利用查询重构识别查询意图*

张晓娟　陆　伟
(武汉大学信息资源研究中心　武汉 430072)

【摘要】基于AOL查询日志数据集,在不给定查询意图类目体系情况下,尝试利用查询重构来识别用户查询意图。主要探讨如何识别出能表达原查询用户意图的查询重构以及如何对识别的查询意图进行聚类两个问题。人工评测结果表明,该方法能够取得较好的实验效果。

【关键词】查询意图　查询重构　随机游走　查询意图聚类
【分类号】G353.4

Identifying Query Intent by Exploiting Query Refinement

Zhang Xiaojuan　Lu Wei
(Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China)

【Abstract】Based on the AOL log dataset, this paper tries to exploit query reformation to identify the concrete query intent of users without given query intent category system. This paper mainly discusses how to identify the query reformation which can express the user intent of original query and how to cluster the query intent. The final results evaluated manually show that this experiment achieves a good effect.

【Keywords】Query intent　Query refinement　Random walk　Query intent clustering

1　引　言

因搜索引擎的搜索方式大多基于关键词组合,而用户提交给搜索引擎的有限关键词常常不能完整地表达用户信息需求,自动识别查询中所包含的用户意图对提高搜索引擎质量有重要意义,其识别结果可用于搜索引擎中的网页排序、查询结果聚类和查询结果呈现等方面。鉴于此,学界对查询意图识别进行了广泛探讨,其研究主要分为给定类目体系的查询意图识别与不给定类目体系的查询意图识别两类,前一类研究主要探讨如何将用户查询映射到某一意图类别,而后一类研究主要探讨用户想要查找的具体内容,如给定查询"汽车"。前一类研究的查询意图识别结果可能是"信息类"、"导航类"、"事务类"或其他类别,而后一类研究的查询意图识别结果可能是"汽车修理"、"汽车清洗"或者"汽车广告"等。因后一类研究能识别更具体的用户查询意图,该研究逐渐成为查询意图识别领域探讨的重点,而如何对查询意图识别结果进行聚类是此类研究的难点[1]。

在查询意图识别过程中,因查询数据具有稀疏性,一般无法直接通过原查询而需通过其他途径来探测用户查询意图。已有研究表明,用户所从事的交互行为(如网页点击、用户反馈、重构查询等)能表达其意图,用户交互行为信息是查询意图识别的主要途径,且目前已有研究大多集中在如何通过用户点击和用户反馈行为信息来识别查询意图,但对重构查询信息探讨甚少。Strohmaier等[2]研究发现,用户向检索系统表达信息需求的主要方式是不断重构查询,用户重构的相关查询是用户意图的直接表达,则查询重构是识别查询意图另一重要途径[3]。

收稿日期:2012-12-25
收修改稿日期:2013-01-18
* 本文系国家自然科学基金面上项目"基于语言模型的通用实体检索建模及框架实现研究"(项目编号:71173164)和武汉大学2012年博士生自主科研项目"网络检索用户查询意图分析与建模研究"(项目编号:2012104010201)的研究成果之一。

本文在不给定查询意图类目体系情况下，利用查询重构来识别查询意图，并对查询意图进行聚类。

2　相关研究

查询意图识别研究主要分为给定类目体系下的查询意图识别与不给定类目体系的查询意图识别两类。其中，前一类研究的主体思想是：通过预先给定的分类体系，选取分类特征、采用自动分类方法将用户查询意图映射到某一类别。其中，陆伟等[4]对此类研究做了大量文献综述。该类方法的最大缺陷是无法获得用户想要查找的具体内容，于是，一些学者尝试在不给定类目体系情况下，自动识别原查询的具体查询意图。而当前此类研究大都基于如下假设：用户在头脑中常常通过“动词 + 名词”形式（如“buy a car”，“repair a car”）表达其意图，于是，扩展与名词相关动词成为后一类研究的主要方法。如 Strohmaier 等[5]根据查询中是否包含动词将查询分为显式意图查询和隐式意图查询，并在此基础上，通过对隐式查询扩展相应动词来识别查询包含的用户意图[6]；Duan 等[7]利用 verb – noun 之间的依存关系来识别非导航类查询的具体意图；He 等[8]利用查询返回的文档片段来获取与该查询相关动词来探测用户查询意图。综合已有相关研究，此类方法虽能识别更加具体的用户查询意图，由于动词之间存在近义词与同义词现象，该方法识别出的结果常常存在概念上的交叉重叠现象，如识别结果“repair a car”与“mend a car”都表达同一查询意图。但又因该方法一般借助相关反馈思想来扩展动词，难以对最终的意图表达形式选取特征对其进行聚类。

