记录时间流，然后 ...
blog.163.com/you_shang2010/blog/static/... 2010-6-10

12款记账理财APP点评（IOS版）| 人人都是产品经理
2013-9-9 · 注：标签记账APP和记账理财APP 均不能保存填写记录，因此部分体验项目未能填写。国外两款记账类APP"Endowment Wealth Management"和"Family Wealth ...
www.woshipm.com/operate/42772.html 2013-9-9

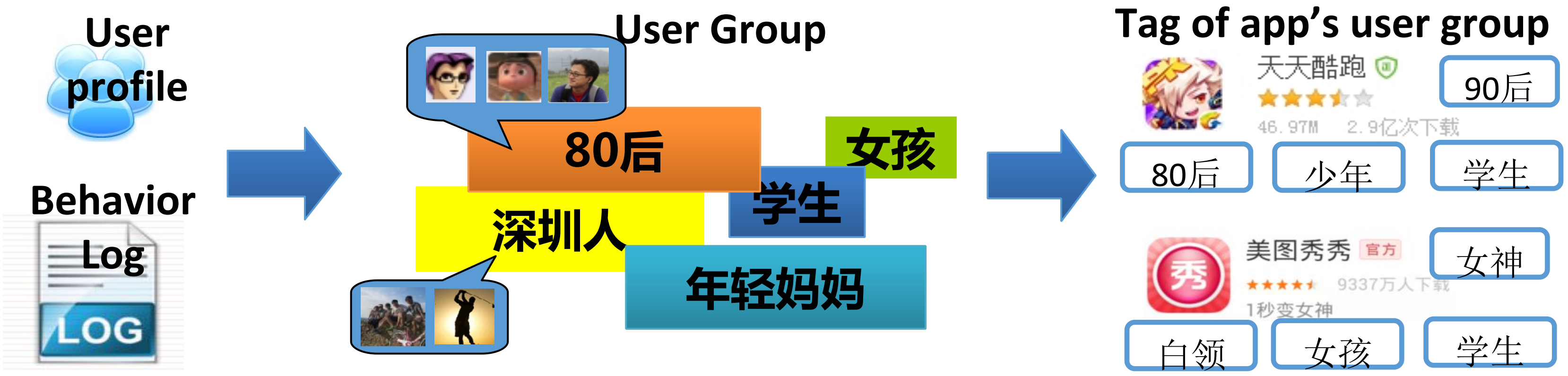
Tag from user behavior logs

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

Download-ratio by gender and category

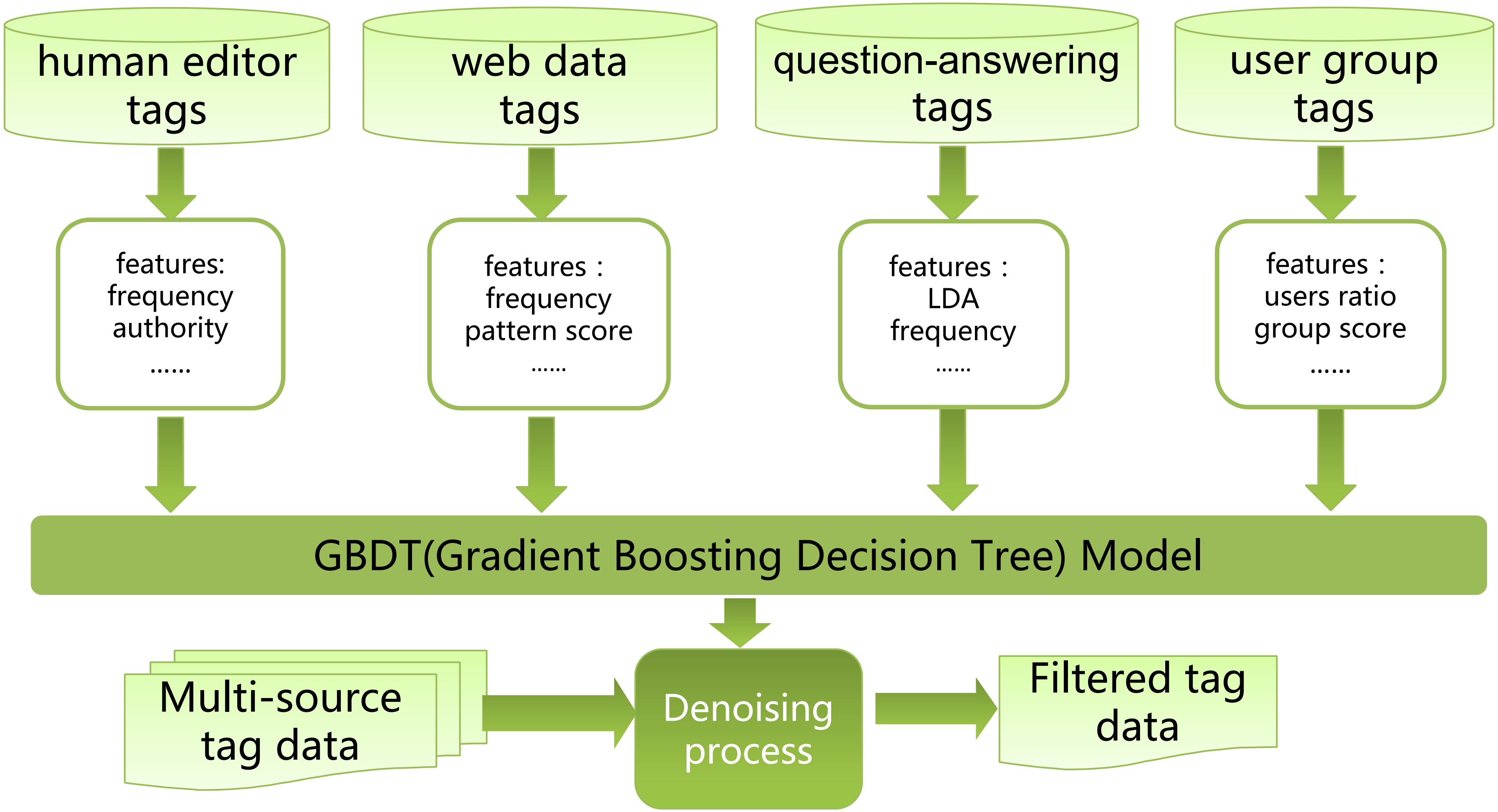


- Based on user-behavior data and natural representation by tag



Tag Denoising

- Using machine learning technology to calculate confidence


human editor tags
web data tags
question-answering tags
user group tags
features: frequency authority ……
features : frequency pattern score ……
features : LDA frequency ……
features : users ratio group score ……
GBDT(Gradient Boosting Decision Tree) Model
Multi-source tag data
Denoising process
Filtered tag data

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%)

