



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

IT168.com

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



DTCC

2016年中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2015

SequeMedia
顺和传媒

168

ChinaUnix

PUB

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

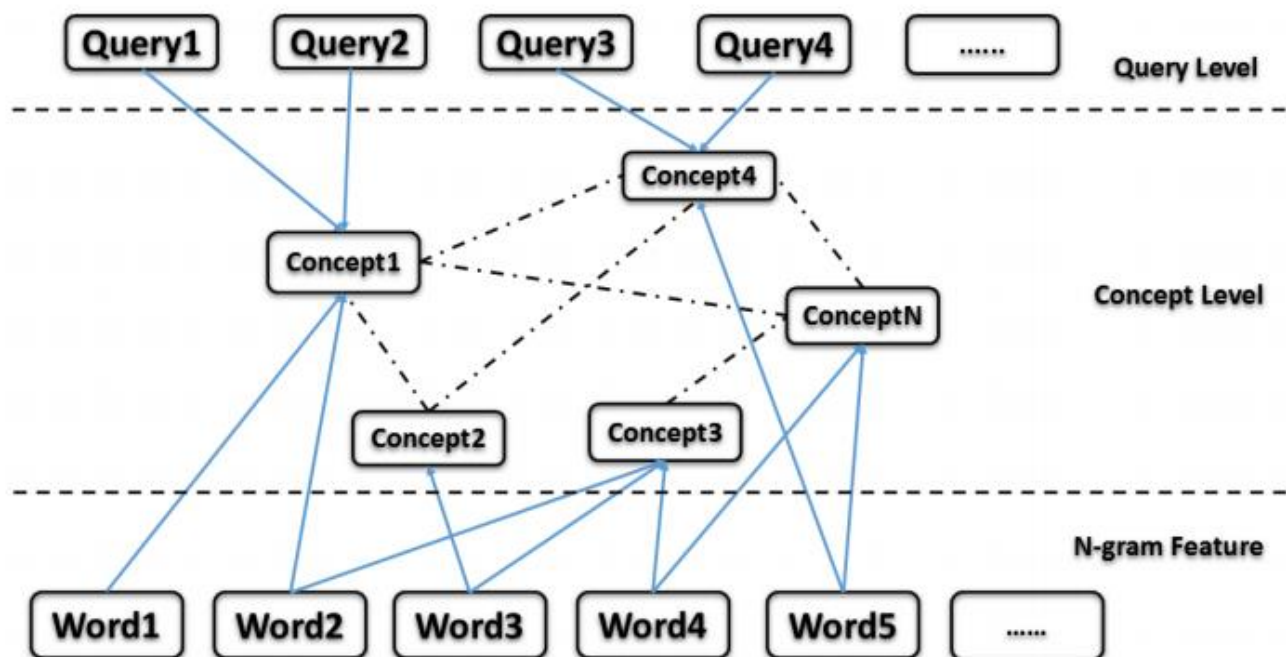


- ✕ 动机与目标
- ✕ 现有方法
- ✕ 框架与方法
- ✕ 效果与应用