关于查询意图识别的途径，相关学者主要探讨了用户点击信息和用户反馈信息，其相关研究主要体现在给定类目体系下的查询意图识别中。如 Lee 等[9]统计得到导航类查询的平均点击次数小于 1.5，信息类的则较大；Liu 等[3]根据 Sogou 搜索引擎日志里查询的点击情况提出两个假设：在执行导航类检索时，用户倾向于进行为数不多的点击，这些被点击的结果往往是靠前的检索结果；Ashkan 等[10]发现商业类查询的广告点击率较大，如果查询为商业导航类点击热度更高；Mendoza 等[11]发现用户花费在导航类查询上的时间比信息类少。总之，目前几乎没有将查询重构行为信息应用到查询意图识别的相关研究，而当前查询重构行为信息主要应用在查询推荐中，如 Shi 等[12]提出了一种基于关联规则的模型来挖掘与原查询相关的查询重构，以此生成候选查询；Jones 等[13]利用根据查询重构与原查询共现信息，利用互信息度量查询间相似性，以此生成候选查询。

本文利用查询重构来表达用户查询意图，将查询意图聚类问题转化为查询聚类问题，则查询聚类是探讨的另一重要问题，其中，如何构建查询向量是研究重点，已有相关研究有：Wen 等[14]利用查询点击文档以及查询点击文档所包含的查询词为查询构建向量；Hosseini 等[15]与 Yi 等[16]通过查询以及文档的点击信息来构建二分图，从而对查询进行聚类，在此基础上，Chan 等[17]通过剔除噪声点击数据来对查询进行向量构建；Baeza – Yates 等[18]通过为查询词加权以此来构建查询的向量，其中查询词权值主要利用查询词在日志中出现频次以及该查询词所出现文档被点击的频次来衡量；Huang 等[19]通过扩展查询内容以及选择 URL 对建立查询词与点击 URL 之间的语义关联为查询构建向量。以上方法适用于对较大数量查询选取特征构建向量，而对较少数量查询聚类时存在局限性。

基于 AOL 查询日志，在不给定分类体系情况下，本文尝试一种新的途径，即利用查询重构来识别原查询的具体用户意图，并主要探讨了如何识别出能表达原查询用户意图的查询重构以及如何对识别出的查询意图进行聚类两个问题。其中，在对用户查询意图进行聚类时，为解决较少数量查询选取特征存在数据稀疏性问题，利用随机游走遍历图思想为每个查询构建向量。需要说明的是：因用户重构查询行为信息主要体现在查询日志中同一 Session 的查询中，在不考虑同一 Session 中查询出现的先后顺序情况下，对查询 q 来说，若某查询 q′与其共现于同一 Session 中，则 q′称为查询 q 的查询重构，并假设该查询可描述 q 的潜在查询意图；另因同一查询相对不同用户会有不同意图，本文所探讨的查询意图识别是指用户查询可能存在的潜在用户意图，而非限定在某一特定用户可能的意图。

3　潜在查询意图识别

本文利用查询重构来表达潜在查询意图的前提假设为：同一 Session 中查询能表达相同的用户意图[20]。因所采用 Session 切分方法难免存在一定缺陷，

以及在同一 Session 中可能存在用户意图转移现象，并非原查询的所有查询重构都能表达其潜在用户意图，于是，综合考虑从查询共现互信息与查询表达式相似性两个方面来识别能描述原查询潜在查询意图的查询重构。

