DTCC

2016中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2016

数 据 定 义 未 来

SequeMedia 盛拓传媒　ITPUB.net　ChinaUnix　ITPUB

叶祺
北京搜狗科技发展有限公司

基于大数据的查询意图识别其应用

目录

- 动机与目标
- 现有方法
- 框架与方法
- 效果与应用

- **动机与目标**
- 现有方法
- 框架与方法
- 效果与应用

动机与目标

- 搜索广告的现状
  - 当前的搜索广告中，搜索引擎主要基于关键字匹配的搜索模式。
- 问题
  - 查询短、特征稀疏、歧义强
  - 字面匹配缺乏意图相关特征
  - 广告缺乏相关性
  - 伤害用户体验、造成客户无效消耗

动机与目标

- 目标
  - 挖掘海量细粒度查询意图
  - 建立查询与意图间映射关系
  - 处理高频与长尾查询
  - 高精确性与较高覆盖率

- 动机与目标
- **现有方法**
- 框架与方法
- 效果与应用

现有方法

- Google的Google Rephil系统
  - Google广告相关性的头号秘密武器
  - 对词或短语片段聚类发现概念
  - 百万量级的概念
  - 基于Bayesian网络的推断方法
  - 细节不公开

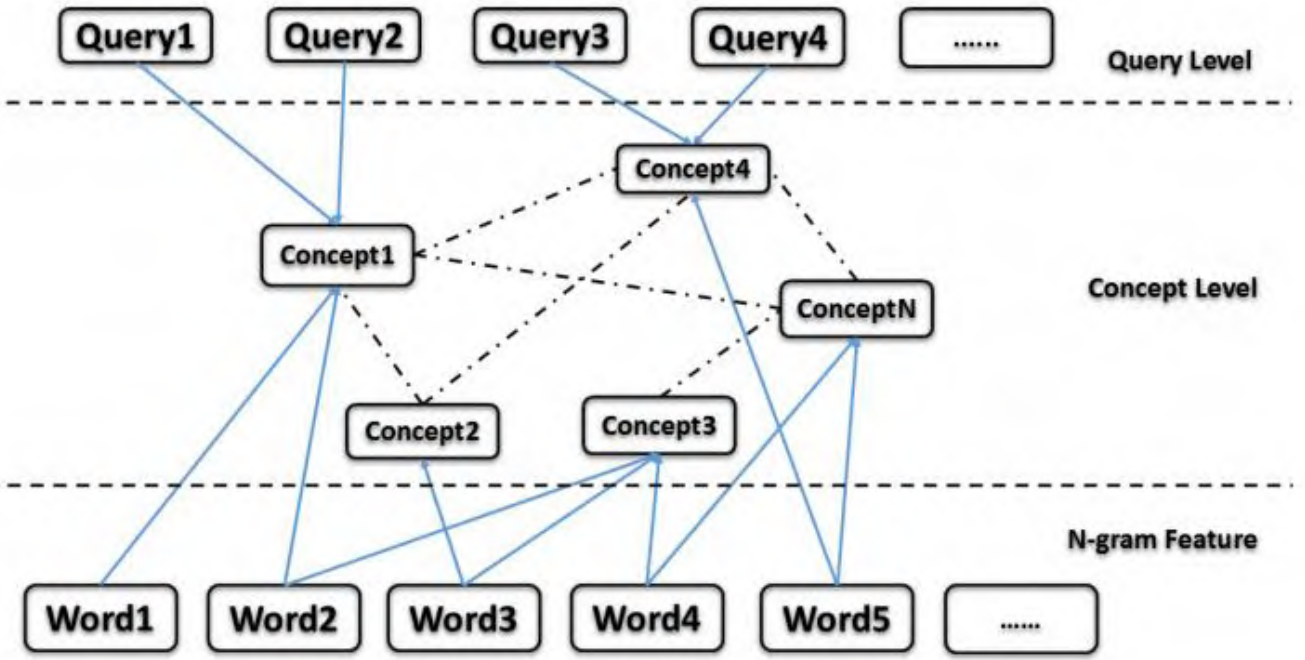
现有方法

- 识别意图的3类方法
  - 短文本聚类
  - Topic Modeling
  - 查询分类
- 特点
  - 可发现细粒度意图、难覆盖长尾查询
  - 不同数据集Topic难对应，短文本分析精确不足
  - 一般含几十到上千个类，粒度较粗

