[电子与信息科学/**Electronics and Information Science**]

# 基于位置的个性化关键词查询推荐

梁耀培，吴定明

深圳大学计算机与软件学院，广东深圳 518060

**摘要**：查询推荐是指根据用户的输入提供若干替代的查询，用户使用推荐的查询去检索，得到更多符合其需求的信息．利用基于位置的关键词查询推荐所提供的替代关键词能够检索到在用户查询位置附近的信息．用户提交的关键词常是多义词且含有各自的背景偏好，采用具有个性化的推荐查询则能检索到符合用户偏好的信息．为同时满足空间位置邻近和个性化需求，提出一种基于位置的个性化关键词查询推荐方法，使推荐查询的关键词能够检索到位于用户附近且符合其偏好的信息．用关键词-文档二部图表示不同关键词查询之间的语义相似性，采用动态边权重调整策略，建立与关键词相关的文档和用户当前位置的空间关系，使用分类向量模型表示用户的兴趣爱好，应用带重启的随机漫步模型，得到与用户输入的关键词具有较高相似度的其他关键词．在 AOL 真实数据集上的测试结果表明，该方法为用户推荐的关键词不仅可以满足用户的信息需求，还可以检索到用户位置附近符合其偏好的文档．

**关键词**：人工智能；数据库；数据结构；关键词推荐；个性化；随机漫步；空间坐标数据；二部图

**中图分类号**: TP319　　**文献标记码**：A　　doi:

# Location-aware personalized keyword query recommendation

LIANG Yaopei and WU Dingming

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, P. R. China

**Abstract**：Query recommendation provides several alternative queries based on the input query. Using the recommended queries, users may retrieve more relevant information. Location-aware keyword query recommendation aims for suggesting queries that are able to retrieve relevant information close to the user's location. When the submitted queries are ambiguous and have various background preferrences, personalized query recommendation can recommend queries that retrieve relevant information matching users' preferences. This work studies a new method of query recommendation, i.e., the location-aware personalized keyword query recommendation. The queries suggested by this approach are able to retrieve nearby relevant information that matches users' preferences. The proposed method establishes the semantic relationships among keyword queries via a keyword-document bipartite graph. The weights of the edges in the keyword-document bipartite graph are dynamically adjusted to represent the spatial proximity of the documents. Users' preferences are modeled by category-based vectors. The random walk with restart model is used to compute recommended queries. This work develops an efficient algorithm and data structures for the computation of the recommendations. Experiments on a

**Received:** 2018-08-28;　　**Accepted:** 2018-11-26

**Foundation**：National Natural Foundation of China (61502310)

**Corresponding author:** Associate professor WU Dingming. E-mail: dingming.wu@szu.edu.cn

**Citation:** LIANG Yaopei, WU Dingming. Location-aware personalized keyword query recommendation [J]. Journal of Shenzhen University Science and Engineering，2019，36(4)：**-**. (in Chinese)

real data set demonstrate the effectiveness of the proposal.

**Keywords**：artificial intelligence; database; data structure; keyword recommendation; personalization; random walk; spatial data; bi-partite graph

在信息检索系统中，关键词查询推荐技术利用根据用户提交查询的关键词，向其推荐若干相关关键词，用户再用这些推荐的关键词去检索，找到所需信息．现有的关键词查询推荐方法大多是基于搜索日志中的点击信息[1-3]和查询会话记录[4]，主要分为 3 类：基于随机漫步（random walk with restart, RWR）的方法[5]、排列学习方法[6]和基于聚类的方法[2]．近年来，也有学者使用神经网络来进行基于位置的查询推荐[7]．在基于位置的信息检索中，用户不仅希望找到与所查询关键词相关的信息，还希望返回的信息在用户位置附近．这个需求来自于空间关键词搜索的增多[8]．个性化的推荐模型在计算的过程中会考虑用户的偏好，因此不同偏好的用户即使提交相同的关键词，也会得到不同的推荐关键词查询．