3.1 查询共现互信息

考虑到某些查询重构与原查询共现是偶然现象，假设当查询 q 与其查询重构 q′共现次数大于其偶然共现次数，则说明二者之间存在一定相关性，且相关性越大，则 q′越有可能描述查询 q 的潜在查询意图。本文利用两查询之间在查询日志用户 Session 中互信息 $I(q,q')$ 来确定查询之间的相关关系，公式如下：

$$I(q,q') = \sum_{X_q,X_{q'}\in\{0,1\}} P(X_q,X_{q'}) \log \frac{P(X_q,X_{q'})}{P(X_q)\,P(X_{q'})} \times \frac{1}{3} \quad (1)$$

其中，X_q 与 $X_{q'}$ 分别表示某 Session 是否包含查询 q 或 q′的二元值(0：没出现，1：出现)，$P(X_q)$ 表示包含或者不包含词 q 的 Session 数与总 Session 数的比值，如 $P(X_q=1)$ 表示包含查询 q 的 Session 数与总 Session 数的比值，$P(X_q=0)$ 表示不包含查询 q 的 Session 数与总 Session 数的比值。$P(X_{q'})$ 的意义与 $P(X_q)$ 相同；$P(X_q,X_{q'})$ 表示查询 q 与 q′在 Session 中的联合分布概率，如 $P(X_q=1,X_{q'}=1)$ 表示同时包含查询 q 与 q′的 Session 数占整个 Session 数的概率，$P(X_q=0,X_{q'}=1)$ 表示不包含词 q 但包含词 q′的 Session 所占的比例。$I(q,q')$ 值越大，则表明两查询之间的相关性越大。

3.2 查询表达式相似性

基于如下事实：用户一般重构查询行为是在保持原查询某些词不变的情况下添加、删除与替换词来表达其潜在意图[2]，尽可能利用与原查询相似的查询表达式来表达其潜在意图，因此，本文假设若 q′与 q 拥有相似词越多，说明 q′越能表达 q 的潜在意图。考虑到查询中某些词拼写错误以及词的前缀后缀现象会影响到这种相似性的计算，综合利用文献[21]提出的计算字符串之间相似度的 Di-gram Jaccard 距离以及标准化编辑距离来计算两查询之间的相似性，公式如下：

$$u_{content}(q_1,q_2) = e^{\frac{-\min(u_{jac},u_{levenstein})}{2}} \quad (2)$$

其中，$\min(u_{jac},u_{levenstein})$ 表示取 u_{jac} 与 $u_{levenstein}$ 二者之间较小值，$u_{content}$ 值越大，表明两查询之间的表达式相似性越大。

综合以上查询共现互信息与查询表达式相似性两方面因素，利用 $p_{intent}(q|q')$ 衡量查询 q′能表达查询 q 潜在意图的概率，公式如下：

$$p_{intent}(q|q') = \partial \times I(q,q') + (1-\partial) \times u_{content}(q,q') \quad (3)$$

其中，$I(q,q')$ 表示两查询之间共现互信息，$u_{content}(q,q')$ 表示两查询之间内容相似性；∂ 表示权值，本文将其设置为 0.6。

4 查询意图聚类

不给定分类体系的查询意图识别结果虽能理解更加具体的查询意图，但识别结果可能存在概念之间的交叉重叠现象，如“southwest airlines”的潜在查询意图识别结果中，“southwest airlines flights”与“southwest airlines schedules”描述同一查询意图。因此，对识别出的查询意图聚类显得尤为重要。本文将查询意图聚类问题转化为查询聚类问题，而最终只对较少数量查询进行聚类，数据存在稀疏性，已有构建查询向量方法[14-19]难以有效解决此问题。基于此，本文通过查询以及查询所点击文档来构图，利用随机游走模型[10]来遍历图以此为每个查询构建向量，再对查询意图进行聚类。

4.1 转移概率设定

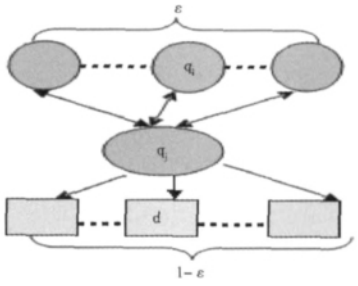

图 1 查询状态与各状态转移情况

构图以及转移概率设定是随机游走模型中的关键部分。本文所构建的二元图 $G=(V,E)$ 中节点 V 由查询词及其所点击文档构成，E 由查询与查询之间共现关系以及查询与文档之间的点击关系构成，其中，共现关系边权值为两查询共现 Session 数与总 Session 数的比值，查询与某文档点击关系边权值为查询点击该文档次数与该查询所点击文档总数的比值。将每个节点当作一个马尔科夫状态，其中，图 1 表示查询状态与各状态之间的转移情况，查询状态 q_j 转移到自身概率为 0，以 ε 的概率转移到文档状态，而以 $1-\varepsilon$ 的概率转

