QCon

计算广告技术之——大数据下短文本相关性计算

王峰　wangfeng@sogou-inc.com

2015-04-25

搜狗搜索

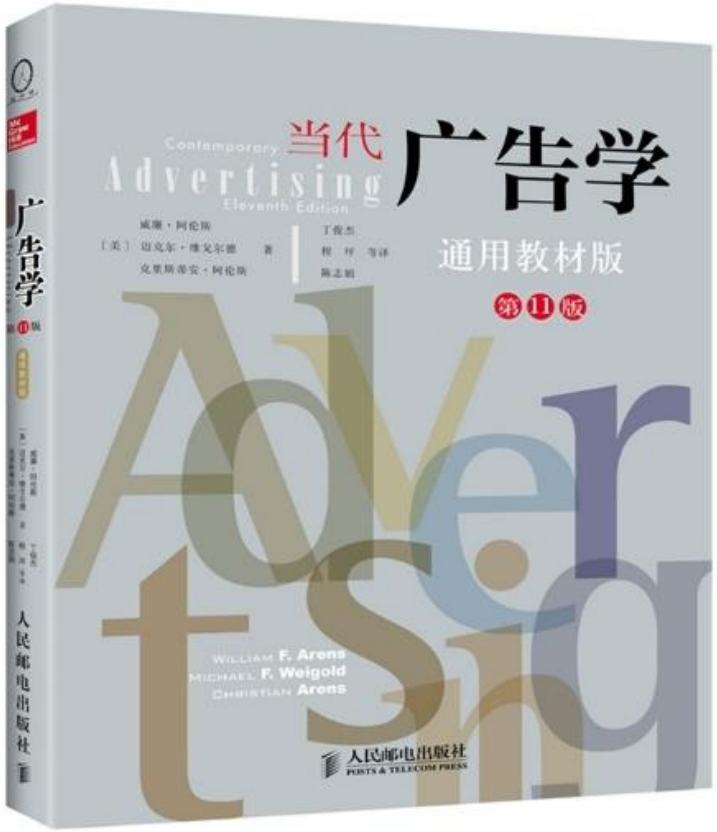
目录

CONTENTS

PART ONE

背景与挑战

广告

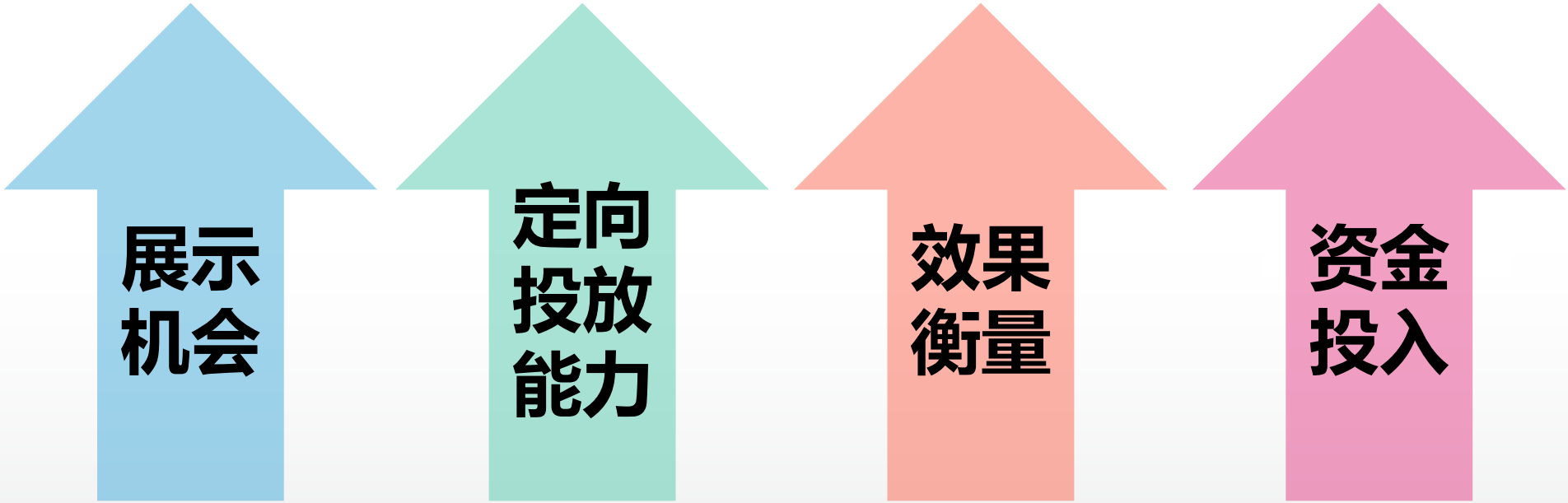


广告是由已确定的出资人(SPONSOR)通过各种媒介进行的有关产品（商品、服务和观点）的，通常是有偿的、有组织的、综合的、劝服性的非人员的信息传播活动。