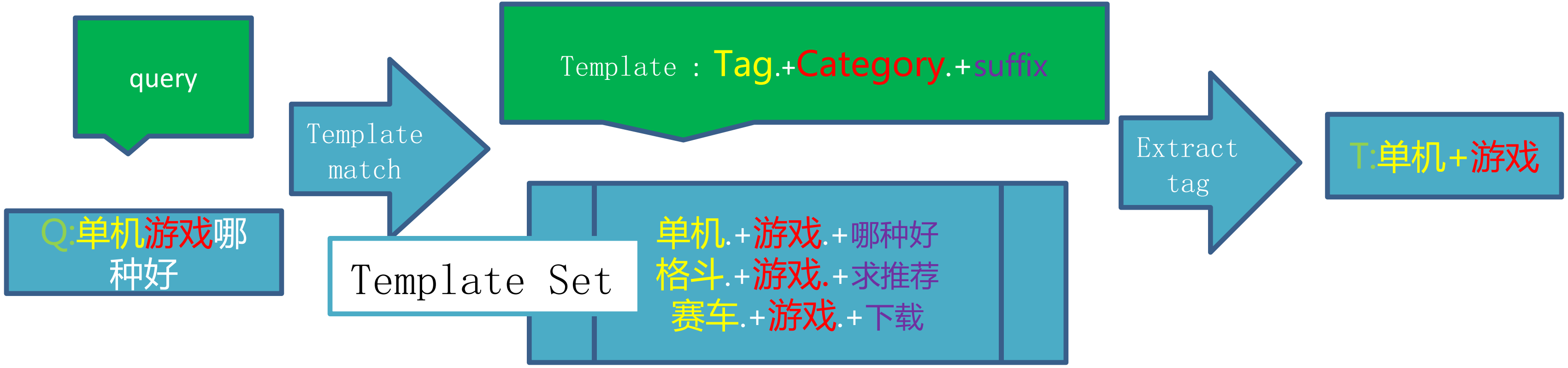
✓Statistic data

- ●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

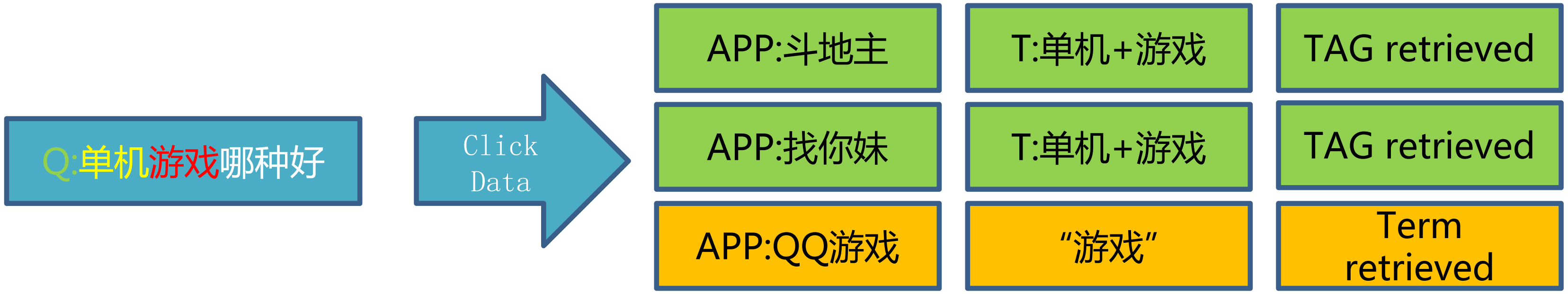
Matching query with tag

✓Using template to map query to tags



✓Using click data to calculate confidence of template

- P(Template) = 2/3



APP:斗地主	T:单机+游戏	TAG retrieved
APP:找你妹	T:单机+游戏	TAG retrieved
APP:QQ游戏	"游戏"	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 traning data
- 3000+ test samples