移到其他查询状态，ε 值的设定是为了在转移矩阵中查询节点 q_j 链出的概率之和为 1。从查询状态 q 转移到查询状态 q′的概率计算公式如下：

$$P(q'|q) = (1-\varepsilon) \times \frac{n(q,q')}{\sum n(q,q'')} \quad (4)$$

其中，$n(q,q')$ 表示查询 q 与查询 q′共同出现的频次，$\sum n(q,q'')$ 表示所有与查询 q 共现的查询与 q 共现的频次之和。

在计算查询与文档之间转移概率时将 URL 加以区别对待，若一个 URL 被多个查询点击，则该 URL 具有模糊性，反之，若一个 URL 总是被一个特定查询点击，表明该 URL 具有专指性。设定某查询转移到专指性 URL 的权值大于该查询转移到模糊类 URL 的权值。利用文献 [22] 提出的 iqf 衡量 URL 的权值，其思想与文档的 idf 思想相似，公式如下：

$$iqf(d_j) = \log \frac{|Q|}{n(d_j)} \quad (5)$$

其中，$n(d_j)$ 表示点击文档 d_j 的查询频次之和，$|Q|$ 表示查询日志中总查询数。

$$cfiqf(q_i,d_j) = c_{ij} \times iqf(d_j) \quad (6)$$

其中，cfiqf [22] 表示点击频次与 iqf 的乘积，c_{ij} 表示查询 q_i 点击文档 d_j 的次数。则查询与文档之间的转移概率如下所示：

$$p(d_j|q_i) = \varepsilon \times \frac{cfiqf(q_i,d_j)}{\sum_{j\in D} cfiqf(q_i,d_j)} \quad (7)$$

其中，$\sum_{j\in D} cfiqf(q_i,d_j)$ 表示查询 q_i 所点击文档的 cfiqf 权值之和。

公式(8)表示文档状态是一个吸收状态(即自转移状态)，即转移到自身的概率为 1，而转移到查询状态的概率为 0。根据马尔科夫吸收链的性质：当状态转移到文档状态后，则会一直停留在此状态；从每一个非吸收状态出发，有限次转移后能以正的概率到达某个吸收状态。

$$P(d|d) = 1 \quad (8)$$

4.2　查询向量构建与聚类

基于如下假设：用户所点击文档能描述用户信息需求，则点击同一文档的两查询具有相似查询意图 [19]，根据每个查询所点击文档分布来构建查询向量。为了解决数据稀疏问题，本文为查询构建向量的特点是不只是考虑某查询点击的文档，而是查询到文档的整个路径，即经过随机游走该查询可能到达的文档状态。查询状态与各状态转移情况见图 1。

以图 2 为例，图中节点由查询 q_1、q_2、q_3、q_4、q_5，以及这些查询所点击的文档 d_1、d_2、d_3、d_4、d_5、d_6 构成，其中，图中查询与点击文档的实线边表示点击关系，而虚线边表示经过随机游走后，查询能到达该文档状态。由图 2 可知，经过随机游走后，q_1 与 q_2 最终到达文档状态相似，q_3、q_4、q_5 可到达的文档状态相似。根据每个查询所到达文档状态相似性对其进行聚类，则对图 2 中查询进行聚类最终需达到的效果为：q_1 与 q_2 属于同一类簇，q_3、q_4、q_5 属于同一类簇。本文利用 KL 距离 [22] 来计算两查询向量之间的相似性，并采用 Complete - link [22] 来对查询向量进行聚类，其中，所需聚成的目标类簇数 k 是其探讨重点。