广告的根本目的，是**广告主**通过**媒体**达到低成本的**用户**接触。

——《当代广告学》 WILLIAM F. ARENS

搜索引擎广告VS传统广告



搜索广告技术

目标

-在给定用户输入的查询以及用户查询上下文的情况下，找出“最佳”的广告展示。

挑战

-如何评价“最佳”？

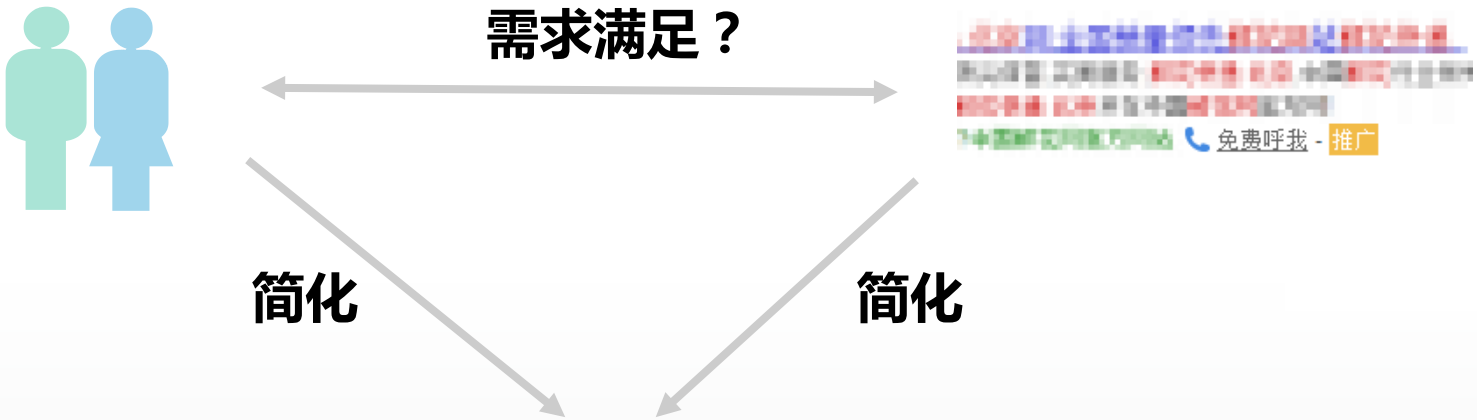
- 用户——相关性
- 搜索引擎——RPM
- 广告主——ROI

核心

-短文本相关性

短文本相关性

- 问题抽象



MatchScore(query, ad_key)∈[0,1]