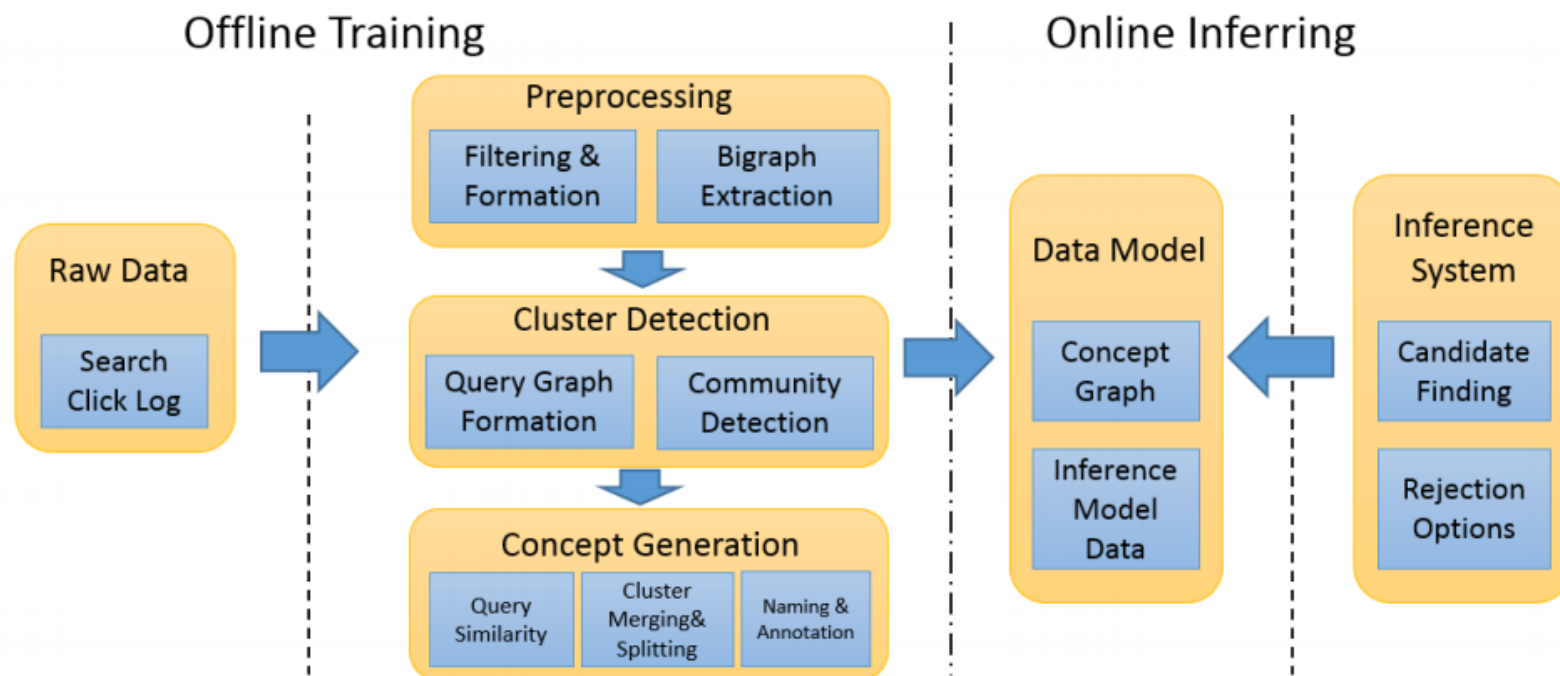
框架与方法

✕ 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。

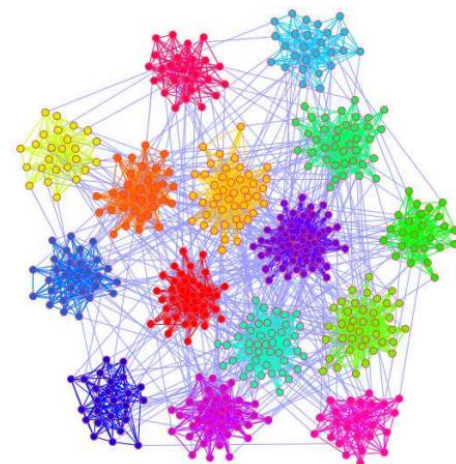
✖ 传统方法

- + 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

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

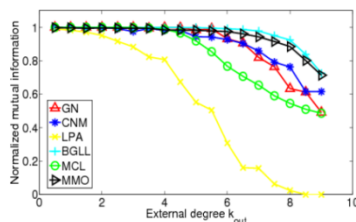
M. Girvan and M. E. J. Newman, PNAS **99**, 7821 (2002).