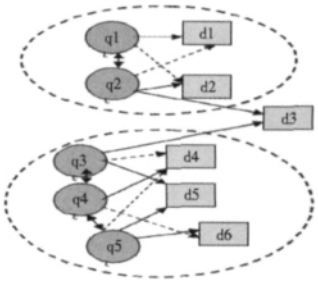

图 2　随机游走后的查询聚类效果示例

5　实验及其结果分析

5.1　数据集

采用 AOL [23] 查询日志作为数据集，其时间跨度为 2006 年 3 月 1 日到 5 月 31 日，其格式如图 3 所示：

217	bestasiancompany.com	2006-03-20 15:15:43	1	http://www.bestasiancompany.com
217	lottery	2006-03-27 14:10:38	1	http://www.calottery.com
217	lottery	2006-03-27 16:34:59	1	http://www.calottery.com
217	ask.com	2006-03-31 14:31:10	1	http://www.ask.com
.....				

图 3　AOL 数据集格式

从左到右分别表示用户 ID、查询表达式、用户点击时间、被点击 URL 在结果列表中排序、点击的 URL 地址。原始数据集中包含许多噪音，本实验首先对其进行清理：如剔除包含色情词查询、只包含单个字符的查询等。笔者通过“15 分钟划分法”[24] 识别 Session 边界。

笔者从查询日志中随机选取 50 个出现频次大于 600 且包含不多于 4 个查询词的非导航类查询作为用于识别用户查询意图的原查询。其中，在进行查询聚

类时，首先获得每个原查询潜在意图识别中排名前 20 的查询重构，再获得每个查询重构在查询日志中所点击的文档，以此来构建每个查询重构的向量。其中，随机游走中参数 ε 决定了查询状态转移到文档状态的概率，本实验中用于聚类的查询较少，实验结果表明，参数 ε 的改变对实验效果影响不大，如当 $\varepsilon = 0.2$ 时，经过 10 步随机游走后大约 90% 达到了吸收状态；当 $\varepsilon = 0.8$ 时，经过 3 步约 90% 达到了吸收状态；同时实验表明，ε 取值在 [0.5，0.7]，随机游走步数 n 设置在 [3，6] 之间，实验效果更佳。于是，将 ε 的值设定为 0.6，n 值设定为 5。

5.2 实验结果评测

笔者选取三名专家对实验结果进行评测，考虑到本实验主要由潜在意图识别和查询意图聚类两步完成，定义以下三个指标来对每个原查询的识别结果与聚类结果进行评测：

(1) 准确度(Precision)：能描述原查询意图的查询个数与该原查询识别结果的总查询数之比；

(2) 内聚合度(Cohesion)：能表达同一查询意图的类簇个数与原查询聚类结果的总类簇之比；

(3) 覆盖度(Coverage)：将识别结果中与类簇中查询相关的所有查询都包含在内的类簇数与该原查询聚类结果的总类簇数之比。

准确度指标用于评价第一步实验结果，则将潜在查询意图识别出的每个查询作为评测对象；而内聚合度、覆盖度指标用于评价查询意图聚类效果，将聚类后的每个类簇作为评测对象。

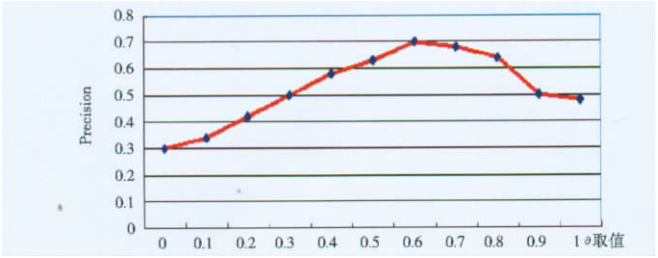

图 4 ∂值对 Precision 值的影响

其中，一些参数值会影响到实验效果。如 Precision 值主要体现公式(3)从查询重构中识别潜在查询意图效果，则∂值对 Precision 有一定影响作用。图 4 表示∂从 0 调节到 1.0 之间，每次调节步长 0.1 而得到不同取值对实验 Precision 的影响，其中，X 轴表示∂的取值，而 Y 轴表示 Precision 值，其计算方法为：首先计算出每位专家的 50 个原查询的 Precision 平均值，再对三位专家的 Precision 平均值再取均值，可以看出，在 [0，0.6] 范围，随着∂值的增加 Precision 值也不断增加，在 [0.6，1.0] 范围，Precision 值随着∂的增加而减少，在本实验中∂的最优值是 0.6。