推荐算法考虑位置和用户偏好，需要解决两个问题：一是在如何在有效地衡量关键词查询的语义相似度的同时，兼顾用户位置与文档之间的空间距离的邻近性．K-D 图是经典的不考虑位置的关键词推荐方法[3-5, 8-10]．本研究构造的关键词-文档（keyword-document）二部图（简称 K-D 图），可将关键词查询和与其相关的文档连接起来，通过调整 K-D 图中边上的权重，使之不仅反映关键词查询之间的语义相关性，还能反映文档位置与用户位置 $\lambda_q$ 的空间距离．通过在 K-D 图上采用个性化的 PageRank（PPR），即带重启的随机漫步模型[11]，来计算与用户提交的关键词 $k_q$ 最相似的 $m$ 个查询，使得推荐的查询关键词可检索到 $\lambda_q$ 附近的信息．二是如何使推荐的查询符合用户偏好．本研究对用户提交过的关键词进行分类，将用户提交最多的类别作为用户的偏好．在用户提交关键词时，将属于用户偏好的历史查询也加入推荐方法的计算中．

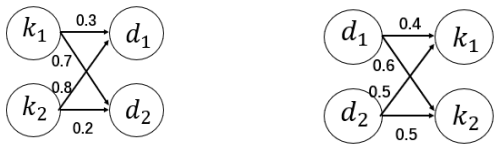
综上所述，本研究的主要工作是：①设计了一个基于位置的个性化关键词查询推荐框架，以获取与用户信息需求相关，符合用户偏好的推荐关键词，这些关键词还能够检索在用户位置附近的文档．②将设计的方法在真实数据上测试，证明基于位置的个性化关键词查询推荐的有用性．

# 1 基于位置的个性化关键词查询推荐

给定一个用户提交的关键词查询 $k_q$（单词或短语），和用户位置 $\lambda_q$，本研究提出的推荐算法需满足 3 个要求：①基于 $k_q$，推荐的关键词能满足用户信息需求，与提交的关键词查询具有语义相似性；②推荐的关键词能够检索与 $\lambda_q$ 在空间上邻近的文档；③推荐的关键词符合用户偏好．

## 1.1 初始 K-D 图

在一个查询日志中，令 $D$ 表示带有位置信息的文档集合，$D$ 中每个文档 $d_i$ 都有一对经纬度坐标 $d_i.\lambda$，$K$ 为查询日志中所有查询关键词 $k_j$ 的集合．首先，构建初始 K-D 图，这是一个有向的加权二部图 $G$=($K$, $D$, $E$)，能够反映关键词查询与文档之间的语义相关性，以及不同查询之间的相似度，能满足①的要求．假设用户提交查询 $k_j$，并点击文档 $d_i$，则 $E$ 包含一条从 $k_j$ 到 $d_i$ 的边 $e$ 和一条从 $d_i$ 到 $k_j$ 的边 $e'$．其中，$e$ 和 $e'$的权重相同，即在日志中提交查询 $k_j$ 后，文档 $d_i$ 被点击的次数[1]．因此，一个查询和一个被点击的文档的联系程度就通过边上权重高低来体现．两个查询的语义相关性则通过它们在 $G$ 上的相似性来体现．查询日志上的任意更新都能被应用到 K-D 图上：一个新的查询或文档，可在图上增加一个新结点；一个新的点击，可更新相应的边的权重．图 2 展示的是一个有 2 个文档 $d_1$、$d_2$ 和 2 个查询 $k_1$、$k_2$ 所对应的二部图．为演示方便，对边上的权重进行了归一化处理（在关键词-文档对中，将点击次数除以最大点击次数）．




(a)关键词到文档的二部图　　（b）文档到关键词的二部图

图 2 K-D 图示例

**Fig. 2 Keyword-document bipartite graph**

### 1.2 对边权重进行基于位置的调整

为让推荐的关键词能够检索到用户位置附近（即文档的位置与用户的位置的欧式距离较小）的文档，本研究基于用户位置和 K-D 图中结点的地理位置关系，调整 K-D 图中边的权重.