- 动机与目标
- 现有方法
- **框架与方法**
- 效果与应用

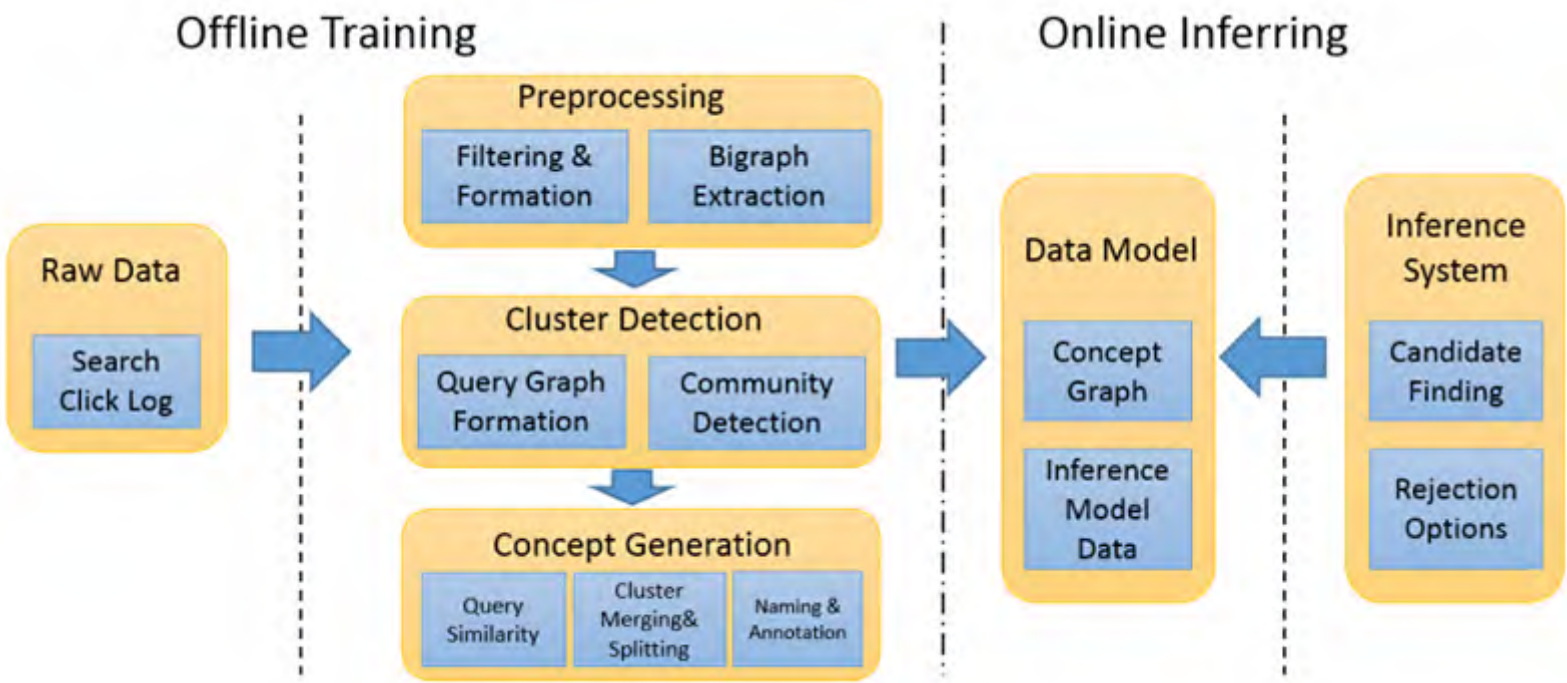
框架与方法

- 细粒度意图识别方法



框架与方法

- 星辰系统整体框架



查询聚类

- 构建Query同点击网络
  - 基本假设：点击相同网页的查询意图相似
- 对网络进行社团划分
  - 查询间的意图会有细微差别、误点情况
  - 聚类算法要具有一定抗噪性
  - 图挖掘中的社团发现算法

