第36卷 第3期
2013年3月

计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

Vol. 36 No. 3
Mar. 2013

# 基于查询意图的长尾查询推荐

白 露 郭嘉丰 曹 雷 程学旗

(中国科学院计算技术研究所网络数据科学与工程研究中心 北京 100190)

**摘 要** 查询推荐是一种提升用户搜索效率的重要工具.传统的查询推荐方法关注频度较高的查询,但对于那些频度较低的长尾查询,由于其信息的稀疏性而难以产生好的推荐效果.另外,传统的方法由于没有考虑查询意图对推荐结果的影响,故对长尾查询的推荐会受到查询中噪声单词的影响.该文提出了一种新的关于词项查询图(term-query graph)概率混合模型,该模型能够准确地发掘出用户的查询意图.另外,文中还提出了一种融合查询意图的查询推荐方法,该方法可以将新查询中单词的推荐结果按查询意图自然地融合起来,从而避免了噪声单词对推荐结果的影响.实验结果表明,通过考虑查询意图,可以显著提高长尾查询推荐的相关性.

**关键词** 查询推荐;长尾查询;概率混合模型;查询意图;词项查询图

**中图法分类号** TP391 **DOI号** 10.3724/SP.J.1016.2013.00636

## Long Tail Query Recommendation Based on Query Intent

BAI Lu GUO Jia-Feng CAO Lei CHENG Xue-Qi

(*Research Center of Web Data Science & Engineering, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

**Abstract** Query recommendation is an important tool for improving searching efficiency. Traditional recommendation methods were mainly care about the frequent queries, but cannot provide good recommendations for long tail queries due to the information sparsity. Without consideration of query intents, traditional methods generated the recommendations for long tail queries, which can be greatly influenced by noise words in queries. A novel probabilistic mixture model of term-query graph was proposed in this paper, which can clearly identify query intents of users. Otherwise, a new method of assembling query intents into recommendation was introduced in the paper, which can prevent the influence from noise words by merging the recommendations of word in newcoming query according to query intents naturally. The result of experiments show the relevance of long tail query recommendation can be greatly improved by taking account of query intent.

**Keywords** query recommendation; long tail query; probabilistic mixture model; query intent; term-query graph

## 1 引 言

随着互联网的蓬勃发展,搜索引擎成为用户访问网络资源必不可少的工具.用户通过向搜索引擎提交查询来表达自己的信息需求.但是构造好的查询并非易事,为了帮助用户构造合适的查询,现代搜索引擎通过分析查询日志来提供查询推荐.现阶段

收稿日期:2012-06-30;最终修改稿收到日期:2012-11-08.本课题得到国家自然科学基金(60933005,61173008,61003166,61203298)、国家"九七三"重点基础研究发展规划项目基金(2012CB316303)资助.**白 露**,男,1983年生,博士研究生,主要研究方向为信息检索、文本挖掘.E-mail:bailu@software.ict.ac.cn.**郭嘉丰**,男,1980年生,博士,副研究员,主要研究方向为网络搜索和挖掘、用户数据挖掘以及社交网络.**曹 雷**,男,1977年生,博士,主要研究方向为实体排序和数据挖掘.**程学旗**,男,1971年生,博士,研究员,博士生导师,主要研究领域为网络科学、互联网搜索与挖掘、对等网络、信息安全、分布式系统.

的查询推荐研究工作，大部分是关注于那些频度较高的用户查询. 然而，已有研究工作表明[1]，用户查询的频度分布服从长尾现象("Long Tail")规律，即很大比例的用户查询的频度都很低，并且大多数用户都会向搜索引擎提交这种长尾查询；但是由于其信息的稀疏性，对于这类查询进行相关查询推荐的结果往往不能令人满意. 因而，为长尾查询进行有效的相关查询推荐，将会极大地提高 Web 搜索效率，进而提升用户查询体验.

针对长尾查询的查询推荐的研究工作开展较晚，一些方法通过单词信息扩充查询间的关系，来进行推荐. 然而这类方法没有考虑到单词在查询中的不同作用，特别是忽略了查询意图对查询推荐的影响，从而导致较差的推荐结果. 另外查询中有很多词是噪声词，过分考虑这类词会造成推荐结果偏离用户的查询意图. 实际上一个单词 $t_i$ 对于查询 $q$ 是否是噪声单词，是由查询 $q$ 的查询意图决定的. 比如有这样的两个查询"www conference deadline"和"www wiki com net"，尽管这两个查询均包含"www"这个单词，但是所起作用明显不同，前者"www"表示该查询是关于国际互联网大会，而在"www wiki com net"查询中，"www"却是无关紧要的噪声词.