用户提交的查询 $q$ 包含 2 个参数：关键词 $k_q$ 和用户查询位置 $\lambda_q$. 对于 $q$，查询结点 $k_i$ 到文档结点 $d_j$ 的边的权重按照式（1）进行调整.

$$\widetilde{w}(e)=\beta w(e)+(1-\beta)\left(1-\mathrm{dist}\left(\lambda_q,\ d_j.\lambda\right)\right) \quad (1)$$

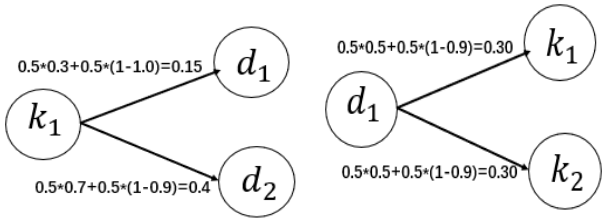
其中，$w(e)$是 K-D 图中原来的边的权重；$\widetilde{w}(e)$是调整后边的权重；dist($\lambda_q$, $d_j.\lambda$)是 $\lambda_q$ 和 $d_j$ 的位置的归一化欧氏距离；参数 $\beta\in[0,\ 1]$，用于平衡原始权重和 $d_j$ 与 $\lambda_q$ 之间的距离权重，$\beta$ 越小，用户的位置对推荐结果的影响越大.

令 $D(k_i)$表示在 K-D 图中与关键词查询 $k_i$ 的连接的文档集合. $D(k_i)$可能包含多个文档. 通过计算 $\lambda_q$ 与 $D(k_i)$的最短距离可调整指向 $k_i$ 的边的权重. 此调整偏向于推荐与 $\lambda_q$ 至少有一个地理位置相近的文档的关键词结点. 从文档结点 $d_j$ 到查询结点 $k_i$ 的边 $e'$ 的权重 $w(e')$的调整公式为

$$\widetilde{w}(e')=\beta w(e')+(1-\beta)\times\left(1-\mathrm{mindist}\left(\lambda_q,D(k_i)\right)\right) \quad (2)$$

其中，mindist($\lambda_q$, $D(k_i)$)是 $\lambda_q$ 到文档集合 $D(k_i)$中所有文档的最短欧氏距离.

假设文档 $d_1$ 和 $d_2$ 距离用户位置$\lambda_q$的欧式距离分别是 1.0 和 0.9 ，其中，$\beta$=0.5. 图 3（b）是从关键词节点到文档节点的边权重调整. 从 $k_1$ 到 $d_1$，dist($\lambda_q$, $d_j.\lambda$)=1.0. 从 $k_1$ 到 $d_2$，dist($\lambda_q$, $d_j.\lambda$)=0.9. 图 3（c）是从文档节点到关键词节点的边权重调整，$k_1$ 与{$d_1$, $d_2$}相连，mindist($\lambda_q$, $D(k_1)$)=0.9.




(b)关键词到文档　　(c)文档到关键词

图 3 基于位置的二部图权重调整

**Fig. 3 Weight adjustment based on bipartite graph**

### 1. 3 基于位置的个性化的关键词查询推荐

#### 1.3.1 在关键词查询推荐中考虑位置

使用 $G_q$ 来表示关键词-文档二部图 $G$ 的边的权重进行调整后的图. 为找到推荐的关键词查询集合，基于随机漫步算法[4, 6]，计算与 $k_q$ 相关的查询的分数. 图 $G_q$ 中结点 $v$ 的 RWR 分数表示一个随机漫步者从 $k_q$ 到达 $v$ 的可能性. 在漫步过程中的每一步，漫步者要么以 1−$\alpha$ 的概率移动到相邻的点，要么以 $\alpha$ 的概率跳转到 $k_q$. 在 $G_q$ 中，排除 $k_q$ 得分最高的前 $m$ 个查询结点，就是算法得到的推荐查询.