✓ Offline Experiment measurement

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

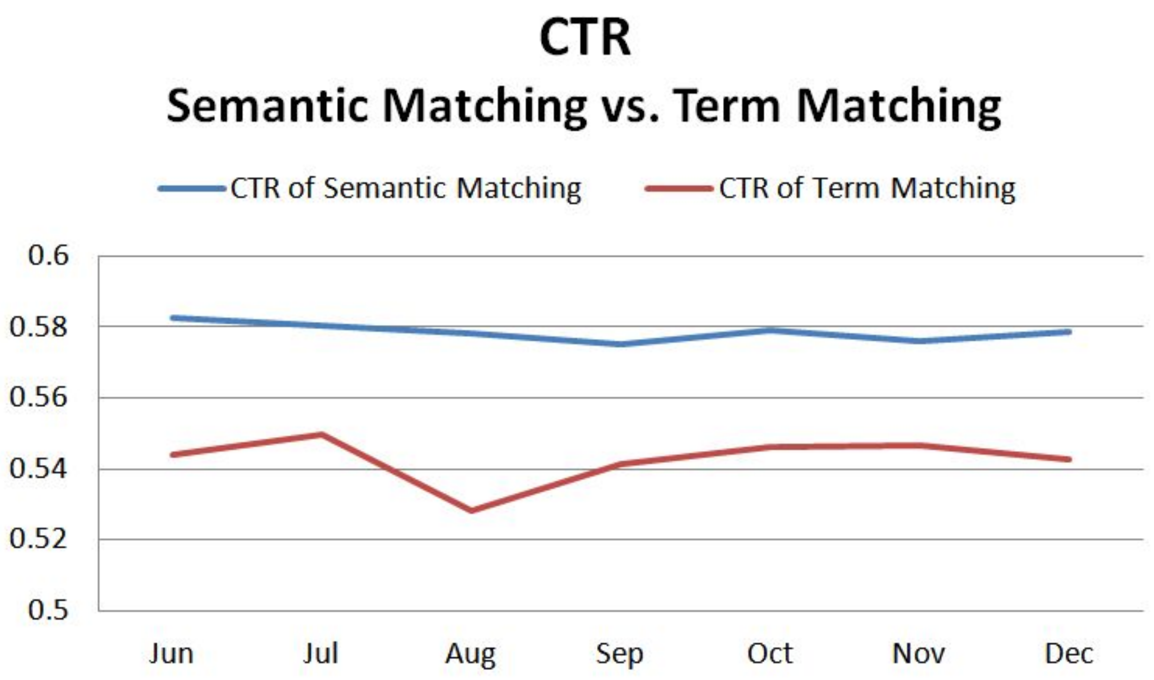
✓ Online A/B Test measurement

- CTR promoted by 6%↑

Semantic Matching Metrics

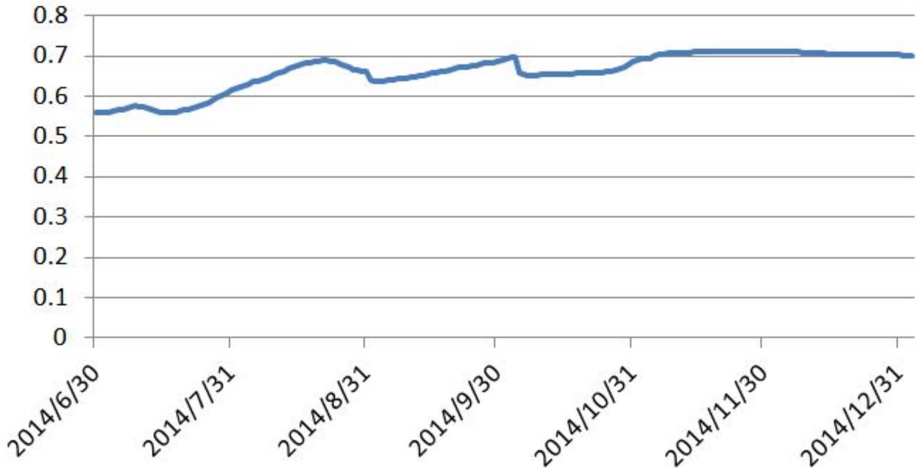
✓ Online A/B Test measurement

- CTR diff. on query samples
- 9.7% querys & 26.9% query views
- CTR promoted by 6%~8%↑