反映用户查询需求被广告满足的概率有多大，需求满足程度越高，相关性越高，MatchScore越趋近于1，反之越趋近于0。

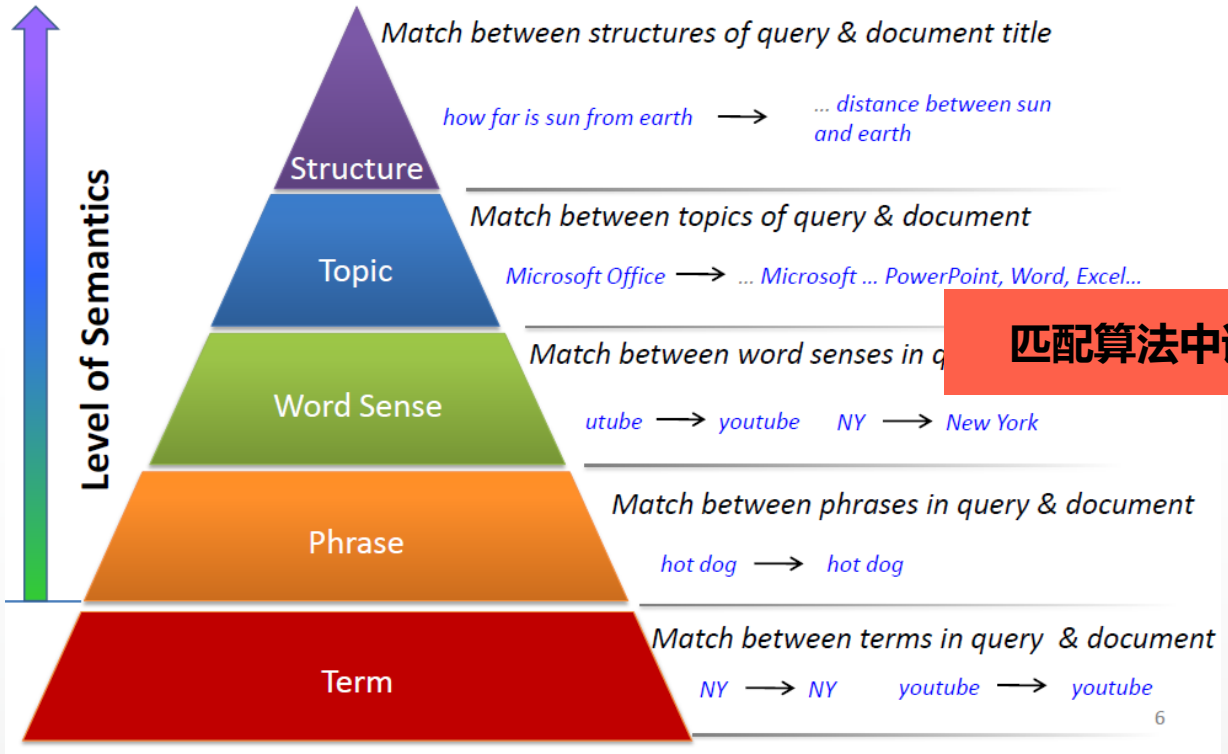
短文本相关性

- 困难与挑战

PART TWO

相关性计算方法

相关性计算方法



匹配算法中语义使用的层次

Li Hang et al, SIGIR 2012 Tutorial

相关性计算方法

解决思路

- 短文本理解——基于外部大数据辅助计算

按数据来源分类

- 文本方法
- 短文本扩展方法
- 基于点击数据
- 组合方法

相关性计算方法

文本相似性

- **分词——Bag of Words**

$$Query = Q(q_1, q_2, \dots, qm)$$

$$Ad = A(a_1, a_2, \dots, an)$$

- **计算方法**

$$Jaccard(Q,A) = \frac{|Q \cap A|}{|Q \cup A|}$$

$$Cosine(Q,A) = \frac{|Q \cdot A|}{|Q||A|}$$

$$Match(Q,A)$$

相关性计算方法

文本相关性改进

- 分词词权
- 同义词
- 相关词矩阵

相关性计算方法

- **短文本扩展方法—网页搜索扩展**

短文本

1. Issue x as a query to a search engine S.
2. Let $R(x)$ be the set of (at most) n retrieved documents $d_1, d_2, \ldots, d_n$
3. Compute the TFIDF term vector v_i for each document $d_i \in R(x)$
4. Truncate each vector v_i to include its m highest weighted terms
5. Let $C(x)$ be the centroid of the L_2 normalized vectors v_i:

$$C(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{\|v_i\|_2}$$

6. Let $QE(x)$ be the L_2 normalization of the centroid $C(x)$:

$$QE(x) = \frac{C(x)}{\|C(x)\|_2}$$

Mehran Sahami et al, WWW 2006

相关性计算方法

- 短文本扩展方法—网页搜索扩展

Query

Ad-Keyword

$$QE(Q) = (w_1, w_2, \ldots, w_m)$$

$$\text{Match}(\text{Q}, A) = QE(\text{Q}) \cdot QE(\text{A})$$

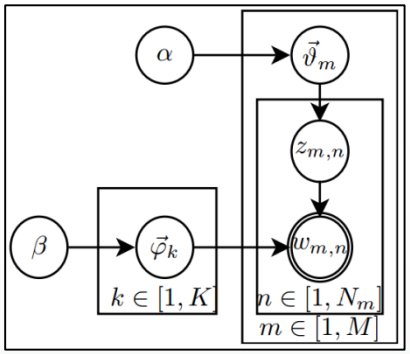
$$QE(A) = (w_1, w_2, \ldots, w_n)$$

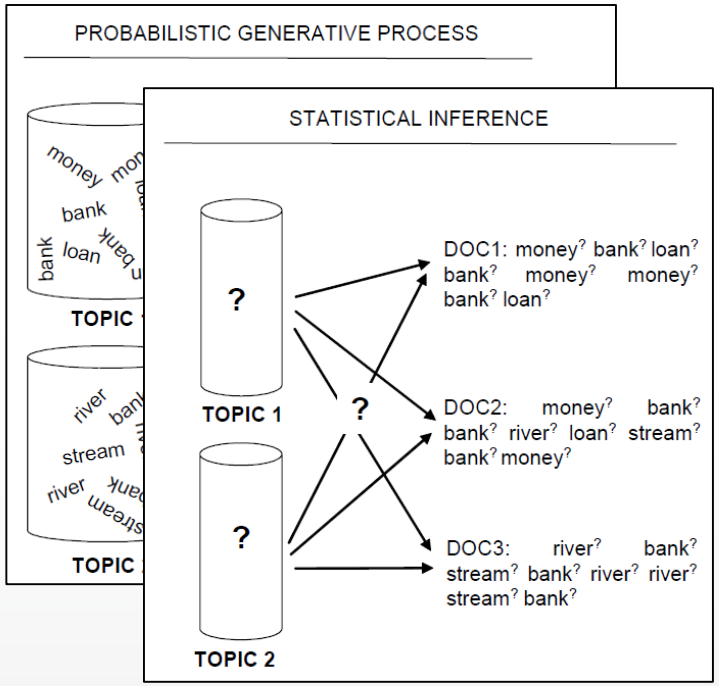
Mehran Sahami et al, WWW 2006

相关性计算方法

- **短文本扩展方法—主题模型：Topic Modeling**

Corpus





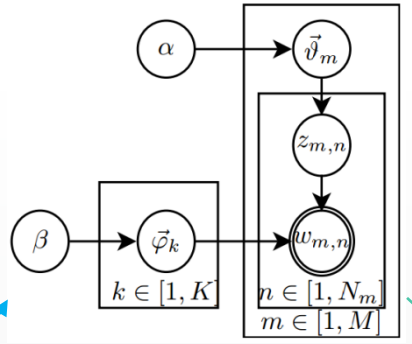
Mark Steyvers et al, 2007

相关性计算方法

- **短文本扩展方法—主题模型：Topic Modeling**

Query

Ad-Keyword



$$P(Z|Q) = (z_1^Q, z_2^Q, \dots, z_k^Q)$$

$$Match(Q, A) = \frac{1}{D_{JS}(P(Z|Q), P(Z|A))}$$

$$D_{JS}(Q||A) = \frac{1}{2}\left[D_{KL}\left(Q||\frac{Q+A}{2}\right) + D_{KL}\left(A||\frac{Q+A}{2}\right)\right]$$

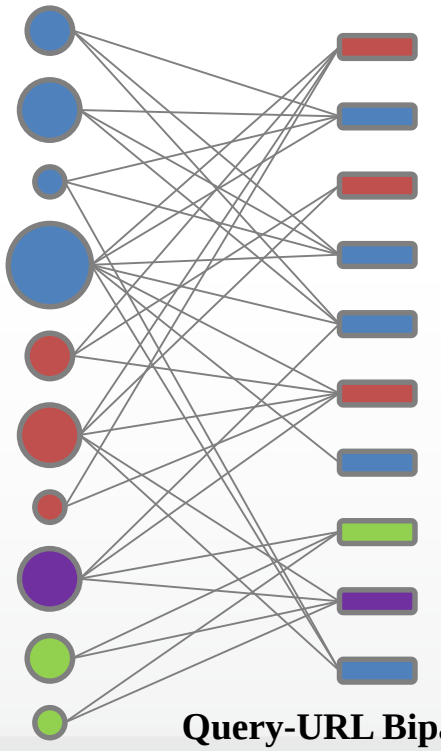
$$D_{KL}(Q||A) = \sum_1^k z_i^Q [log_2 z_i^Q - log_2 z_i^A]$$

$$P(Z|A) = (z_1^A, z_2^A, \dots, z_k^A)$$

Mark Steyvers et al, 2007

相关性计算方法

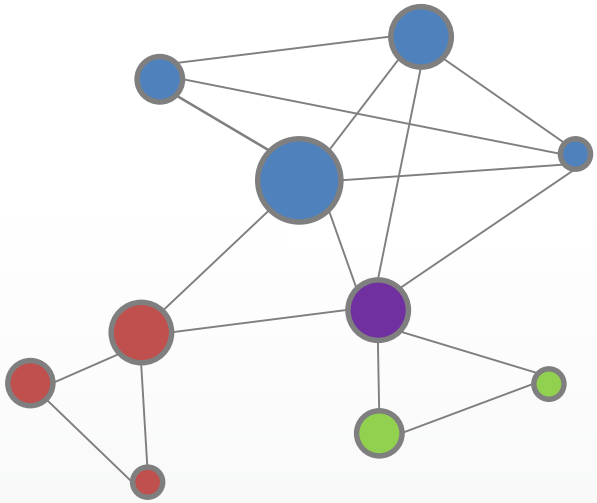
- 基于点击数据—Co-click Graph



Query-URL Bipartite-Graph

$$W(q_1, q_2) = \left| \frac{\Gamma(q_1) \cap \Gamma(q_2)}{\Gamma(q_1) \cup \Gamma(q_2)} \right|$$