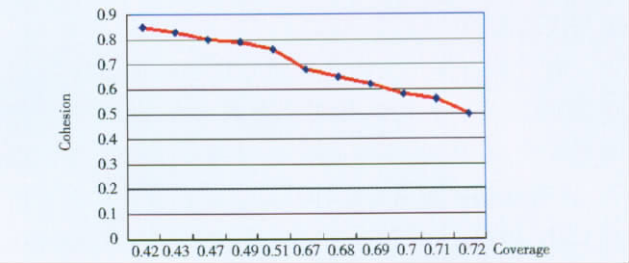

图 5 k 取不同值时其相应的 Coverage 与 Cohesion 值

图 5 表示目标类簇数 k 取不同值时对查询聚类效果的影响。其中，X 轴表示 Coverage 取值，Y 轴表示 Cohesion 取值，二者计算方法与 Precision 值类似。其中，曲线上的 11 个点从左到右分别表示类簇数取值从 15 以差值 1 递减到 5 过程中 Coverage 与 Cohesion 的取值情况。可以看出，随着类簇数不断减少，Cohesion 值不断减少，而 Coverage 值不断增加。经综合对比，当类簇数取值为 10 时，该实验的 Cohesion 值与 Coverage 取得最优解。

当∂取值为 0.6，k 取值为 10 时，三位专家针对以上三个指标对 50 个查询的评测结果，如表 1 所示：

表 1 三位专家的评测结果

指标	A1	A2	A3	Average
Precision	0.73	0.70	0.68	0.70
Cohesion	0.68	0.64	0.72	0.67
Coverage	0.68	0.62	0.71	0.68

可以看出，三位专家的平均准确度、平均聚合度与平均覆盖度分别为 0.70、0.67 与 0.68，说明实验返回的 20 个查询中约 7 个与原查询的意图相关，且聚类结果的 10 个类簇中约 7 个类簇能表达同一查询意图，约 7 个类簇能包含与该类簇查询相关的所有查询，以上数据说明本实验取得了较好的效果。

另外，利用 Cohen kappa(δ)[14] 来获取评测者两两之间同意对方评测结果的概率，公式如下：

$$\delta = \frac{P(O) - P(C)}{1 - P(C)} \quad (9)$$

其中，P(O) 表示两评测者实际同意对方的概率，

而 P(C) 表示期望两评价者同意对方的概率。δ 的取值在 [-1,1] 之间,1 表示二者之间的判断结果完全一致,0 表示二者的判断结果是偶然性造成的,-1 表示二者判断结果完全不一致。

表 2 为三位专家两两之间同意对方的概率,可以看出,三个值都大于 0.6,说明三专家之间同意对方评测结果的概率较高。

表 2　三位专家两两之间的 Cohen kappa(δ) 值

	X-Y	X-Z	Y-Z
Cohen kappa(δ)	0.72	0.65	0.62

当∂值设置为 0.6,类簇数值设置为 10 时,从 50 个原查询中随机选取两查询的最终实验结果示例如表 3,表 4 所示,其中,从表 3 的类簇 C1 与 C2 中可知,用户分别想获得有关 "southwest airline" 的航班与工作信息;另从表 4 的类簇 C1 与 C2 中可以得知,用户想分别获得 "myspace layouts" 相关的颜色背景与设计者信息。

表 3　查询 "southwest airlines" 的最终实验结果

类簇编号	所包含查询	类簇编号	所包含查询
C1	southwest airlines flights southwest airline schedules	C2	southwest airlines employment southwest airlines flight attendant
C3	southwest international airlines	C4	discount airlines tickets southwest airlines specials
C5	southwest airlines club	C6	cheap air flights cheap airline tickets cheap tickets. com southwest airline tickets southwest airlines pricing and restrictions
C7	southwest airlines telephone number southwest airlines customer service	C8	American airlines reservations
C9	southwest airlines fun fares	C10	southwest airlines last minute deals last minute airline tickets last minute flights to atlanta