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

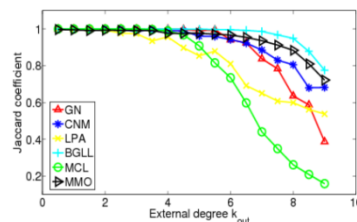


MMO算法

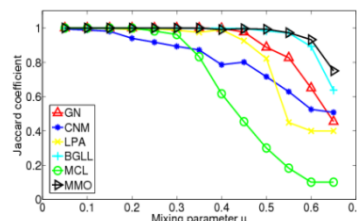
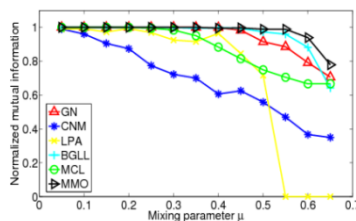
✗ MMO算法



(a) NMI (GN graphs)



(b) Jaccard (GN graphs)



✗ MMO算法的优点

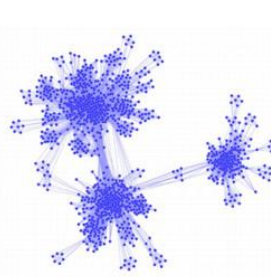
- + 易于实现
- + 时间复杂度近似线性，空间复杂度为线性。
- + 推广到 Hadoop 并行环境中的运行
- + 避免生成极大的社团

Algorithm 1 MMO algorithm

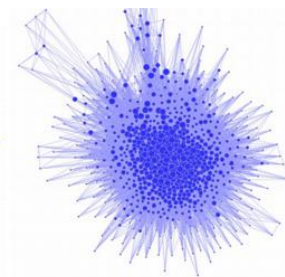
```

1:  $C = \{\{v\} | v \in V\}$ 
2: for  $i = 1$  to  $I_{max}$  do
3:    $Q_{old} = Q(C)$ 
4:   for  $v \in V$  do
5:     for  $v' \in N_v$  do
6:        $c' = C_{v'}$ 
7:        $m_{v'}^v = m_{v'}^v + 1$  {The default value of  $m_{v'}^v$  is 0}
8:     end for
9:      $c_{old} = C_v$ 
10:     $c_{best} = c_{old}$ 
11:    remove  $v$  from  $c_{old}$ 
12:     $Q_{mas} = \Delta Q(c_{best}, v)$ 
13:    for  $v' \in N_v$  do
14:       $Q' = \Delta Q(C_{v'}, v)$ 
15:      if  $Q'$  is better than  $Q_{mas}$  then
16:         $Q_{mas} = Q'$ 
17:         $c_{best} = C_{v'}$ 
18:      end if
19:    end for
20:    insert  $v$  into  $c_{best}$ 
21:  end for
22:   $Q_{new} = Q(C)$ 
23:   $\delta = Q_{new} - Q_{old}$ 
24:  if  $\delta < \epsilon$  then
25:    break
26:  end if
27:   $Q_{old} = Q_{new}$ 
28: end for
29: for  $e_{i,j}$  in  $E$  do
30:   if  $C_i \neq C_j$  then
31:     mark  $e_{i,j}$  as unlinked
32:   end if
33: end for
34: Find all the components as communities  $C$ 
35: return  $C$ 