基于以上考虑，本文提出了一种关于词项查询图的概率混合模型. 具体地讲，该模型用概率的方法建模查询意图，并解释了词项查询图中每个查询、每个单词以及每条边的产生过程. 通过该模型，我们可以清晰地表示出蕴藏于词项查询图中的不同查询意图. 另外，我们还提出一种新的长尾查询的推荐方法. 该方法通过查询意图来衡量单词在查询中所起的作用，将不同单词的推荐结果按查询意图整合起来. 实验数据表明我们的方法能够推荐出更加相关的查询.

## 2　相关工作

概率混合模型是近年来出现的一种比较流行的数据分析方法. 借助其清晰的定义和解释，概率混合模型被用到多个研究领域，如文本分析[2]、协同过滤[3]、社会网络分析[4]等等. 其中比较有代表性的工作是 Hofmann[5] 提出的 PLSA 模型和 Blei 等人[6]提出的 LDA 模型. 最近人们开始尝试用混合模型去解决社区发现的问题，并取得了一定的效果. 其中 Newman 等人[7]提出的社区发现算法，能够发现可部分重叠的网络社区. Ramasco 等人[8]基于 Newman 的方法提出了一种可以即建模单向和双向边的社区发现算法. Ren 等人[9]从 PLSA 的角度出发，给出了一种概率化的社区发现算法，该算法假设有边相连的两个结点来自于同一个社区. 受以上工作影响，我们构建了一种关于词项查询图的概率混合模型，该模型不仅能够建模查询和查询之间，以及查询和单词之间的关系，而且能够发掘出用户的查询意图.

查询推荐一直是检索领域中一个重要问题. 面对不同的推荐需求，人们提出了各种各样的模型和算法. 比如 Mei 等人[10]使用 hitting time 在 click-though 图上推荐语义相关查询. Zhu 等人[11]通过利用带停止点的流行排序来推荐多样性查询. Guo 等人[12]提出了一种结构化的推荐方法，该方法将推荐进行聚类来帮助用户更好地理解推荐. 最近由 Boldi 等人[13]提出的查询流图(query-flow graph)也被用于查询推荐，他们用在查询流图上进行个性化随机游走的方法进行查询推荐. 在 Boldino 等人[14]的工作中，他们将查询流图映射到低维的欧式空间中进行查询推荐.

长尾查询的推荐是近年来人们比较关注的问题. 尽管这类查询中单个查询发生的频率较低，但其整体却占所有查询中不小的一部分. Song 等人[15]利用伪反馈的方法扩展长尾查询的信息，通过 url 信息将 click graph 和 skip graph 的推荐结果融合在一起. Pandey 等人[16]对长尾查询中的广告效用进行了分析. Szpektor 等人[17]利用模板和额外的本体信息对长尾查询进行推荐. 在 Bonchi 等人[18]的工作中，长尾查询被分解到单词，然后通过随机游走推荐查询. 受该项工作的启发，我们也使用了查询中的单词信息，但我们的方法建模了查询意图，因而能产生更加相关的推荐结果.

## 3　基于查询意图的长尾查询推荐方法

在这一部分，我们首先介绍词项查询图的定义，然后我们将介绍一种基于词项查询图的概率混合模型，通过该模型可以得到用户的查询意图. 最后，我们基于查询意图，用词项查询图上的个性化随机游走进行长尾查询推荐.

### 3.1　词项查询图

由于词项查询图是查询流图的一种扩展，因此我们先介绍一下查询流图. 查询流图是由 Boldi 等人[13]提出的一种新的用于挖掘查询日志信息的工

具.查询流图利用查询日志中的会话信息,将不同的查询连接起来构成一同质有向图.具体的说,如果我们在查询日志中有一查询会话集合 $S$,我们用 $G=(V,E,W)$ 表示与其对应一个查询流图,其中结点集合 $V=Q\cup\{s,t\}$,$Q$ 是集合 $S$ 中的所有查询,$s,t$ 分别表达两个特殊的虚拟结点:开始结点和终止结点.$E$ 是查询流图中边的集合,如果两个查询 $q_a,q_b$ 是同一个查询会话中的连续的两个查询,那么在查询流图中就有一条从 $q_a$ 指向 $q_b$ 的有向边.$W$ 是边的权重集合.不同的应用场景权重的定义不同,这里我们仅用查询会话集合中的查询间转移频率来表示边的权重.图 1 展示了从一组查询会话中构建查询流图的过程.