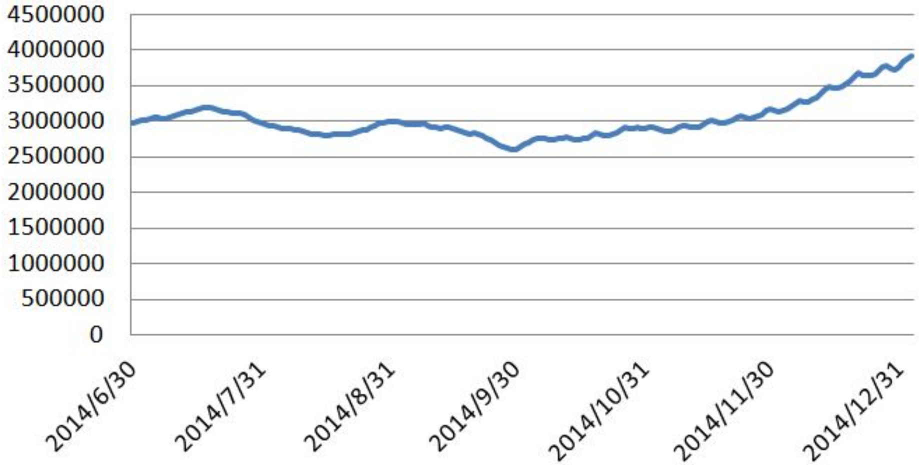


App Search with Semantic Matching

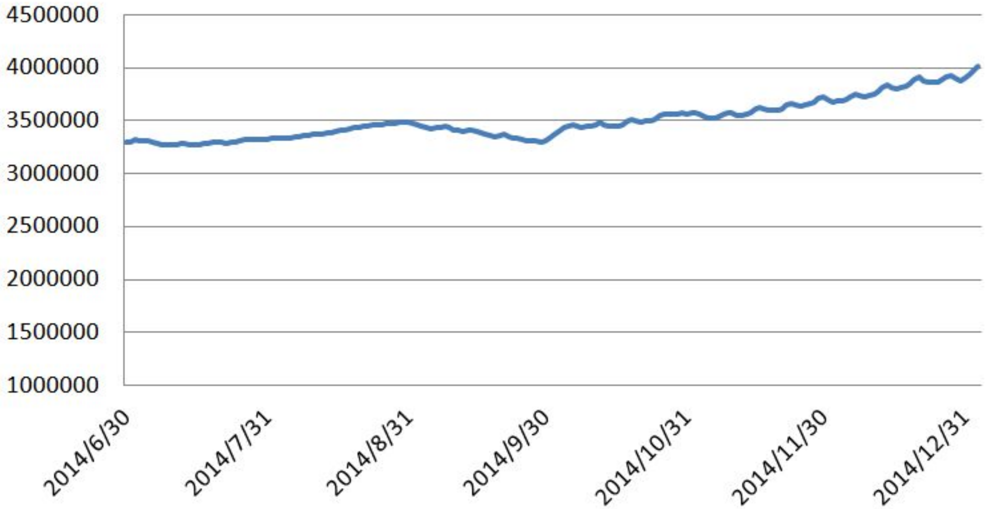
CTR



Search UV

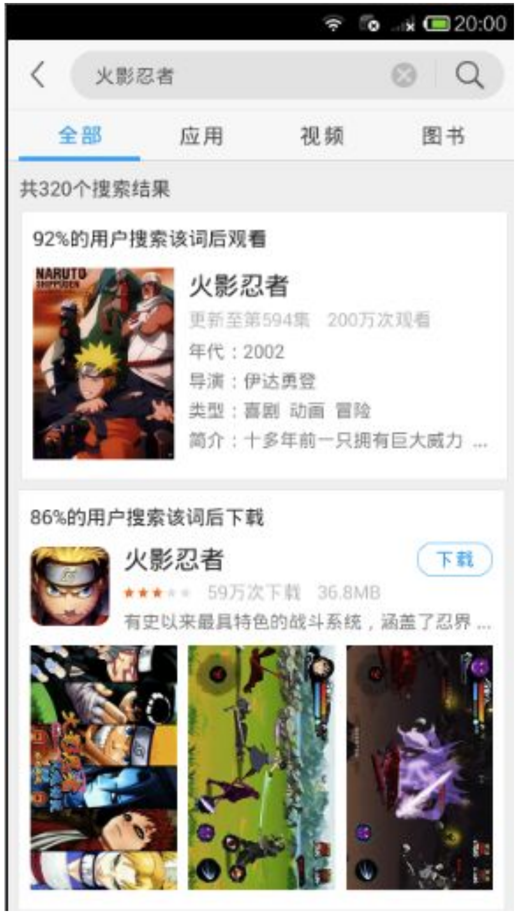


Download Apps



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%↑



共320个搜索结果

92%的用户搜索该词后观看

火影忍者