```



(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

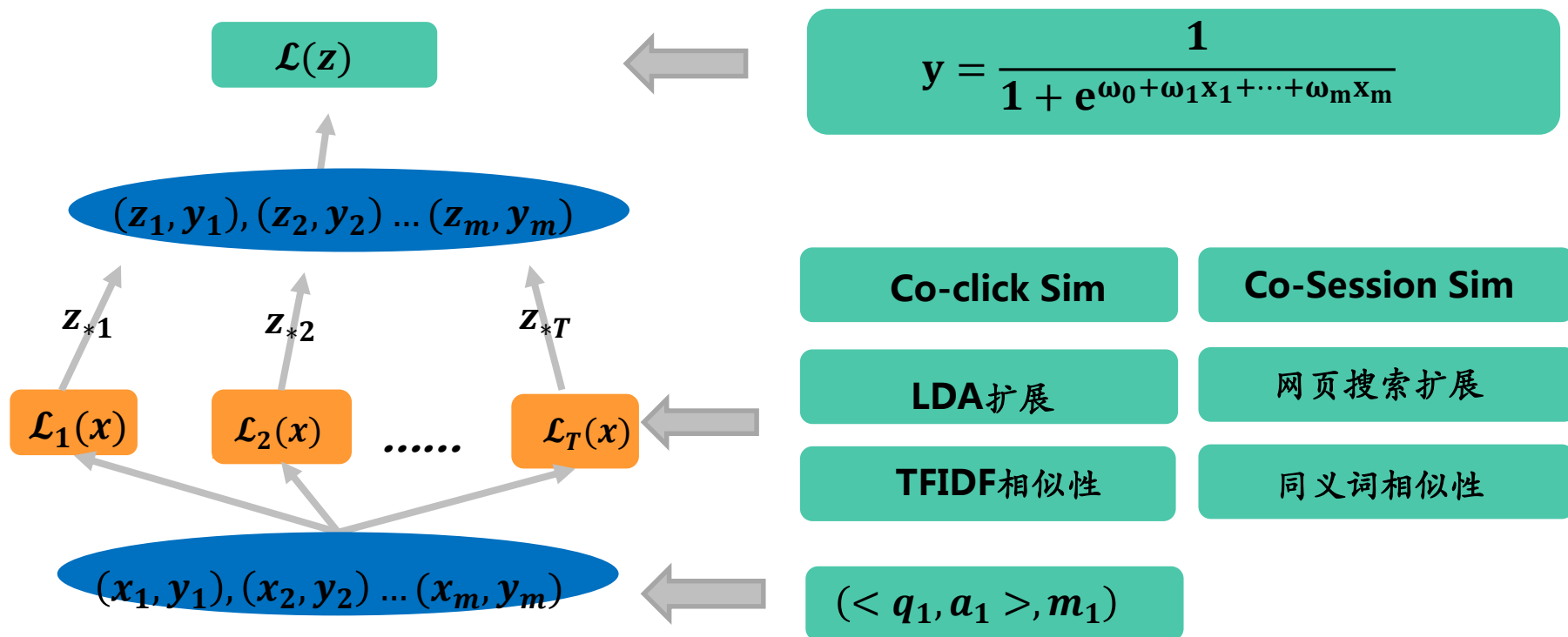
$$r(c) = \frac{\sum_{q,s \in c, q \neq s} f(q, s)}{|c| \times (|c| - 1)},$$

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



文本相关性计算方法

- 组合方法
——Stacking Learning



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

$$c = \operatorname{argmax}_{c \in C} p(c | \mathbf{x}_q)$$

$$\propto \operatorname{argmax}_{c \in C} p(\mathbf{x}_q | c) \times p(c)$$

$$= \operatorname{argmax}_{c \in C} \prod_{i=1}^n p(x_i | c) \times p(c)$$

$$\propto \operatorname{argmax}_{c \in C} \sum_{i=1}^n \log p(x_i | c) + \log p(c).$$

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

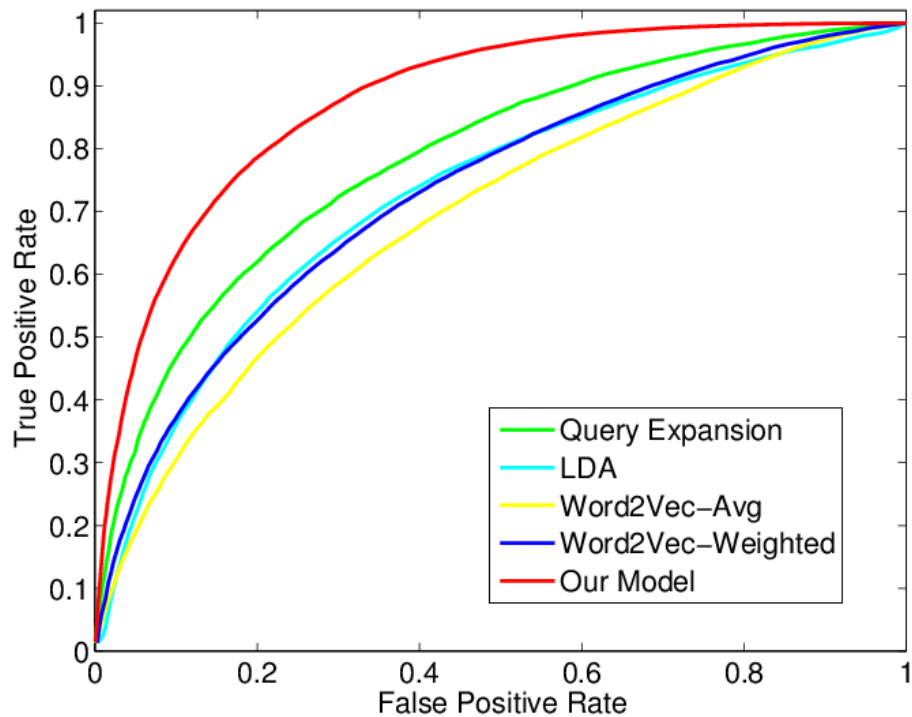
✕ 对比方法:

+ 查询扩展

+ w2v

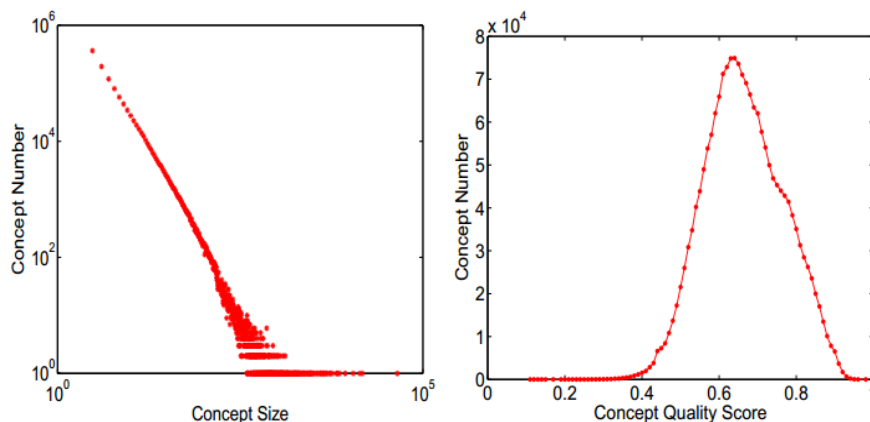
+ 字面匹配

+ LDA



概念发现的结果

✖ 概念分布



✖ 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

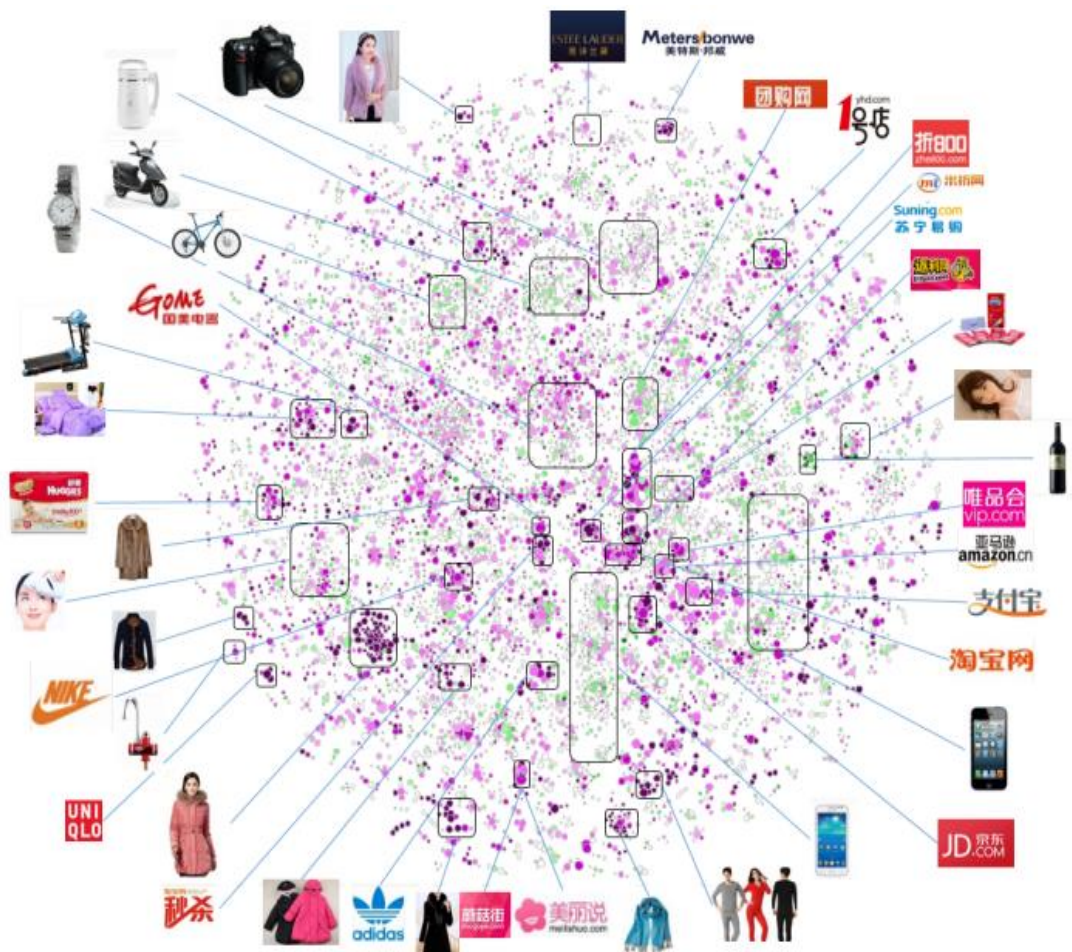
+ 统计查询次数