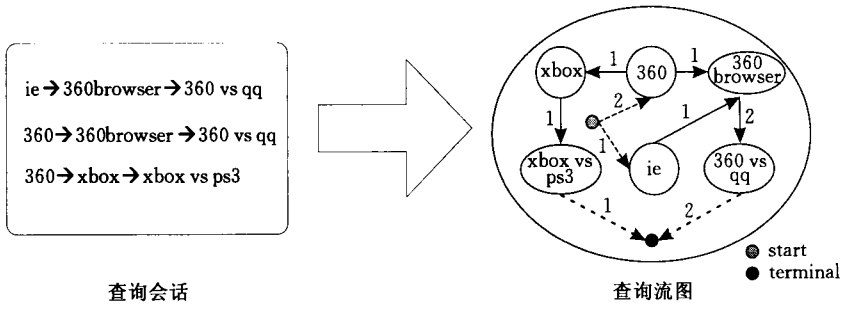


图 1 查询流图的构建

在查询流图中,查询结点之间的关系只能通过查询会话中连续共现信息来建立.实际上在查询流图中有大量的查询虽然没有边相连,但是却因为包含同样的单词而具有相似或相关的语义.于是,Bonchi 等人[18]提出一种被称为词项查询图的包含单词信息的扩展查询流图.如果还用 $G=(V,E,W)$ 表示某一查询流图,而用 $G'=\{V',E',W'\}$ 表示其对应的一个词项查询图(如图 2 所示),那么 $V'=V\cup T$,$T=\{w_1,w_2,\cdots\}$ 表示所有查询中的单词集合.$E'=E\cup E_{TQ}$,$E_{TQ}$ 表示词到查询的边集合,这样的边可定义为,如果某一查询 $q$ 包含某个词 $w_i$,那么就有一条边 $w_i\rightarrow q$.$W'=W\cup W_{TQ}$,$W_{TQ}$ 表示从单词到查询的转移频率集合.图 2 给出了图 1 对应的词项查询图.简洁起见,在词项查询图中,我们没有将起始节点 $s$ 和终止节点 $t$ 绘出.

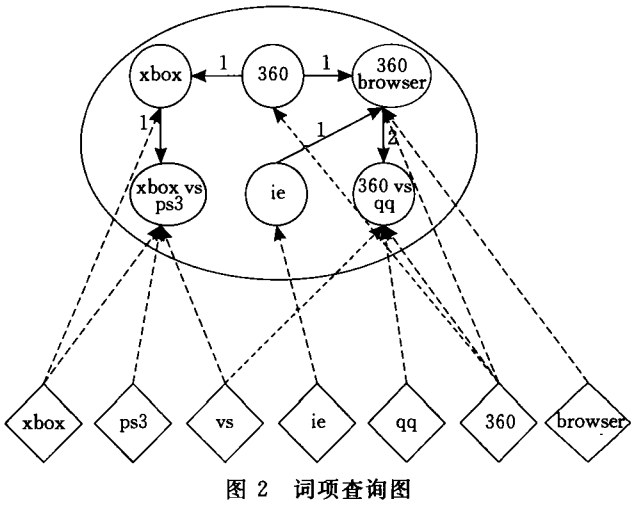


图 2 词项查询图

### 3.2 基于词项查询图的概率混合模型

这一节将具体介绍我们提出的一种新型的关于词项查询图的概率混合模型.实际上,我们的模型基于如下几点假设:(1)每一查询都是产生于某种查询意图.(2)在某一查询会话中,相邻的两个查询来自于同样的查询意图.(3)组成某个查询的所有单词都来自同样的查询意图.这几条假设比较简单直观,符合人们对查询及查询间关系的理解.具体的说,每一查询都是产生于某种查询意图,查询中单词蕴含了该种查询意图,用户在查询提交给搜索引擎之后,如果对搜索结果不满意,将会提交不同的查询来重新表述其查询意图.因而,查询会话中连续的查询是用户对其某一查询意图的多种表述.基于这样的认识,我们可以分析词项查询图中查询与查询、查询和单词间的关系来挖掘出查询意图和单词的相互关系.同传统的话题模型一样,这里我们也采用单词的分布来表达用户的查询意图.但是,相对于传统的话题模型(LDA[6],PLSA[5]),我们所学得的单词分布不仅能反映出单词的共现语义性质,而且能通过建模词项查询图中查询和查询之间的相互关系来影响查询中单词的产生方式,从而达到建模查询意图的目的.