$$Match(Q, A) = W(Q, A)$$

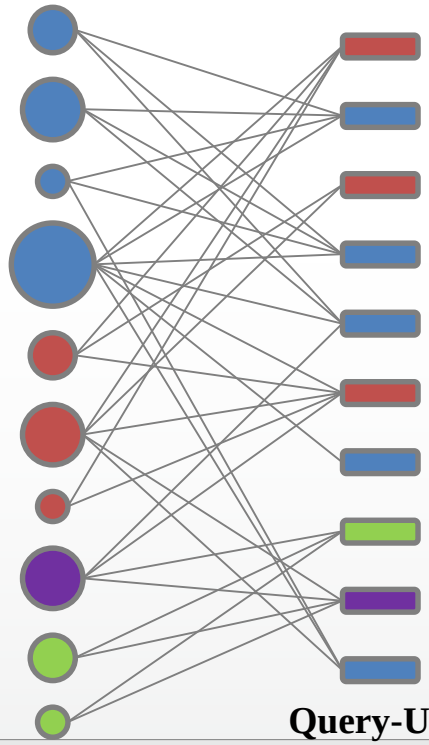


Query Co-click Graph

David Liben-Nowell et al, CIKM 2003

相关性计算方法

- 基于点击数据—Bipartite-Graph SimRank



Query-URL Bipartite-Graph

SimRank:

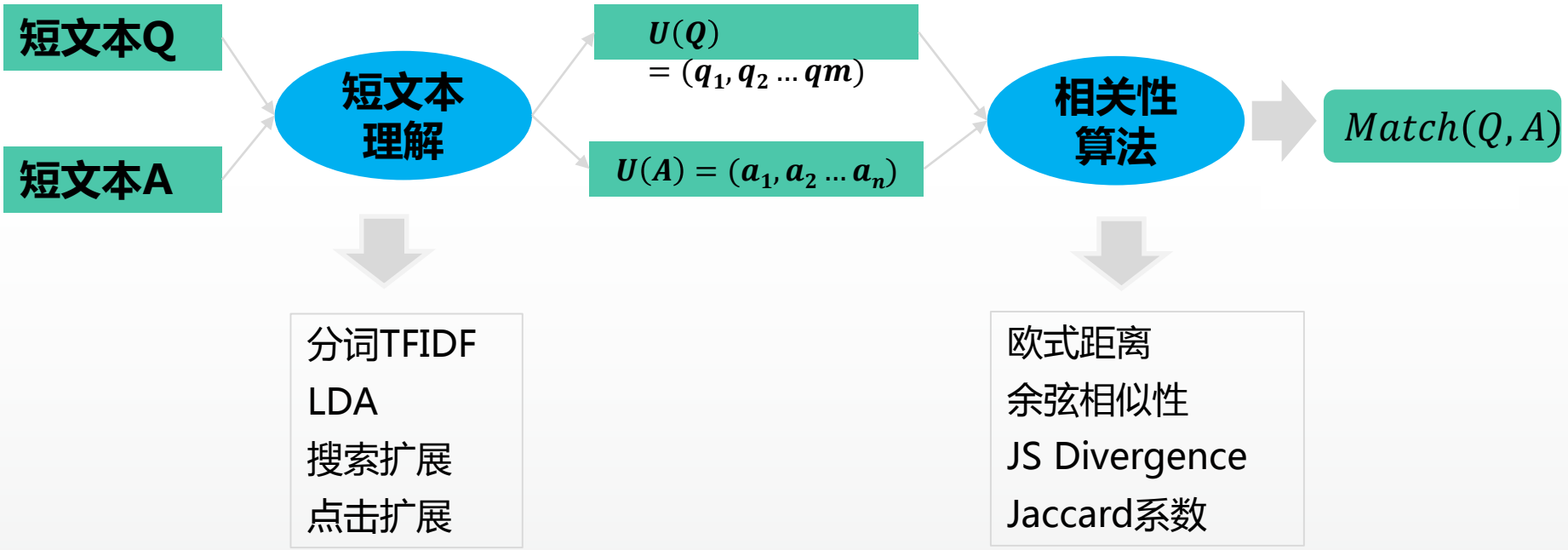
$$Sim_k(q,q') = \frac{C_1}{|E(q)||E(q')|} \sum_{i_u \in E(q)} \sum_{j_u \in E(q')} Sim_{k-1}(i_u, j_u)$$

$$Sim_k(u,u') = \frac{C_1}{|E(u)||E(u')|} \sum_{i_q \in E(u)} \sum_{j_q \in E(u')} Sim_{k-1}(i_q, j_q)$$

Ioannis Antonellis, et al, WWW 2004

相关性计算方法

- 小结


短文本Q
短文本A
短文本理解
$U(Q) = (q_1, q_2 \dots qm)$
$U(A) = (a_1, a_2 \dots a_n)$
相关性算法
$Match(Q, A)$
分词TFIDF
LDA
搜索扩展
点击扩展
欧式距离
余弦相似性
JS Divergence
Jaccard系数