更新至第594集 200万次观看

年代：2002

导演：伊达勇登

类型：喜剧 动画 冒险

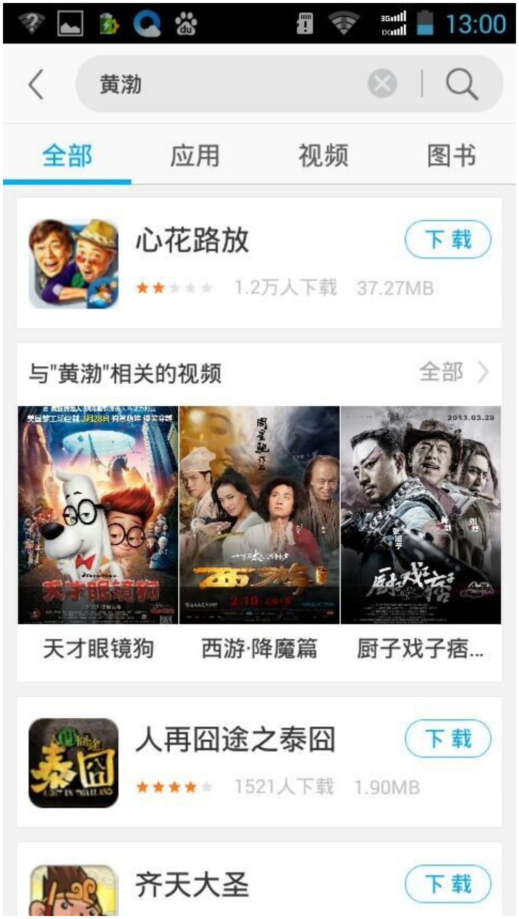
简介：十多年前一只拥有巨大威力 ...

86%的用户搜索该词后下载

火影忍者

★★★★★ 59万次下载 36.8MB

有史以来最具特色的战斗系统，涵盖了忍界 ...



心花路放

★★★★★ 1.2万人下载 37.27MB

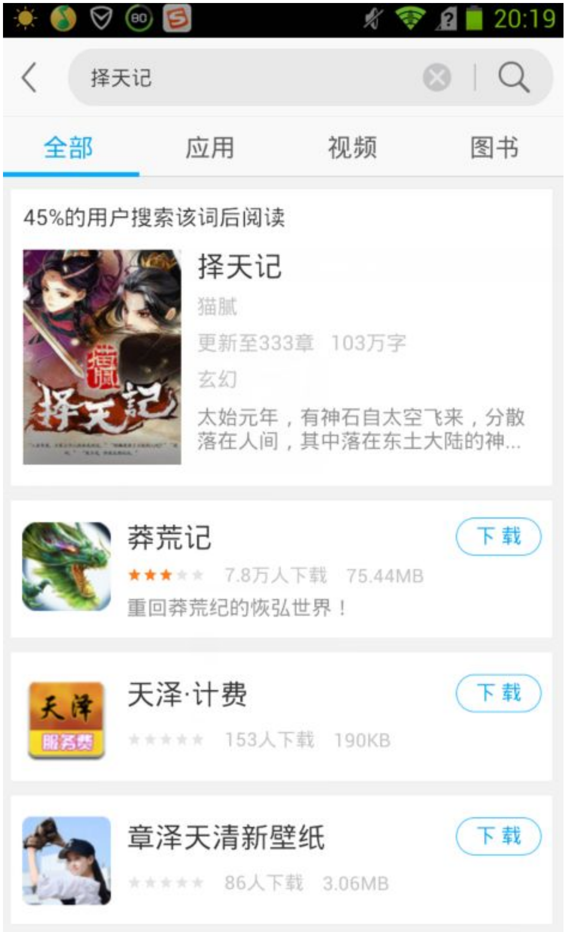
与"黄渤"相关的视频 全部 >

天才眼镜狗　西游·降魔篇　厨子戏子痞...