图 3 展示了产生词项查询图的概率图模型.具体的说,如果我们有一词项查询图 $G'$,其中查询节点的个数是 $N$,查询和查询之间的边的数目是 $M$,单词集合是 $T$.我们假设这样的 $G'$ 是由 $K$ 个查询意图产生的,$\beta_r$ 是关于查询意图 $r$ 的 $|T|$ 维多项分布,$|T|$ 表示单词集合的大小.我们用 $e_{ab}$ 来表示 $G'$ 中连接查询 $a,b$ 的边.$g_{ab}$ 是一指示器变量,用来说明边 $e_{ab}$ 来自于哪个查询意图.$\pi$ 是整个 $G'$ 中关于查询意图的 $K$ 维多项分布.$w$ 表示在查询中出现的单词,$|a|$ 和 $|b|$ 分别表示查询 $a$ 和 $b$ 中单词的个数.方便

起见，我们用 $Q_{d,i}$ 来表示在查询 $d$ 中单词 $w_i$ 出现的次数，用 $G_{ab}$ 来表示连接节点 $a$ 和 $b$ 边的权重，使用如上的符号的定义，我们给出一条连接查询 $a$ 和 $b$ 的边 $e_{ab}$ 的产生过程：

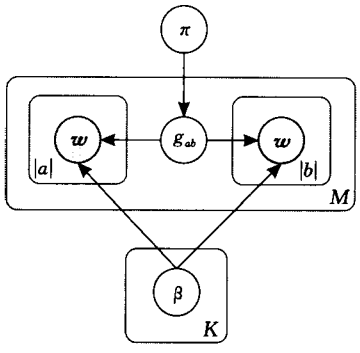


图 3　产生词项查询图的概率图模型

**过程 1.**　产生边 $e_{ab}$ 的过程.

1. 对于边 $e_{ab}$，我们首先按照 $K$ 维多项分布 $\pi$ 选取某个意图 $r$.

2. 对于查询 $a$，我们按照 $|T|$ 维多项分布 $\beta_r$ 选取 $|a|$ 个单词组成查询 $a$.

3. 对于查询 $b$，我们按照 $|T|$ 维多项分布 $\beta_r$ 选取 $|b|$ 个单词组成查询 $b$.

根据上述过程，我们可以写出产生边 $e_{ab}$ 且 $g_{ab}=r$ 的概率如下：

$$Pr(e_{ab},g_{ab}=r)=Pr(g_{ab}=r)\cdot \prod_{i:w_i\in q_a} Pr(w_i|r)^{Q_{a,i}} \prod_{i:w_i\in q_b} Pr(w_i|r)^{Q_{b,i}} = \pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}} \tag{1}$$

将变量 $g_{ab}$ 积掉，我们可以得到关于边 $e_{ab}$ 的概率：

$$Pr(e_{ab})=\sum_{r=1}^{K}\pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}} \tag{2}$$

通过建模每一条边的产生概率，我们可以得到关于整个 $G'$ 似然函数如下定义：

$$Pr(G'|\pi,\beta)=\prod_{e_{ab}\in E}\Big(\sum_{r=1}^{K}\pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}}\Big)^{G_{ab}} \tag{3}$$

从上式可以看出，$\pi$ 和 $\beta$ 是控制似然函数 $Pr(G'|\pi,\beta)$ 参数. 一般来说，通过最大化似然函数可以将模型参数求出，但是由于上式中连乘的每一项都是由累加计算而来，这使得直接最大化似然函数比较复杂. 基于这种考虑，这里我们采用 EM 算法来迭代的最大化似然函数，同时求出各个参数的值.

在模型中，我们用隐变量 $g_{ab}$ 指示边属于哪个查询意图. 令 $\eta_{ab,r}=Pr(g_{ab}=r|e_{ab})$，那么关于 $\eta_{ab,r}$ 就可以写成：

$$\eta_{ab,r}=Pr(g_{ab}=r\mid e_{ab})=\frac{Pr(g_{ab}=r,e_{ab})}{Pr(e_{ab})} = \frac{\pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}}}{\sum_{r=1}^{K}\pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}}} \tag{4}$$

从该式可以看出 $\eta_{ab,r}$ 反映了第 $r$ 个查询意图对产生边 $e_{ab}$ 的贡献程度.

有了以上的关于隐变量 $g_{ab}$ 的后验估计，我们就可以写出如下关于 $G'$ 的对数似然函数：

$$\ln Pr(G|\pi,\beta)=\sum_{e_{ab}\in E} G_{ab}\eta_{ab,r}\log\Big(\sum_{r=1}^{K}\pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}}\Big) \tag{5}$$

我们在上式中加入 $\pi$ 和 $\beta$ 的概率归一性约束，就得到如下目标函数：

$$LL=\sum_{e_{ab}\in E}\sum_{r=1}^{K} G_{ab}\eta_{ab,r}\log\Big(\pi_r \prod_{i:w_i\in q_a} \beta_{r,i}^{Q_{a,i}} \prod_{i:w_i\in q_b} \beta_{r,i}^{Q_{b,i}}\Big)- \alpha\Big(\sum_{r=1}^{K}\pi_r-1\Big)-\sum_{r=1}^{K}\mu_r\Big(\sum_{i=1}^{|T|}\beta_{r,i}-1\Big) \tag{6}$$