设 $\boldsymbol{\Psi}$ 为记录了 $G_q$ 中 $K$ 的所有查询分数的列向量，其计算式为

$$\boldsymbol{\Psi}=(1-\alpha)\boldsymbol{M}_{DK}^{\mathrm{T}}\boldsymbol{M}_{KD}^{\mathrm{T}}\boldsymbol{\Psi}+\alpha\boldsymbol{\Psi}_q \quad (3)$$

其中，$\boldsymbol{M}_{DK}$ 是一个文档量×查询量的矩阵；$\boldsymbol{M}_{KD}$ 是一个查询量×文档量的矩阵，两者存储 $G_q$ 中边的权重.

#### 1.3.2 在关键词查询推荐中考虑个性化

令 $\boldsymbol{\Psi}_q$ 为原始的分数向量. 为在推荐中考虑用户偏好，将属于用户偏好的 $m$ 个历史查询记录，在矩阵中赋值为（1−$\gamma$）/$m$，并设 $k_q$=$\gamma$.

如图 4，$q_1$ 到 $q_5$ 是用户的查询记录，分属类别 $C_1$ 和 $C_2$. 用户的偏好是类别 $C_2$. $q_6$ 是用户新提交的查询. 在本算法中，$q_6$ 的初始值是 $\gamma$，$q_3$ 到 $q_5$ 属于用户偏好类别，初始值是（1−$\gamma$）/3. 其中，$q_1$ 到 $q_5$ 是关键词-文档二部图的关键词节点，分别对应 $k_1$ 到 $k_5$.

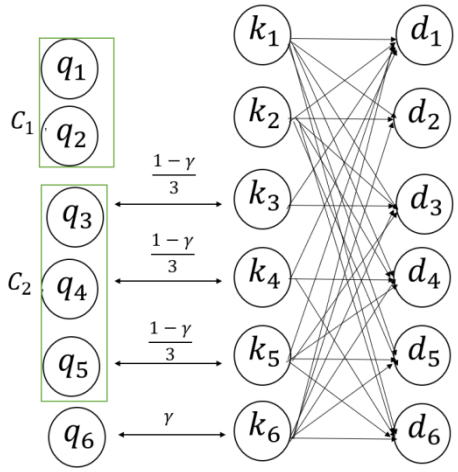



**图 4 对历史查询和现有查询的权重分配**

**Fig. 4 Weight allocation to history queries and current query**

## 2 基于位置的个性化关键词查询推荐算法

本研究通过扩展 Bookmark-Coloring 算法[12]（BCA），计算基于 RWR 算法的 $m$ 个推荐查询．本算法的 K-D 图 $G_q$ 有 2 种类型的结点：查询结点和文档结点．与 BCA 不同的是，本算法只对查询结点评分，查询结点保留 $\alpha$ 部分的活跃墨水，然后向与其相连的结点散播 $1-\alpha$ 部分的活跃墨水；而文档结点将它所有的活跃墨水散播到与其相连的结点．算法的伪代码见附加材料图 1（请扫描论文末页右下角二维码查看图）．优先队列 $Q$ 以活跃墨水量的降序顺序来处理结点，最初包含 $m$+1 个结点．其中，$m$ 个属于用户偏好的历史查询的关键词结点，墨水量为（$1-\gamma$）/$m$，1 个为用户提交的关键词结点，墨水量为 $\gamma$．优先队列 $C$ 初始为空，按保留墨水量的降序来存储推荐查询的候选者．出现以下任一情况，算法结束：①队列 $C$ 中第 $m$ 个结点的活跃墨水量，比第 $m$+1 个结点的活跃墨水量加上剩余的活跃墨水量多；②每个结点的活跃墨水量低于 $\varepsilon$（$\varepsilon$=1×10$^{-3}$）．所有结点的活跃墨水的总和被设成 1（算法伪代码第 3 行）．终止条件①和②分别在第 8 行和第 4 行检验．对查询结点的处理是：保留 $\alpha$ 部分的活跃墨水，然后向与它相连的文档结点的散发（$1-\alpha$）部分的活跃墨水．活跃墨水的总量作对应的修改（第 14 行）．若查询结点已保留墨水，则进入优先队列 $C$．文档结点的处理是：根据调整后边的权重，向与它相连的关键词结点散播其全部活跃墨水．算法返回 $C$ 中除 $k_q$ 外前 $m$ 个关键词查询．