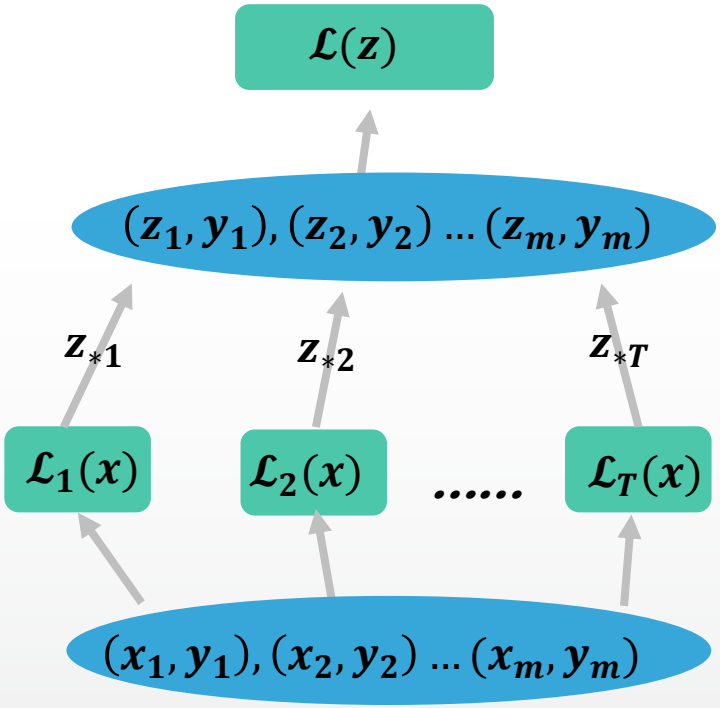
相关性计算方法

- 小结

查询理解	优点	缺点
文本方法	扩展性强，可覆盖长尾查询	词义歧义影响较大，准确率低
短文本扩展	准确率较高。LDA可用推演等方法增加算法扩展能力	覆盖率低，网页搜索扩展数据维护成本高
点击数据	准确率最高	覆盖率最低，扩展能力最差

相关性计算方法

- 组合方法
  ——Stacking Learning



Input: Data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$;
First-level learning algorithms $\mathcal{L}_1, \cdots, \mathcal{L}_T$;
Second-level learning algorithm $\mathcal{L}$.

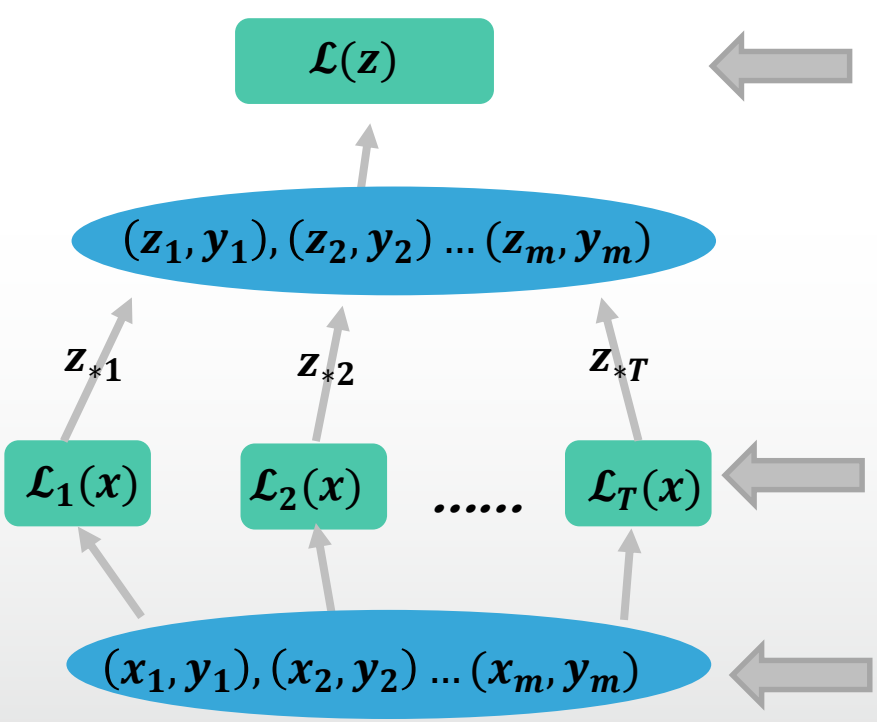
Process:
for $t = 1, \cdots, T$:
  $h_t = \mathcal{L}_t(\mathcal{D})$ % Train a first-level individual learner h_t by applying the first-level
end; % learning algorithm $\mathcal{L}_t$ to the original data set $\mathcal{D}$
$\mathcal{D}' = \emptyset$; % Generate a new data set
for $i = 1, \cdots, m$:
  for $t = 1, \cdots, T$:
    $z_{it} = h_t(\boldsymbol{x}_i)$ % Use h_t to classify the training example $\boldsymbol{x}_i$
  end;
  $\mathcal{D}' = \mathcal{D}' \cup \{((z_{i1}, z_{i2}, \cdots, z_{iT}), y_i)\}$
end;
$h' = \mathcal{L}(\mathcal{D}')$. % Train the second-level learner h' by applying the second-level
% learning algorithm $\mathcal{L}$ to the new data set $\mathcal{D}'$