人再囧途之泰囧

★★★★★ 1521人下载 1.90MB

齐天大圣



45%的用户搜索该词后阅读

择天记

猫腻

更新至333章 103万字

玄幻

太始元年，有神石自太空飞来，分散落在人间，其中落在东土大陆的神...

莽荒记

★★★★★ 7.8万人下载 75.44MB

重回莽荒纪的恢弘世界！

天泽·计费

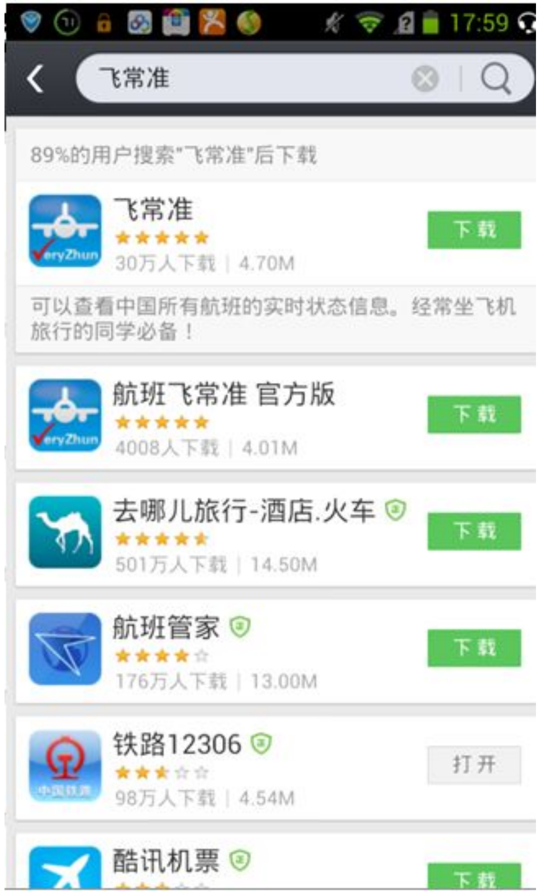
★★★★★ 153人下载 190KB

章泽天清新壁纸

★★★★★ 86人下载 3.06MB

Semantic Matching Application Example

- ✓ Deeper understanding colloquial form of query



飞常准

89%的用户搜索"飞常准"后下载

飞常准
30万人下载 | 4.70M
可以查看中国所有航班的实时状态信息。经常坐飞机旅行的同学必备！

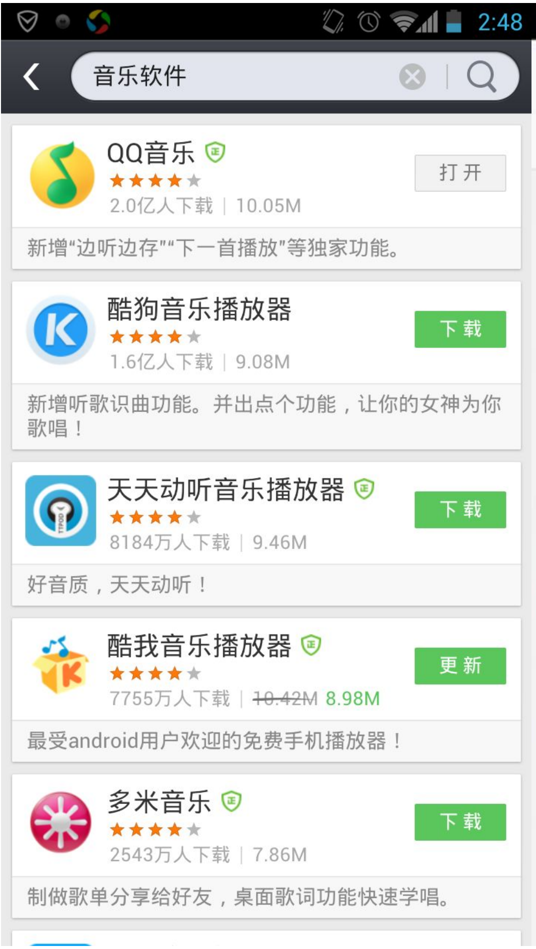
航班飞常准 官方版
4008人下载 | 4.01M

去哪儿旅行-酒店.火车
501万人下载 | 14.50M

航班管家
176万人下载 | 13.00M

铁路12306
98万人下载 | 4.54M

酷讯机票



音乐软件

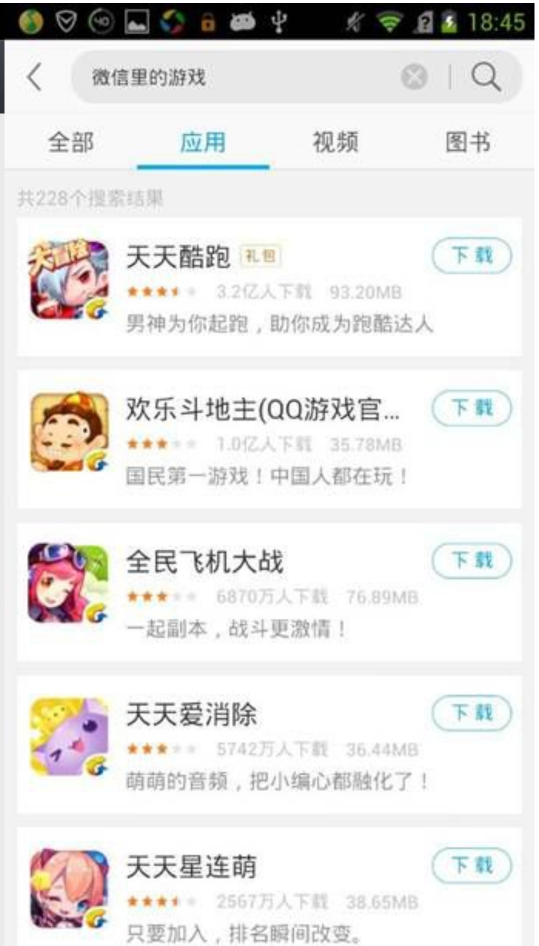
QQ音乐
2.0亿人下载 | 10.05M
新增“边听边存”“下一首播放”等独家功能。

酷狗音乐播放器
1.6亿人下载 | 9.08M
新增听歌识曲功能。并出点个功能，让你的女神为你歌唱！

天天动听音乐播放器
8184万人下载 | 9.46M
好音质，天天动听！

酷我音乐播放器
7755万人下载 | ~~10.42M~~ 8.98M
最受android用户欢迎的免费手机播放器！

多米音乐
2543万人下载 | 7.86M
制做歌单分享给好友，桌面歌词功能快速学唱。



微信里的游戏

共228个搜索结果

天天酷跑
3.2亿人下载 93.20MB
男神为你起跑，助你成为跑酷达人

欢乐斗地主(QQ游戏官...
1.0亿人下载 35.78MB
国民第一游戏！中国人都在玩！

全民飞机大战
6870万人下载 76.89MB
一起副本，战斗更激情！

天天爱消除
5742万人下载 36.44MB
萌萌的音频，把小编心都融化了！

天天星连萌
2567万人下载 38.65MB
只要加入，排名瞬间改变。

Conclusion

✓ From **Term matching** to **Semantic matching**

- Richer representation of semantic

✓ Methodology of Enriching infomation

- 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