其中 $\alpha,\mu$ 是拉格朗日因子. 我们将上式分别对 $\pi$ 和 $\beta$ 求导数，并令各自结果为 0. 化简后可得到如下的公式：

$$\pi_r=\frac{\sum_{e_{ab}\in E} G_{ab}\eta_{ab,r}}{\sum_{r=1}^{K}\sum_{e_{ab}\in E} G_{ab}\eta_{ab,r}} \tag{7}$$

$$\beta_{r,i}=\frac{\sum_{e_{ab}\in E}(Q_{a,i}+Q_{b,i})G_{ab}\eta_{ab,r}\,\mathbb{I}(w_i\in a\vee w_i\in b)}{\sum_{i=1}^{|T|}\Big(\sum_{e_{ab}\in E}(Q_{a,i}+Q_{b,i})G_{ab}\eta_{ab,r}\,\mathbb{I}(w_i\in a\vee w_i\in b)\Big)} \tag{8}$$

上式中，$\mathbb{I}(\cdot)$ 是指示函数，定义如下：

$$\mathbb{I}(x)=\begin{cases}1, & x=\text{true}\\ 0, & x=\text{false}\end{cases} \tag{9}$$

在实际计算中，我们首先随机初始化 $\pi$ 和 $\beta$ 的值，然后按照式(4)、(7)、(8)迭代计算，就可以得到关于模型参数 $\pi$ 和 $\beta$ 的估计值. EM 算法保证了迭代的收敛性，并且每次迭代都会使对数似然函数增加. 因此我们可以通过设定某一阈值作为终止迭代的条件，当对数似然函数的增幅小于这一阈值之后便认为收敛. EM 算法是确定性算法，其迭代终止时得到的结果是由参数初始化值而定的. 因此我们可以进行多轮迭代，将其中拥有最大似然函数值的那轮迭代的参数作为最后结果.

### 3.3　结合查询意图的长尾查询推荐

在文献[18]中作者提出了一种利用词项查询图对长尾查询进行推荐的方法. 该方法首先从查询会

话中构建出词项查询图，然后对于某个长尾查询 $q=\{t_1,t_2,\cdots,t_m\}$，利用如下概率转移矩阵 $\boldsymbol{A}_i$ 计算每个单词 $t_i$ 在词项查询图上通过个性化随机游走而得到的 PageRank 向量 $\boldsymbol{z}_{(i)}$：

$$\boldsymbol{A}_i=(1-\lambda)\boldsymbol{W}+\lambda\boldsymbol{1}\boldsymbol{e}_i^{\mathrm{T}} \tag{10}$$

其中，$\boldsymbol{W}$ 是由 $\boldsymbol{G}'=\{V',E',W'\}$ 的邻接矩阵按照行进权重归一化得到的矩阵，$\lambda\in[0,1]$ 是个性化随机游走中远距传输概率(teleportation probability)．$\boldsymbol{e}_i$ 是一个 $|V'|$ 维的行向量，该向量除了第 $i$ 个元素是 1，其余全是 0．$\boldsymbol{1}$ 是各元素全是 1 的 $|V'|$ 维行向量．

传统的方法将 $q$ 中每个 $t_i$ 计算出来的 $\boldsymbol{z}_{(i)}$ 按照 Hadamard 积相乘：

$$\boldsymbol{z}_q=\boldsymbol{z}_{(1)}\circ\boldsymbol{z}_{(2)}\circ\cdots\circ\boldsymbol{z}_{(m)} \tag{11}$$

其中，$\boldsymbol{z}_q$ 即是 $G'$ 中每个结点关于查询 $q$ 的推荐得分．然后将得分较高的结点所对应的查询取出作为查询 $q$ 的推荐．

从式(11)可以看出，传统方法使用 Hadamard 积时，将每个单词 $t_i$ 等同看待，且不考虑 $q$ 的查询意图和每个单词 $t_i$ 的相互关系，这样就会造成一些噪音单词 $\boldsymbol{z}_{(i)}$ 一定程度上影响了 $q$ 的最后推荐结果．

基于这样的考虑，我们提出利用查询意图来衡量每个单词 $t_i$ 对查询 $q$ 的重要程度，从而达到消除或减缓噪音单词对查询推荐的影响．

在上一节中，我们可以计算出查询意图的分布 $\pi$ 和查询意图关于单词的分布 $\beta$．对于一个新来的查询 $q_c$，利用贝叶斯公式可得到该查询关于查询意图的分布：

$$Pr(r\mid q_c)=\frac{Pr(r)Pr(q_c\mid r)}{\sum_{r=1}^{K}Pr(r)Pr(q_c\mid r)}=\frac{\pi_r\prod_{i:w_i\in q_c}\beta_{r,i}^{Q_{c,i}}}{\sum_{r=1}^{K}\left(\pi_r\prod_{i:w_i\in q_c}\beta_{r,i}^{Q_{c,i}}\right)} \tag{12}$$