Output: $H(\boldsymbol{x}) = h'(h_1(\boldsymbol{x}), \cdots, h_T(\boldsymbol{x}))$

Zhou,et al, Ensemble Methods 2012

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

$(< q_1, a_1 >, m_1)$

PART THREE

举例与应用

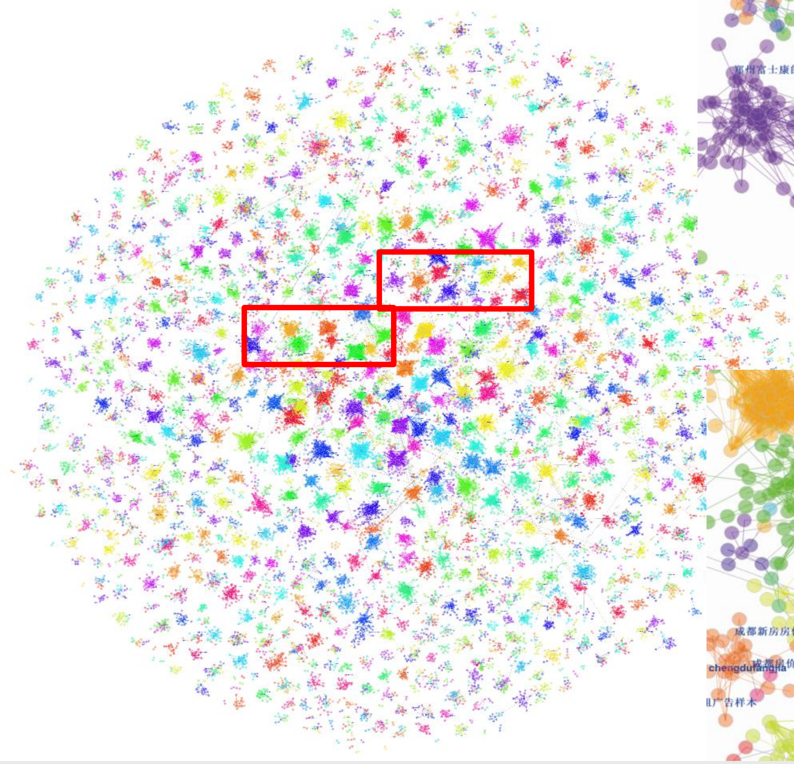
举例与应用

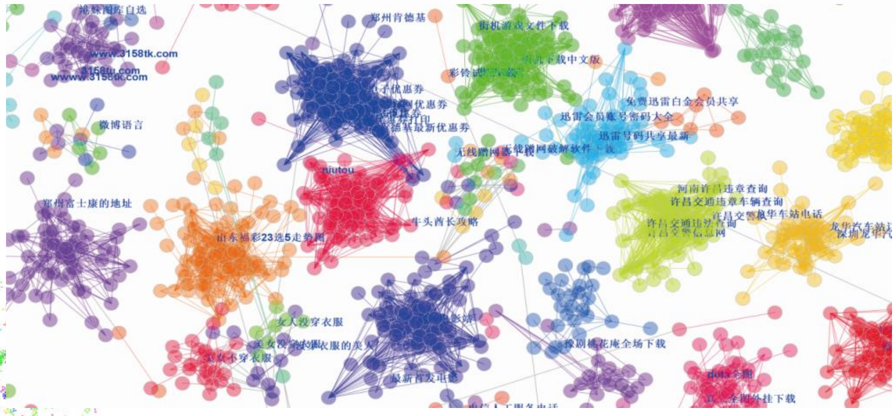
- **相关性举例**

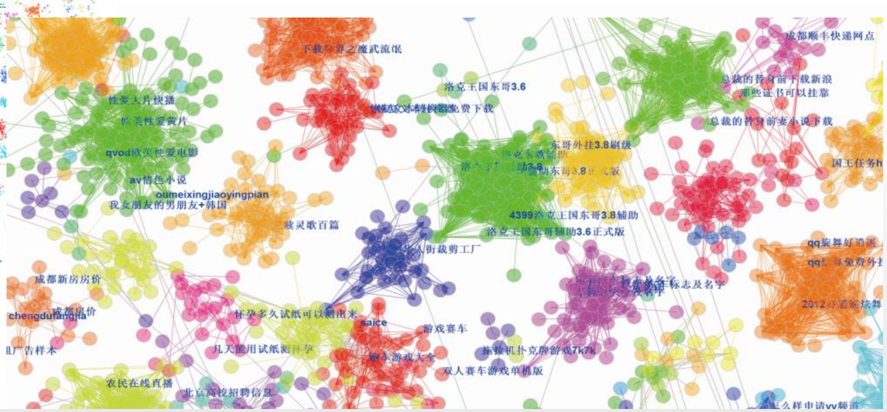