社团发现算法

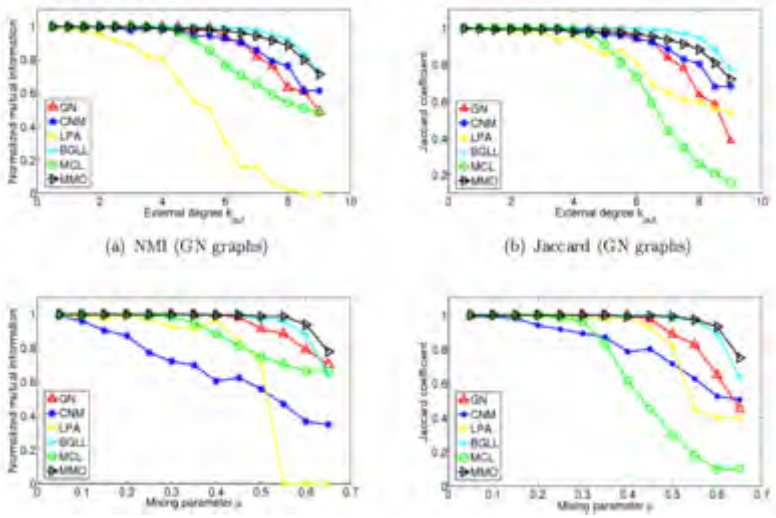
- 社团的定义
  - 网络中一群节点集合。
  - 集合中节点间的内部链接很多，而集合中节点与外部网络的链接却很少。
- 传统方法
  - 主要发端于 Girvan 与 Newman 于 2002 年提出的开创性工作
  - 定义了一个质量函数

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

M. Girvan and M. E. J. Newman, PNAS **99**, 7821 (2002).
M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

MMO算法

- MMO算法

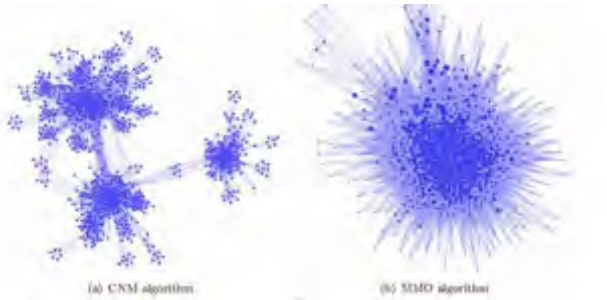


(a) NMI (GN graphs)　(b) Jaccard (GN graphs)

Algorithm 1 MMO algorithm

```
C = {{v}|v ∈ V}
for i = 1 to I_max do
  Q_old = Q(C)
  for v ∈ V do
    for v' ∈ N_v do
      c' = C_v'
      m^c'_v = m^c'_v + 1 {The default value of m^c'_v is 0}
    end for
    c_old = C_v
    c_best = c_old
    remove v from c_old
    Q_max = ΔQ(c_best, v)
    for v' ∈ N_v do
      Q' = ΔQ(C_v', v)
      if Q' is better than Q_max then
        Q_max = Q'
        c_best = C_v'
      end if
    end for
    insert v into c_best
  end for
  Q_new = Q(C)
  δ = Q_new − Q_old
  if δ < ε then
    break
  end if
  Q_old = Q_new
end for
for e_i,j in E do
  if C_i ≠ C_j then
    mark e_i,j as unlinked
  end if
end for
Find all the components as communities C
return C
```

- MMO算法的优点
  - 易于实现
  - 时间复杂度近似线性，空间复杂度为线性。
  - 推广到 Hadoop 并行环境中的运行
  - 避免生成极大的社团



(a) CNM algorithm　(b) MMO algorithm

同点击网络构造

- 数据集
  - 2年的匿名点击日志
- 具体步骤
  - 抽取query-URL的关系（1300万查询，1650万URL）
  - 如果两个query间有一个同点击，则在两个query间连接一条边
  - 得到查询同点击网络（1300万查询节点，8亿条边）