因此对于每一种查询意图 $r$，$q_c$ 中的每个单词 $t_i$ 关于该意图的重要性可以用 $Pr(t_i\mid r)=\beta_{r,i}$ 来表示．利用 $q_c$ 中每个 $t_i$ 的 PageRank 值 $\boldsymbol{z}_{(i)}$，我们可以计算在该种意图 $r$ 下 $q_c$ 的推荐得分：

$$\boldsymbol{z}_{q_c}^{r}=\boldsymbol{z}_{(1)}^{\beta_{r,1}}\circ\boldsymbol{z}_{(2)}^{\beta_{r,2}}\circ\cdots\circ\boldsymbol{z}_{(m)}^{\beta_{r,m}} \tag{13}$$

将不同的 $\boldsymbol{z}_{q_c}^{r}$，以 $Pr(r\mid q_c)$ 为权重累加起来，就得到最后我们关于 $q$ 的推荐得分：

$$\boldsymbol{z}_{q_c}=\sum_{r=1}^{K}\left(Pr(r\mid q_c)\cdot\left(\boldsymbol{z}_{(1)}^{\beta_{r,1}}\circ\boldsymbol{z}_{(2)}^{\beta_{r,2}}\circ\cdots\circ\boldsymbol{z}_{(m)}^{\beta_{r,m}}\right)\right) \tag{14}$$

从式(14)中我们可以看到，不同单词的 $\boldsymbol{z}_{(i)}$ 会由于当前查询 $q_c$ 属于不同的查询意图 $r$ 而表现出不同的重要性．那些关于意图 $r$ 比较重要的单词 $t_i$，由于其 $\beta_{r,i}$ 比较大，故 $\boldsymbol{z}_{(i)}^{\beta_{r,i}}$ 中各个结点得分的差异也比较大，因而能够对 $\boldsymbol{z}_{q_c}^{r}$ 产生相对重要影响．同时由于我们用 $Pr(r\mid q_c)$ 来权衡不同意图的重要程度，故对于 $q_c$ 蕴含的那些不明显的意图(即 $Pr(r\mid q_c)$ 比较小)，其对 $\boldsymbol{z}_{q_c}$ 的最后得分影响也较小．

## 4 实 验

### 4.1 数据集

我们实验数据来源于某商业搜索引擎 3 个月的搜索日志．我们从中剔除了非英语的查询，并将所有的查询中的单词转化为对应的小写单词．我们按照 30 min 的时间间隔将查询切分成查询回话，构建查询流图．为了减小噪音所带来的影响，我们去除了所有发生频率小于 3 次的边，并使用 nltk 工具包中的 lancaster 算法对所有单词提取词干．我们选取了其中最大的一个连通子图来构造词项查询图，该子图包含 16980 个不同的查询、51214 条边以及 8713 个不同的单词．实验中我们经验选取意图的个数为 60，80，100，120，并均取得相似的实验效果，这里我们仅报告话题个数 $K=100$ 的结果．

### 4.2 评价查询意图

在表 1 中，我们展示了 3 个随机选取的查询意图，并将每个词按照 $Pr(w_i\mid r)$ 的大小进行排序，这里我们仅展示了每种查询意图中排序最高的 10 个单词．为了方便辨认，我们将词干还原成对应的单词．我们可以看出这些词清晰地反映了各个查询意图，其中第一列是关于移动通信的，第二列是关于汽车的，第三列是关于航空旅行的．

**表 1 在 3 个查询意图中的高频词**

| 移动通信 | 汽车 | 航空旅行 |
|---|---|---|
| verizon | toyota | airline |
| cingular | honda | southwest |
| wireless | ford | delta |
| sprint | nissan | ticket |
| mobile | dodge | expedia |
| nextel | bmw | cheap |
| tmobile | saturn | continental |
| phone | chevrolet | travelocity |
| verizonwireless | lexus | orbitz |
| cell | chevy | air |

### 4.3 评价长尾查询的推荐

这一节我们将评价对长尾查询的查询推荐质量．我们将文献[18]中提到的方法作为基准方法．基准方法和我们的方法均使用相同的 $\lambda=0.7$ 来计算长尾查询 $q=\{t_1,\cdots,t_m\}$ 中每词 $t_i$ 对应的 PageRank 向量 $\boldsymbol{z}_{(i)}$．

表 2 中列出了两个随机选出的长尾查询："yahoo mail inbox"，"zip code for Washington dc" 为例子，来比较我们的方法和基准方法产生的推荐有效性．从这两个例子中我们可以看出，我们方法产