## 3 基于真实数据的测试结果

### 3.1 运行前的数据处理

本研究使用 AOL 数据集（https://jeffhuang.com/search_query_logs.html）对所提出的方法进行检验．AOL 数据集是一个基于 AOL 搜索引擎的真实查询日志，每条记录包含用户 ID、用户提交的查询和点击的 URL．为使日志数据适于使用，需进行以下处理：①对日志的 URL 定位，通过 Freegeoip 项目（https://freegeoip.io/）获得 URL 位置，该项目可确保获取位置的可靠性；②确定用户偏好．本研究使用基于开放目录项目（open directory project）的分类器来对用户的查询记录进行分类，将查询次数最多的类别作为用户的偏好．当一条记录中包含用户 $u_i$、查询 $k_q$ 和文档 $d_j$ 时，在 $u_i$ 和 $k_q$ 之间添加边，边的权重是 $u_i$ 对 $k_q$ 的点击次数除以用户的总查询次数．在 $k_q$ 和 $d_j$ 添加边，其权重是包含 $k_q$ 和 $d_j$ 的记录条数除以 $k_q$ 的记录条数．最终，构建的 K-D 图有 187 050 个查询，79 376 个文档，126 889 个用户．

### 3.2 算法衡量指标

本研究提出的推荐算法除了需满足第 1 章的 3 个要求外，还需有较好的时间性能．为此，采用以下 4 个标准来衡量算法性能．

衡量标准 1：使用推荐的关键词进行检索，检索返回属于用户偏好且距离用户位置 0.1 欧氏距离范围内的文档数量．为了能够使用关键词进行检索，使用 Lucene 工具包对所有文档进行索引，并将推荐返回的关键词作为查询关键词，确定检索返回的符合衡量标准 1 的文档数量．

衡量标准 2：衡量原始查询和推荐查询检索的属于用户偏好并且在离用户位置 0.1 范围内的文档相似性．令$\{d^o\}$表示原始查询检索返回的在查询位置 0.1 范围内且属于用户偏好的前 10 个文档，令$\{d^s\}$表示推荐的关键词检索返回的在查询位置 0.1 范围内且属于用户偏好的前 10 个文档．计算$\{d^o\}$和$\{d^s\}$的返回列表在同一位置的文档之间的余弦相似度，通过 nDCG(normalized discounted cumulative gain)方法[12]来聚合这些余弦的相似度，计算式为

$$S_{\cos}=\frac{\cos(d_1^o+d_1^s)+\sum_{i=2}^{10}\cos(d_i^o,d_i^s)/\lg\ i}{1+\sum_{i=2}^{10}(1/\lg\ i)}$$

衡量标准 3：衡量推荐查询的类别与用户偏好的相关性．使用分类器将关键词查询分类，计算关键词的共同前缀，根据共同前缀确定关键词与用户偏好的相关性，计算式为

$$\mathrm{sim}(C_1,C_2)=\frac{|P(C_1,C_2)|}{\max(|C_1|,|C_2|)}$$

本研究使用分类器对两个推荐算法返回的前 5 个关键词进行分类，计算这些关键词类别与用户偏好的相似度的平均值．

衡量标准 4：记录算法的运行时间作为衡量标准之一．

### 3.3 测试结果及分析

#### 3.3.1 算法运行参数

$\alpha$ 为 RWR 中随机漫步重新开始的概率，依次设

置为 0.20、0.35、0.50、0.65 和 0.80，默认值为 0.50；$\beta$ 为边权重调整参数，即散播墨水中用户位置对推荐结果的影响，依次设置为 0、0.25、0.50、0.75 和 1.00，默认值为 0.50；$\gamma$ 为用户偏好对推荐结果的影响权重，依次设置为 0.10、0.30、0.50、0.70 和 0.90，默认值为 0.50.