概念质量优化

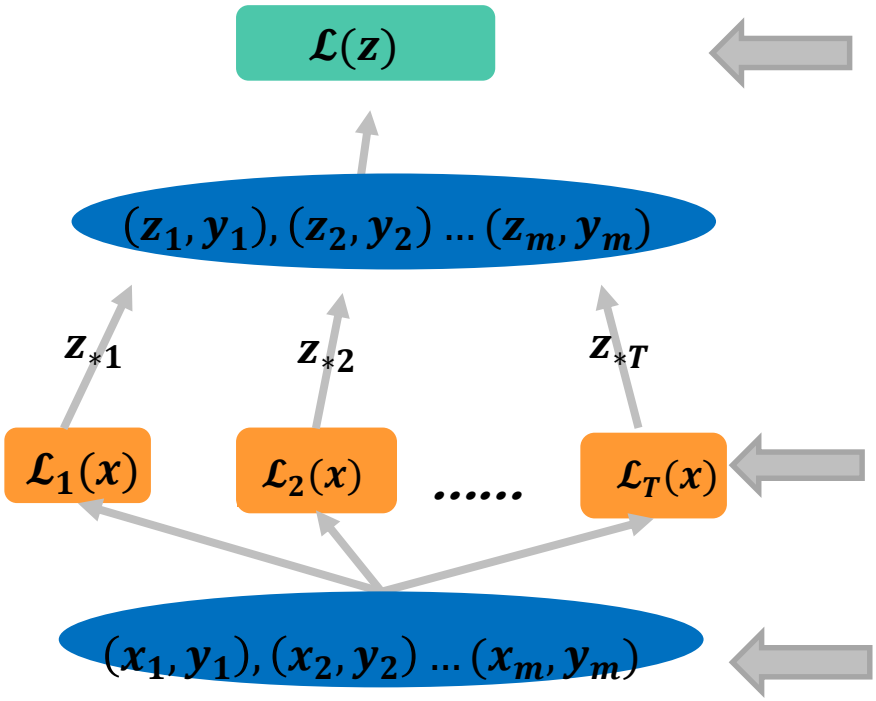
- 聚类存在的问题
  - 过大的不纯类
  - 太多的细粒度聚类
- 聚类质量评估
  - 聚类纯度
  - 聚类间的相关性

$$r(c) = \frac{\sum_{q,s \in c, q \neq s} f(q,s)}{|c| \times (|c| - 1)},$$

$$r(c_i, c_j) = \frac{\sum_{q \in c_i, s \in c_j} f(q,s)}{|c_i| \times |c_j|}$$

文本相关性计算方法

- 组合方法
  ——Stacking Learning



$$y = \frac{1}{1 + e^{\omega_0 + \omega_1 x_1 + \cdots + \omega_m x_m}}$$

Co-click Sim　　Co-Session Sim

LDA扩展　　网页搜索扩展

TFIDF相似性　　同义词相似性

$(< q_1, a_1 >, m_1)$

查询意图推断

- 问题定义
  - 将query的意图识别变为一个大规模多分类问题
- 关键步骤
  - 候选分类概念
  - 拒绝分类结果

查询意图推断

- 候选概念的发现

$$
\begin{aligned}
c &= \underset{c \in C}{\operatorname{argmax}}\, p(c|\mathbf{x}_q) \\
&\propto \underset{c \in C}{\operatorname{argmax}}\, p(\mathbf{x}_q|c) \times p(c) \\
&= \underset{c \in C}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|c) \times p(c) \\
&\propto \underset{c \in C}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_i|c) + \log p(c).
\end{aligned}
$$

x_q: feature vector of query q.

查询意图推断

- 拒绝项
  - Query侧相关性：

$$s_q(\mathbf{v}_q, c) = \frac{\sum_{x_i \in \mathbf{x}_q, \mathbf{x}_c} \mathbf{v}_q(x_i)}{\sum_{x_i \in \mathbf{x}_q} \mathbf{v}_q(x_i)}$$