生的推荐查询同原查询更相关. 而基准方法给出的查询由于等同看待查询中的每个单词，因而推荐出的查询是那些与每个单词都尽量相关的查询. 从而导致了在整个词项查询图中比较流行的查询(如mapquest，bank of america)被推荐出来. 我们的方法通过查询意图来衡量单词关于查询的重要性，因而对于那些比较重要的单词 $w$(如“zip code for Washington dc”中的“zip”和“code”)赋予更多的权重($Pr(w|r)=\beta_{r,w}$较大)，使得计算出来这些节点的PageRank值在该权重的作用下具有较强的区分能力，对最后的排序结果的影响也较大，从而推荐效果也就更好. 如图4的例子所示，我们有一PageRank向量 $z_{(i)}=[0.6 \quad 0.3 \quad 0.1]^{\mathrm{T}}$，相对于不同的单词权重 $\beta_{r,w}=0.1$，$\beta_{r,w'}=0.9$ 得到的两个排序值对每个节点有明显不同的区分性，那些 $Pr(w|r)=\beta_{r,w}$ 较大的 $z_{(m)}^{\beta_{r,w}}$ 区分力较大.

**表 2　两个长尾查询推荐的对比**

| “yahoo mail inbox” | | “zip code for Washington dc” | |
|---|---|---|---|
| 基准方法 | 我们的方法 | 基准方法 | 我们的方法 |
| yahoo mail | yahoo mail | mapquest | zip code |
| hotmail | mail yahoo com | map quest | area code |
| www hotmail com | www mail yahoo com | yahoo mail | zip code finder |
| cingular | yahoo mail com | bank of america | zip code |
| mapquest | yahoo e mail | american idol | zip code map |
| mail | www yahoo mail com | zip code | detroit michigan zip code |
| e bay | yahoo web mail | white page | zip code directory |
| yahoo personal | sbc yahoo mail | yellow page | area code lookup |
| bank of america | yahoo maili | realtor com | zip code lookup |
| gmail | keen | home depot | area code finder |

$$z_{rw'}^{0.1}=\begin{bmatrix}0.6\\0.3\\0.1\end{bmatrix}^{0.1}=\begin{bmatrix}0.95\\0.89\\0.79\end{bmatrix}\qquad z_{rw'}^{0.9}=\begin{bmatrix}0.6\\0.3\\0.1\end{bmatrix}^{0.9}=\begin{bmatrix}0.63\\0.34\\0.13\end{bmatrix}$$

图 4　查询意图对排序值的影响

另外，我们请了3位评价者来人工地标记推荐查询的相关性. 这里我们采用3级标注法，其中0分表示根本不相关，0.5分表示部分相关，1分表示完全相关. 我们从已有的查询日志中随机地产生100个在词项查询图中没有出现过的查询，同时要求组成这些查询的单词包含在词项查询图内. 我们按照式(15)计算关于查询 $q$ 的相关性得分($RS$). 其中 $q$ 表示当前查询，$k$ 表示关于 $q$ 的推荐查询个数，$L=\{l_1,l_2,\cdots,l_k\}$ 表示各个推荐所得分数.

$$RS_q=\frac{\sum_{i=1}^{k}l_i}{k} \tag{15}$$

其中 $q$ 表示当前查询，$k$ 表示关于 $q$ 的推荐查询个数，$L=\{l_1,l_2,\cdots,l_k\}$ 表示各个推荐所得分数.

我们按照推荐分数进行排序，并选取前 $k=1$，3，5，7，10 个查询让用户标注同源查询的相关性. 图5展示了我们的方法和基准方法随机选取100个查询的平均相关性得分的对比. 可以看出随着推荐查询个数的增多，越来愈多的不相关的查询被推荐出来，所以基准方法和我们的方法的相关性得分都在变差. 但显而易见，由于我们的方法考虑了查询意图的影响，推荐查询的相关性明显好于基准方法的推荐查询.

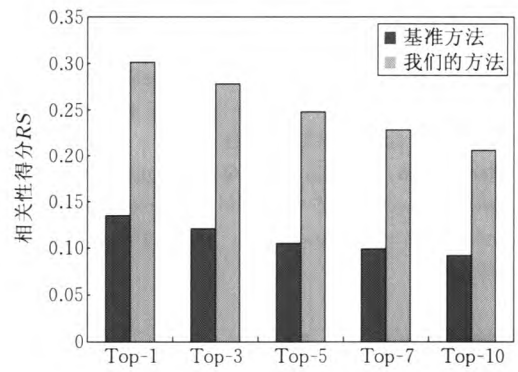


图 5　相关性得分对比

## 5　总　结

本文提出了一种新的基于查询意图的长尾查询推荐方法. 不同于以往关于长尾查询的推荐方法，本文提出了一种概率混合模型来挖掘词项查询图中的查询意图，在该模型中我们将查询意图定义为单词的分布，因而可以从单词的角度去预测长尾查询的查询意图. 此外我们还提出了一种根据查询意图来集成单词的个性化随机游走的方法，该方法通过衡量单词在查询中的重要程度，对长尾查询进行推荐. 实验验证了我们推荐方法的有效性. 在以后的工作中，我们将会考虑 url 等信息，进一步提高推荐效果.