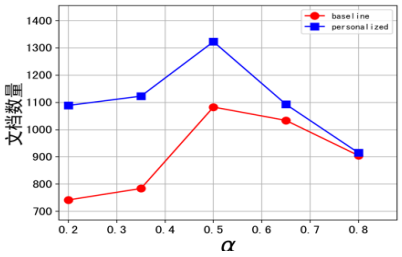
3.3.2 结果分析

图 5 展示了不同参数设置对算法运行结果的影响. 其中，baseline 表示只考虑用户位置，不考虑用户偏好的推荐算法；personalized 是本算法，即考虑用户偏好和位置的推荐算法.

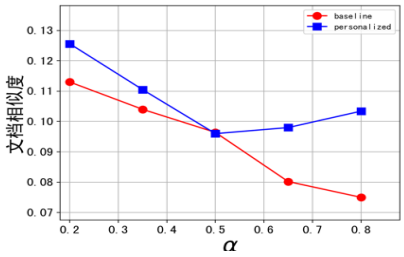
1）图 5（a）、（b）和（c）显示了 $\alpha$ 对检索结果的影响. 参数 $\alpha$ 与关键词节点的保留墨水的部分相关. 对于不同的 $\alpha$，在衡量标准 1、2 和 3 中，personalized 算法的表现都优于 baseline 方法. 在时间性能方面，personalized 算法运行时间较 baseline 方法长（图 5（j）），这是因为 personalized 算法的初始队列中包含更多的查询关键词. 算法的运行时间随着 $\alpha$ 的增大而缩短. 这是因为 $\alpha$ 越大，越多墨水留在关键词节点中，活跃墨水的总量下降的也越快，使得终止条件更早被满足. 当 $\alpha$=1.00 时所有墨水被保留在关键词节点中，算法结束，返回结果为空. 当 $\alpha$=0 时，每一步都无墨水被保留，因此墨水总是被重新分配，直到随机漫步过程达到稳定状态. 在这种情况下，节点的最终得分依赖于图的结构，而非开始节点（关键词节点），即无论输入什么关键词，最终结果都相同. 因此，$\alpha$=0 或 1.00 不能给出有效的结果.

2）图 5（d）、（e）和（f）显示了参数 $\beta$ 对检索结果的影响. $\beta$ 是对用户提交的查询关键词具有语义相似性的关键词分配的权重，1−$\beta$ 表示与查询位置相近的关键词分配的权重. 当 $\beta$ 接近 0 或 1.00 时，只有与查询位置相近的或与用户提交的关键词语义相似度较高的关键词才会参与计算. 但是，当 $\beta\in$(0, 1.00)时，与查询位置稍远但与查询具有高语义相似度的文档，或与查询位置相近但语义相似度低的文档都被考虑到了. 对于不同的 $\beta$，在衡量标准 1 中，personalized 算法的表现优于 baseline 算法；在衡量标准 2 中，$\beta$=1.00 时 baseline 算法表现好于 personalized 算法，原因是此时两个算法都只考虑与用户提交查询相似度高的文档，未考虑位置信息；在衡量标准 3 中，$\beta$=1.00 时，两个算法表现相近，原因同衡量标准 2. 在时间性能方面，personalized 算法运行时间高于 baseline 算法（图 5（k）），原因同（1），且在 $\beta$=0.75 时，算法运行时间最长.

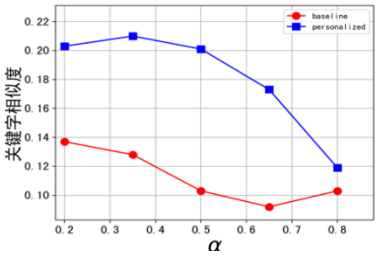
3）图 5（g）、（h）和（i）显示了 $\gamma$ 对检索结果的影响. 参数 $\gamma$ 为对用户提交的关键词查询分配的墨水量，1−$\gamma$ 为对属于用户偏好的历史查询分配的墨水量. 因为 baseline 算法不需要参数 $\gamma$，所以在 3 个图中 baseline 算法的表现没有变化. 对于衡量标准 1 和 2，personalized 算法的表现优于 baseline 算法. 对于衡量标准 3，personalized 算法只有在 $\gamma$=0.9 时表现优于 baseline 算法. 在时间性能方面，personalized 算法的运行时间较 baseline 算法长，原因与 1）相同，且在 $\gamma$=0.5 时，算法运行时间最长.