  - 概念侧相关性：

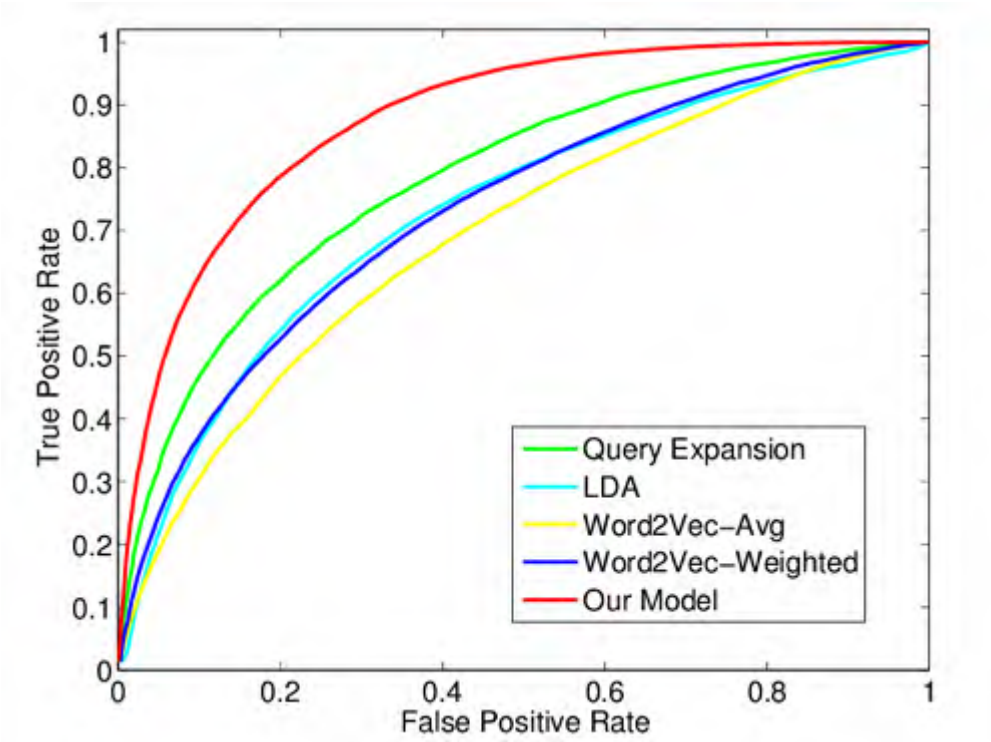
$$s_c(\mathbf{v}_q, c) = \frac{\sum_{x_i \in \mathbf{x}_q, \mathbf{x}_c} \mathbf{v}_c(x_i)}{\sum_{x_i \in \mathbf{x}_c} \mathbf{v}_c(x_i)} \propto \sum_{x_i \in \mathbf{x}_q, \mathbf{x}_c} \mathbf{v}_c(x_i).$$

$$\lambda_c = \frac{\sum_{q \in c} s_c(\mathbf{v}_q, c)}{|c|} \propto \frac{\sum_{q \in c} \sum_{x_i \in \mathbf{x}_q, \mathbf{x}_c} \mathbf{v}_c(x_i)}{|c|}$$