## 参　考　文　献

[1]　Goel S, Broder A, Gabrilovich E, Pang B.　Anatomy of the

long tail: Ordinary people with extraordinary tastes//Proceedings of the ACM international conference on Web search and Data Mining. New York, USA, 2010: 201-210

[2] Huh S, Fienberg S E. Discriminative topic modeling based on manifold learning//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2010: 653-662

[3] Kawamae N, Takahashi K. Information retrieval based on collaborative fltering with latent interest semantic map// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York, USA, 2005: 618-623

[4] Mei Q, Cai D, Zhang D, Zhai C. Topic modeling with network regularization//Proceedings of the International Conference on World Wide Web. New York, USA, 2008: 101-110

[5] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 2001, 42(1-2): 177-196

[6] Blei D M, Ng A Y, Jordan M I, Lafferty J. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993-1022

[7] Newman M E J, Leicht E A. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences, 2007, 104(23): 9564-9569

[8] Ramasco J J, Mungan M. Inversion method for content-based networks. Physical Review E, 2008, 77(3): 036122

[9] Ren W, Yan G, Liao X. A simple probabilistic algorithm for detecting community structure in social networks. Physical Review E, 2009, 79(3): 036111

[10] Mei Q, Zhou D, Church K. Query suggestion using hitting time//Proceedings of the ACM Conference on Information and Knowledge Management. New York, USA, 2008: 469-478

[11] Zhu X, Guo J, Cheng X, Du P, Shen H W. A unified framework for recommending diverse and relevant queries//Proceedings of the International Conference on World Wide Web. New York, USA, 2011: 37-46

[12] Guo J, Cheng X, Xu G, Shen H. A structured approach to query recommendation with social annotation data//Proceedings of the ACM International Conference on Information and Knowledge Management. New York, USA, 2010: 619-628

[13] Boldi P, Bonchi F, Castillo C, Donato D, Gionis A, Vigna S. The query-flow graph: Model and applications//Proceedings of the ACM Conference on Information and Knowledge Management. New York, USA, 2008: 609-618

[14] Bordino I, Castillo C, Donato D, Gionis A. Query similarity by projecting the query-flow graph//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 2010: 515-522

[15] Song Y, He L W. Optimal rare query suggestion with implicit user feedback//Proceedings of the International Conference on World Wide Web. New York, USA, 2010: 901-910

[16] Pandey S, Punera K, Fontoura M, Josifovski V. Estimating advertisability of tail queries for sponsored search//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 2010: 563-570

[17] Szpektor I, Gionis A, Maarek Y. Improving recommendation for long-tail queries via templates//Proceedings of the International Conference on World Wide Web. New York, USA, 2011: 47-56

[18] Bonchi F, Perego R, Silvestri F, Vahabi H, Venturini R. Recommendations for the long tail by term-query graph// Proceedings of the International Conference Companion on World Wide Web. New York, USA, 2008: 15-16

**BAI Lu**, born in 1983, Ph. D. candidate. His research interests include Web search and text mining.

**GUO Jia-Feng**, born in 1980, Ph. D., associated professor. His research interests include Web search and mining, user data mining, machine learning, and social network.

**CAO Lei**, born in 1977, Ph. D.. His main research interests include entity ranking and data mining.

**CHENG Xue-Qi**, born in 1971, Ph. D., Ph. D. supervisor. His main research interests include network science, Web search and data mining, P2P and distributed system, information security.

**Background**

This paper focuses on the recommendation for long-tail query. Query recommendation has been recognized as an important tool that helps users seek their information needs. Many approaches have been proposed to generate query recommendations by leveraging query logs. Traditional recommendation methods were mainly care about the frequent queries, but the low frequency and sparse correlation hurt the performance of recommendation for the long-tail queries. Recently, some methods are proposed for recommendation of long-tail ones by modeling the words of queries. But without of modeling the query intent explicitly, the recommendations are often contaminated by noise words. In this paper, the authors propose a novel probabilistic mixture model of term-query graph that can clearly identify the query intents. Moreover, an intent-biased rank method is introduced for recommendation that can greatly reduce the influences of noise words. Experiments are conducted based on real world query logs, and both the qualitative and quantitative results demonstrate the effectiveness of our approach.

This work is funded by the National Natural Science Foundation of China under Grant Nos. 60933005, 61173008, 61003166, and 61203298, the National Basic Research Program (973 Program) of China under Grant No. 2012CB316303.