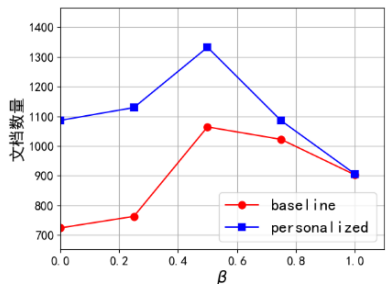
（a）$\alpha$ 对检索文档数量的影响

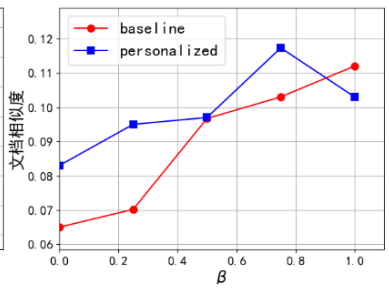


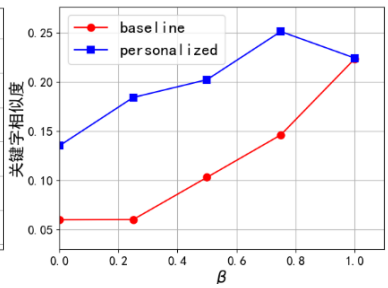
（b）$\alpha$ 对检索文档相似度的影响

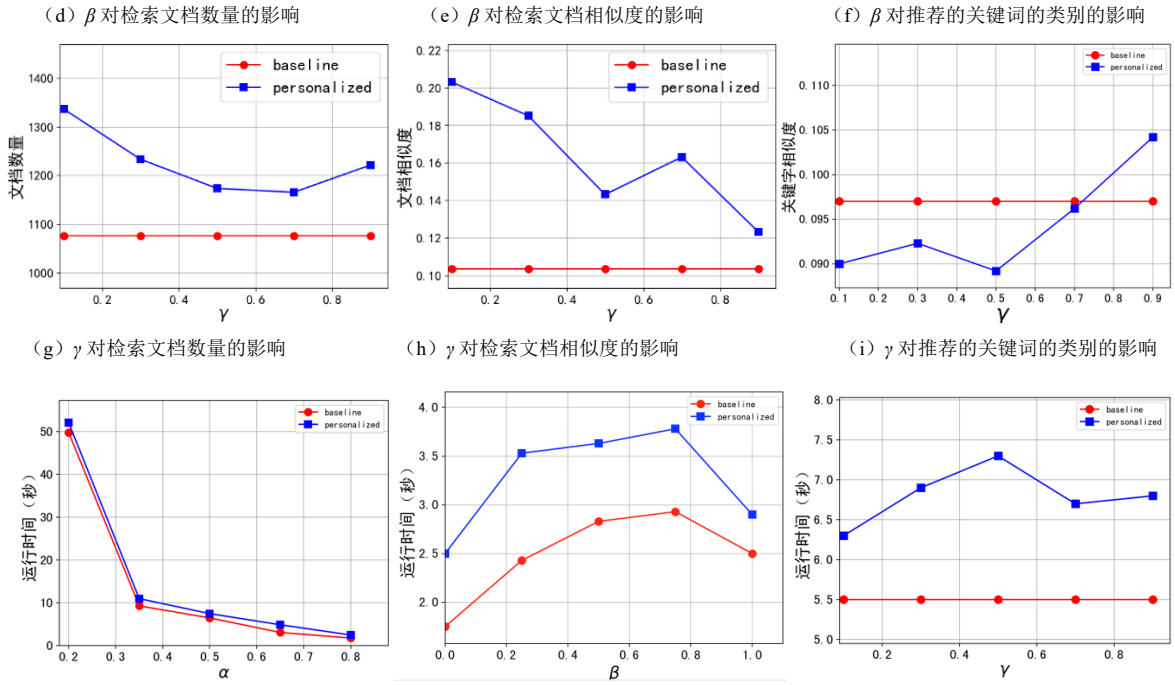


（c）$\alpha$ 对推荐的关键词的类别的影响






（d）$\beta$ 对检索文档数量的影响　（e）$\beta$ 对检索文档相似度的影响　（f）$\beta$ 对推荐的关键词的类别的影响

（g）$\gamma$ 对检索文档数量的影响　（h）$\gamma$ 对检索文档相似度的影响　（i）$\gamma$ 对推荐的关键词的类别的影响

（j）$\alpha$ 对运行时间的影响　（k）$\beta$ 对运行时间的影响　（l）$\gamma$ 对运行时间的影响

**图 5　不同参数设置对算法运行结果**

**Fig. 5 Algorithm running result base on various parameters setting**

## 结　语

为推荐考虑了位置信息和用户偏好信息的关键词，本研究在文档-关键词二部图上，使用带重启的随机漫步算法，计算关键词查询之间的语义相关性，以及档位置与用户位置的空间距离．在 AOL 数据集上对该算法进行测试，得到较好的结果．但是，用户的偏好会随时间发生变化，过去的偏好不能代表用户现在的喜好．下一步我们将加入时间因素后，用户偏好变化时的关键词查询推荐．

**基金项目**：国家自然科学基金资助项目(61502310)

**作者简介**：梁耀培(1996)，深圳大学硕士研究生．E-mail:

**引　　文**：梁耀培，吴定明．基于位置的个性化关键字查询推荐[J]．深圳大学学报理工版，2019，36(4)：00-00．

**参考文献/References：**

[1] BAEZA-YATES R, HURTADO C, MENDOZA M. Query recommendation using query logs in search engines [C]// Proceedings of the international Conference on Current Trends in Database Technology. Berlin: Springer-Verlag, 2004: 588-596.

[2] BEEFERMAN D, BERGER A. Agglomerative clustering of a search engine query log [C]// Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2000: 407-416.

[3] CAO Huanhuan, JIANG Daxin, PEI Jian, et al. Context-aware query suggestion by mining click-through and session data [C]// Proceedings of the14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008: 875-883.