- 动机与目标
- 现有方法
- 框架与方法
- **效果与应用**

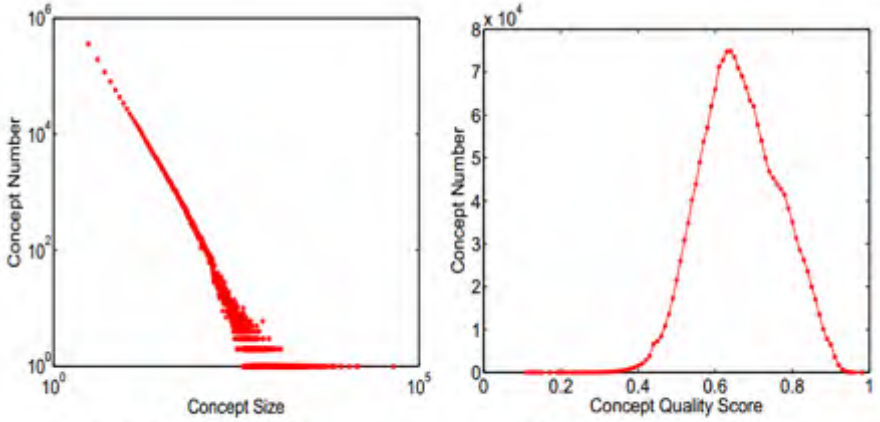
文本相关性模型的效果

- 对比方法：
  - 查询扩展
  - w2v
  - 字面匹配
  - LDA



概念发现的结果

- 概念分布



(a) Concept Size (b) Concept Quality

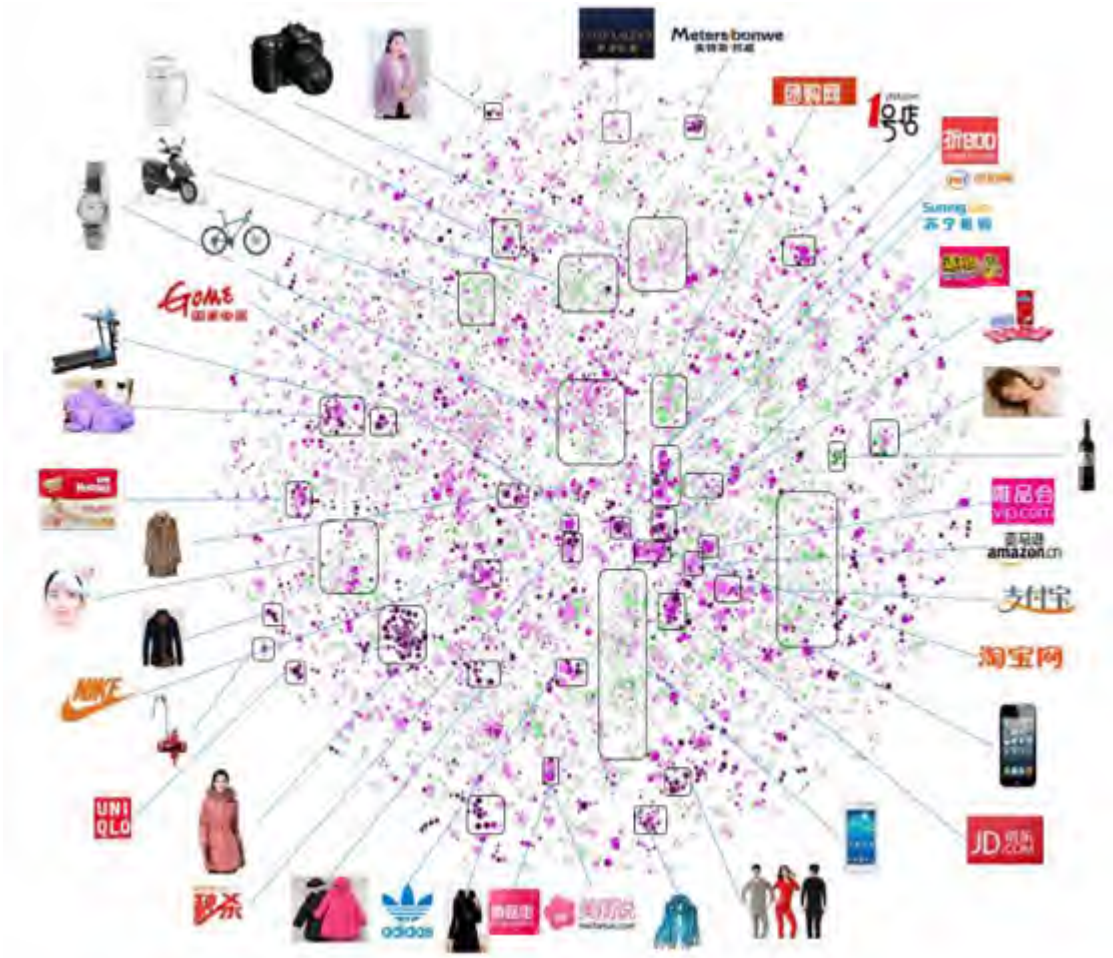
- 概念举例

ID	查询抽样	最高频查询	关键短语
265211	苹果批发价, 苹果批发价格, 红富士批发价格	苹果批发价	苹果批发价, 苹果批发价格
195748	苹果配件批发, 苹果手机配件批发, 苹果手机配件批发网	苹果手机配件批发网	苹果配件批发
403304	减肥抽脂, 抽脂手术, 吸脂减肥的价格, 吸脂整形	吸脂	抽脂减肥, 吸脂手术, 抽脂手术, 吸脂减肥
1399473	1111购物狂欢节, 双11天猫, 11.11淘宝	双十一	双十一, 天猫双十一, 双11, 双十一网购