	Query	KWD	SimScore
1	摩托罗拉	笔记本维修	0.15
2	笔记本	笔记本维修	0.45
3	修理电脑	笔记本维修	0.56
4	汽车修理技术	汽车驾校	0.21
5	托福培训	新东方	0.71
6	厨师培训	新东方	0.51

举例与应用

- **相关性举例**

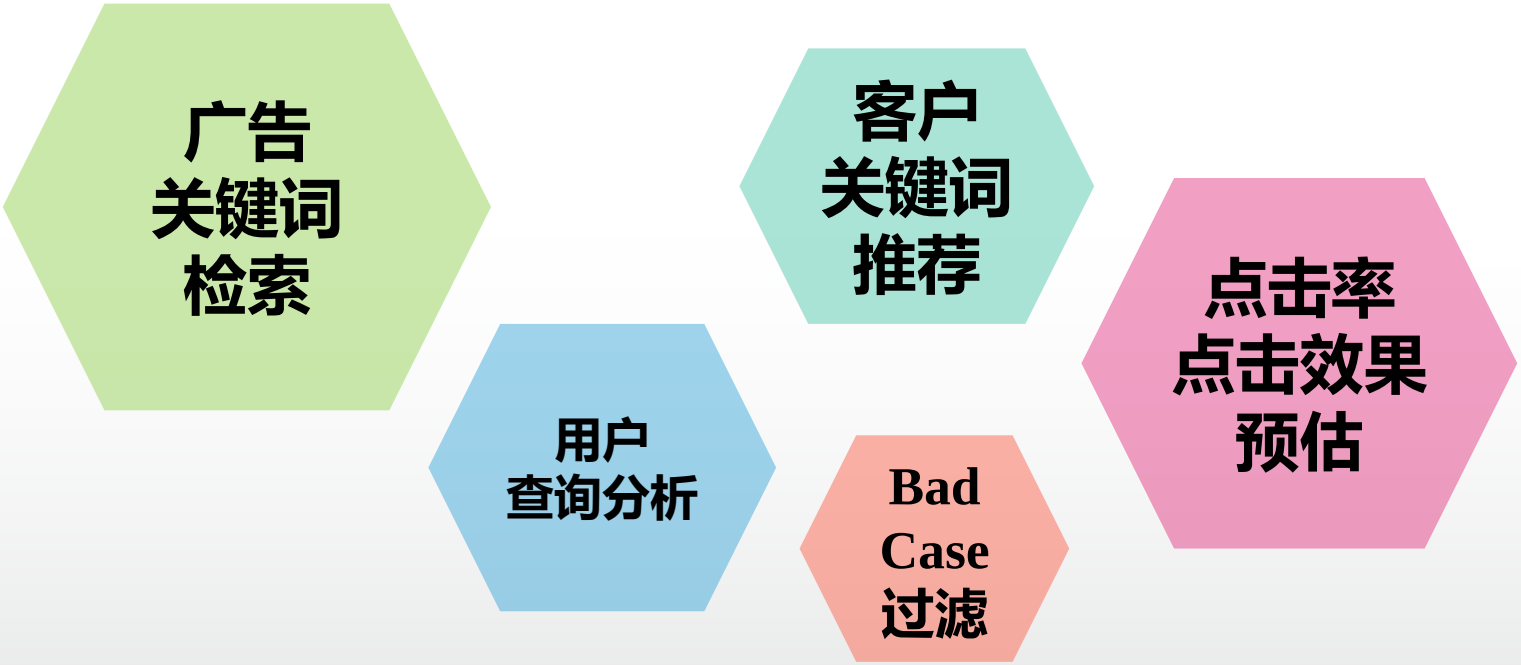






举例与应用

- 在广告系统中的应用



举例与应用

- 短文本相关性之外

PART FOUR

未来展望

未来展望

两个可能方向

- 基于实体、语法分析的推理
- Deep Learning的应用

QCon

www.qconferences.com

THANKS! Q&A