✕ 精确性 97.4%

✕ 覆盖率 61.3%

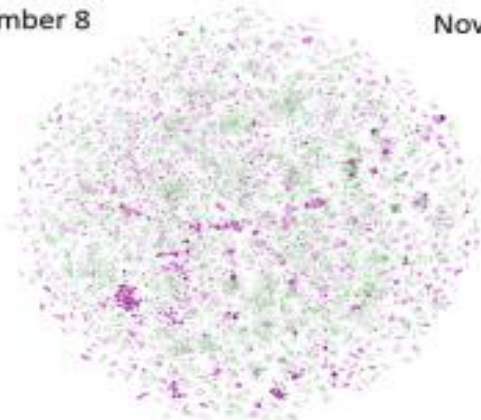


查询意图追踪

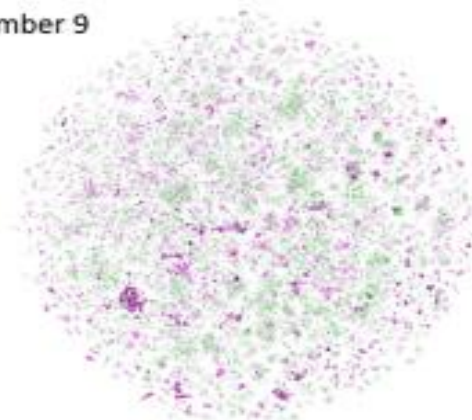


查询意图追踪

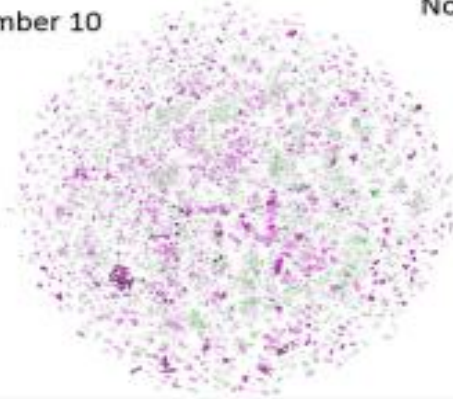
November 8



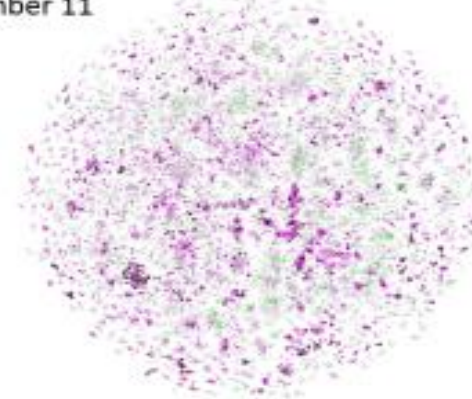