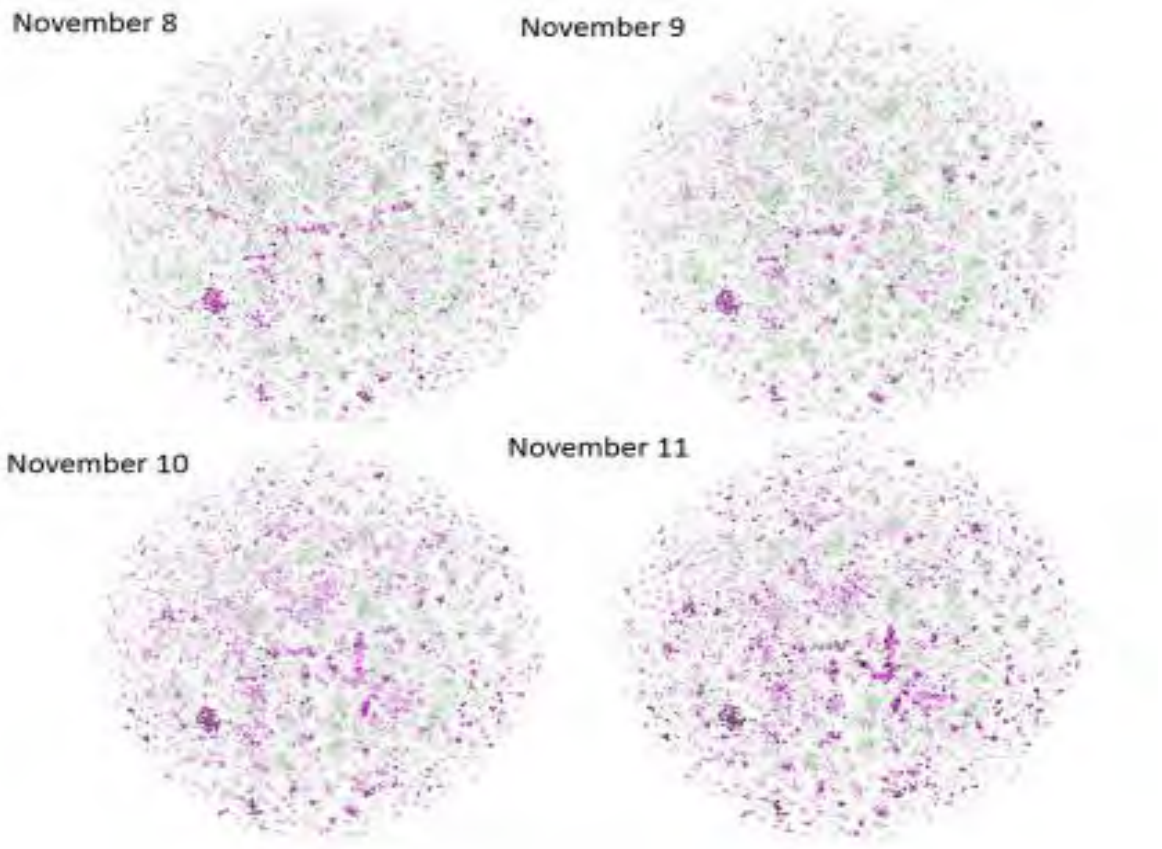
精确性与覆盖率

- 星辰系统的精确性与覆盖率
  - 统计查询次数
    - 精确性 97.4%
    - 覆盖率 61.3%

查询意图追踪


Metersbonwe
团购网
1号店
Suning
苏宁易购
GOME
国美中国
唯品会
vip.com
amazon.cn
支付宝
淘宝网
NIKE
UNIQLO
JD.COM
adidas

查询意图追踪



线上DEMO系统

类目	结果
概念	540545
商业性	1
top query	撸啊撸
核心短语	lol盒子 lol 英雄联盟 英雄联盟盒子 多玩
相关概念	lol.duowan.com lol多玩视频站 lol盒子视频 盒子