[4] MIYANISHI T, SAKAI T. Time-aware structured query suggestion [C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Develop in Information Retrieval. Dublin: [s. n.], 2013: 809-812.

[5] CRASWELL N, SZUMMER M. Random walks on the click graph [C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2007: 239-246.

[6] LI Lin, XU Guandong, YANG Zhenglu, et al. An efficient approach to suggesting topically related web queries using hidden topic model [J]. World Wide Web, 2013, 16: 273-297.

[7] SONG Jun, XIAO Jun, WU Fei, et al. Hierarchical contextual attention recurrent neural network for map query suggestion [J] IEEE Transactions on Knowledge and Data Engineering, 2017, 29(9): 1888-1901.

[8] WU Dingming, CONG Gao, JENSEN C S. A framework for efficient spatial web object retrieval [J] The Internal Journal on Very Large Data BAses, 2012, 21(6): 797-822.

[9] JÄRVELIN K, KEKÄLÄINEN J, Cumulated gain-based evaluation of IR techniques [M] ACM Trans Actions on Information Systems, 2002, 20(4): 422-446.

[10] TONG Hanghang, FALOUTSOS C, PAN Jiayu. Fast random walk with restart and its applications [C]// Proceedings of the 6th International Conference on Data Mining. Washington D C: IEEE Computer Society, 2006: 613-622.

[11] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine [J]. Computer Networks ISDN Systems, 1998, 30(1/2/3/4/5/6/7): 107-117.

[12] OZERTEM U, CHAPELLE O, DONMEZ P, et al. Learning to suggest: a machine learning framework for ranking query suggestions [C]// Proceedings of the 35th International ACM SIGIR Conference on Research Development Information Retrieval. New York, USA: ACM, 2012: 25-34.

[中文责编： 英 子； 英文责编： ]