November 9



November 10



November 11

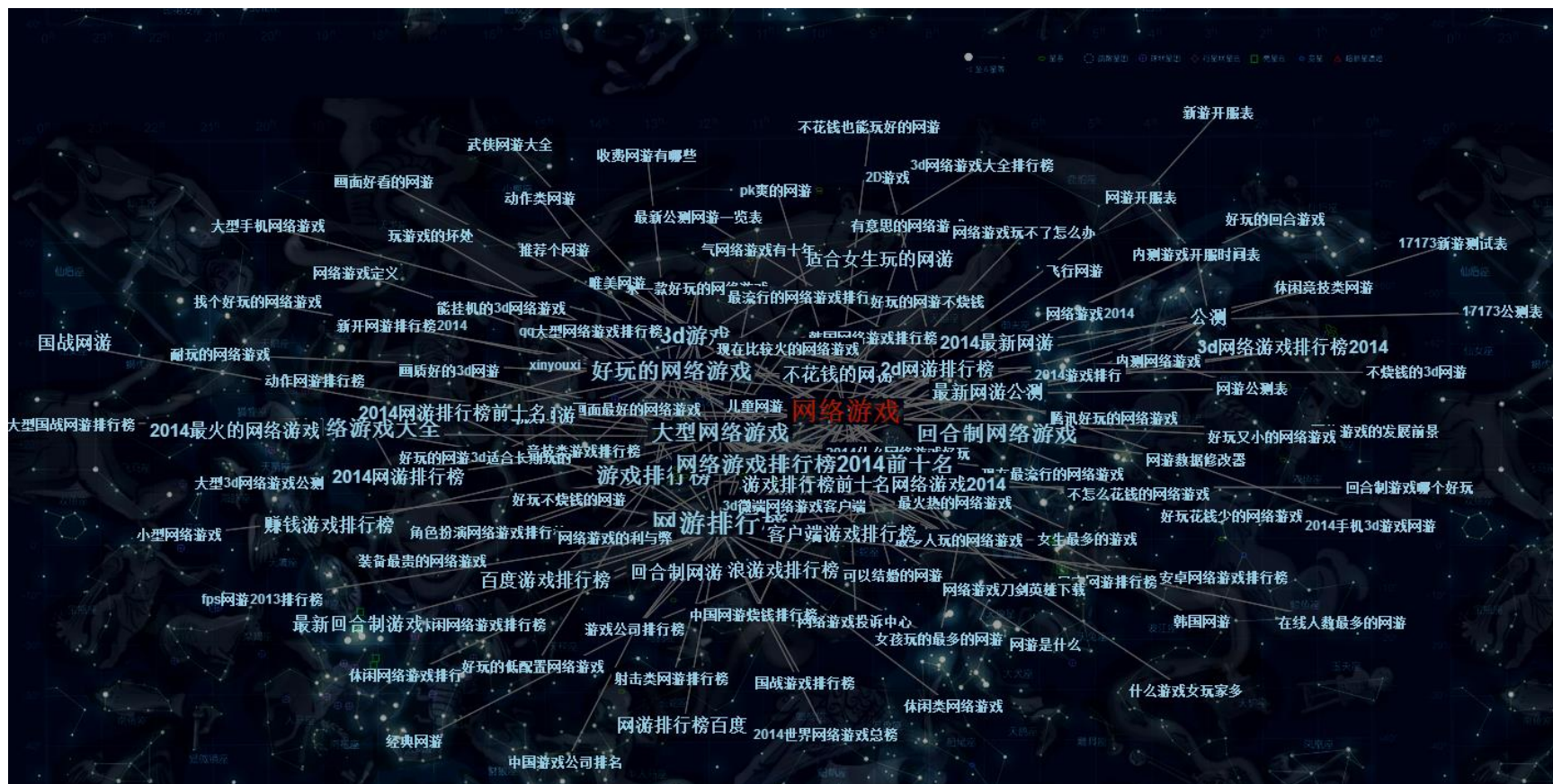


线上DEMO系统



[illegible]

线上DEMO系统



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.



SequeMedia
盛拓传媒