线上DEMO系统


百丽正品凉鞋2010款
哈森官方旗舰店女靴
圣思熙女鞋正品专柜
意尔康女鞋中筒靴短靴
莎诗特女鞋
意尔康皮鞋女鞋冬款
厂家10元时尚女鞋批发
真皮马丁靴短靴达芙妮
伊伴女鞋官方旗舰店
百丽官方旗舰店女鞋
百丽新款
圣思熙
佰莫文
2012指物猫正品单鞋
达芙妮女鞋断码清仓
达芙妮生皮靴子中筒靴
月芽儿官方旗舰店
品牌女鞋2013新款冬鞋
百丽+2010凉+鞋
达芙妮
广州女鞋品牌大全
卡美多
达芙妮专柜正品
百丽
卓之雅秀
冬季女鞋
百丽+凉鞋+专柜正品
意尔康皮鞋专卖店女鞋
意尔康鞋业集团有限公司
达芙妮短靴
指物猫女鞋官方旗舰店
鞋子批发网
指物猫
巨圣女鞋新款靴子
中国女鞋之都网站
约修亚女鞋
品牌女鞋有哪些牌子
春达皮鞋女鞋
百丽女鞋2014新款靴子
女鞋
戈美其
佰纳皮鞋
达芙妮女鞋折扣店
月芽儿品牌女鞋官网
女鞋新款
思加图2010凉+鞋
吉尔达皮鞋专卖店
妙丽官方旗舰店
自由星期女鞋
卓诗尼
国际女鞋十大品牌排名
百丽凉+鞋+2010年新款
达芙妮女鞋单鞋
名牌女鞋
佰纳鞋业有限公司
真皮女鞋
百丽2010新款凉女+鞋
国际品牌女鞋
女鞋什么牌子好
思加图+2010+单鞋
女士皮鞋品牌大全
国际女鞋品牌大全
卓诗尼凉鞋+2010新品

线上DEMO系统


不花钱也能玩好的网游
武侠网游大全
收费网游有哪些
画面好看的网游
动作类网游
2D游戏
3d网络游戏大全排行榜
pk要的网游
网游开服表
大型手机网络游戏
最新公测网游一览表
有意思的网络游戏
网络游戏玩不了怎么办
好玩的回合游戏
玩游戏的环境
推荐个网游
适合女生玩的网游
内测游戏开服时间表
17173新游测试表
网络游戏定义
飞行网游
体育高技类网游
找个好玩的网络游戏
能打机的3d网络游戏
最流行的网络游戏排行
好玩的网游不烧钱
网络游戏2014
公测
17173公测表
国战网游
新开网游排行榜2014
3d游戏
2014最新网游
3d网络游戏排行榜2014
耐玩的网络游戏
现在比较火的网络游戏
内测网络游戏
动作网游排行榜
画质好的3d网游
好玩的网络游戏
不花钱的网游
2d网游排行榜
不烧钱的3d网游
最新网游公测
网游公测表
2014网游排行榜前十名
网络游戏
大型网络游戏
回合制网络游戏
2014最火的网络游戏
网络游戏大全
腾讯好玩的网络游戏
好玩又小的网络游戏
游戏的发展前景
网络游戏排行榜2014前十名
网游数据能改号
2014网游排行榜
游戏排行榜前十名网络游戏2014
最流行的网络游戏
回合制游戏哪个好玩
不怎么花钱的网络游戏
好玩不烧钱的网游
最火热的网络游戏
好玩花钱少的网络游戏
小型网络游戏
赚钱游戏排行榜
网游排行
客户端游戏排行榜
2014手机3d游戏网游
网络游戏的利与弊
女生最多的游戏
装备最贵的网络游戏
百度游戏排行榜
回合制网游
浪游戏排行榜
可以结婚的网游
网络游戏刀剑英雄下载
安卓网络游戏排行榜
fps网游2013排行榜
最新回合制游戏
休闲网络游戏排行榜
中国网游烧钱排行榜
网络游戏投诉中心
韩国网游
在线人数最多的网游
女孩玩的最多的网游
网游是什么
好玩的低配置网络游戏
射击类网游排行榜
国战游戏排行榜
什么游戏女玩家多
休闲类网络游戏
网游排行榜百度
2014世界网络游戏总榜
中国游戏公司排名

广告召回中的应用

- 广告召回应用
  - 训练针对概念的商业性分类器
  - 判断每个概念是否适合召回广告
  - 线下计算每个概念和关键词的相关性
  - 线下选择每个概念适合召回的关键词链
  - 线上判断query所属概念，根据概念召回

广告质量保证中的应用

- 广告质量保证
  - 线下确定概念是否适合展示广告
  - 线下确定概念不适合展示的关键词
  - 线下确定概念不适合展示的广告类别
  - 黑名单过滤

一般的技术性研究

- 增强查询分类准确性
  - 为概念中的每个查询分类
  - 确定每个概念的类别分布
  - 选择出类别纯净的概念
  - 该星辰系统置于分类器前端

Qi Ye, Feng Wang, and Bo Li. 2016. StarrySky: A Practical System to Track Millions of High-Precision Query Intents. In Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion). 961-966.

THANKS

SequeMedia 盛拓传媒 IT168.com ChinaUnix ITPUB