表 4　查询 "myspace layouts" 的最终实验结果

类簇编号	所包含查询	类簇编号	所包含查询
C1	colorul myspace layouts myspace orange layouts myspace layouts pink	C2	myspace layout editor designer myspace layouts
C3	free myspace layouts	C4	website layouts free web layouts
C5	myspace cool tools	C6	myspace picture upload myspace glitter graphics
C7	myspace layouts of women outdoorsmen myspace layouts girl myspace layouts	C8	icons for myspace myspace font color
C9	myspace background myspace background	C10	myspace layout codes

6　结　语

本文基于 AOL 查询日志,利用查询共现互信息、查询内容相似性来识别潜在查询意图;并通过查询与查询以及查询与文档之间关系来构图,再对该图进行随机游走遍历,以此来为每个查询构建向量,最后对查询意图进行聚类,并取得了较好的实验效果。但该实验仍存在一些不足之处,也是笔者在以后工作中需深入研究的几个方面:

(1) 同一 Session 中,出现在某查询前一时间段或后一时间段的查询对表达该查询用户意图的不同作用;

(2) 后续工作将利用众包思想来对结果进行评测;

(3) 将识别结果应用到搜索引擎优化中,如查询推荐与检索排序中。

参考文献:

[1] Duan R, Wang X, Hu R, et al. Dependency Relation Based Detection of Lexicalized User Goals [C]. In: *Proceedings of the 7th International Conference on Ubiquitous Intelligence and Computing (UIC'10)*. Berlin, Heidelberg: Springer-Verlag, 2010: 167-178.

[2] Strohmaier M, Lux M, Granitzer M. How do Users Express Goals on the Web? - An Exploration of Intentional Structures in Web Search [C]. In: *Proceedings of the 2007 International Conference on Web Information Systems Engineering (WISE'07)*. Berlin, Heidelberg: Springer-Verlag, 2007: 67-78.

[3] Gonzalez-Caro C, Calderon-Benavides L, Baeza-Yates R. Web Queries: The Tip of the Iceberg of the User's Intent [C]. In: *Proceedings of the 2011 International Conference on Web Search and Web Data Mining*. 2011.

[4] 陆伟,周红霞,张晓娟. 查询意图研究综述 [J]. *中国图书馆学报*,2013,39(1):100-111. (Lu Wei, Zhou Hongxia, Zhang Xiaojuan. Review of Research on Query Intent [J]. *Journal of Library Science in China*, 2013, 39(1): 100-111.)

[5] Strohmaier M, Prettenhofer P, Lux M. Different Degrees of Explicitness in Intentional Artifacts: Studying User Goals in a Large Search Query Log [C]. In: *Proceedings of International Workshop on Commonsense Knowledge and Goal Oriented Interfaces (CSK-GOI'08)*. 2008.

[6] Strohmaier M, Kröll M, Körner C. Intentional Query Suggestion: Making User Goals More Explicit During Search [C]. In: *Proceedings of the 2009 Workshop on Web Search Click Data (WSCD'*

09). New York, NY, USA: ACM, 2009: 68 – 74.

[7] He K Y, Chang Y S, Lu W H. Improving Identification of Latent User Goals Through Search – Result Snippet Classification [C]. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2007: 683 – 686.

[8] Lee U, Liu Z, Cho J. Automatic Identification of User Goals in Web Search [C]. In: *Proceedings of the* 14*th International Conference on World Wide Web*. New York, NY, USA: ACM, 2005: 391 – 400.

[9] Liu Y Q, Zhang M, Ru L, et al. Automatic Query Type Identification Based on Click Through Information [C]. In: *Proceedings of Asia Information Retrieval Symposium – AIRS*. Berlin, Heidelberg: Springer, 2006: 593 – 600.

[10] Ashkan A, Clarke C L A, Agichtein E, et al. Classifying and Characterizing Query Intent [C]. In: *Proceedings of the* 31*st Annual European Conference on Information Retrieval Research* (*ECIR'09*), Berlin, Heidelberg: Springer – Verlag, 2009: 578 – 586.

[11] Mendoza M, Zamora J. Identifying the Intent of a User Query Using Support Vector Machines [C]. In: *Proceedings of the* 16*th International Symposium on String Processing and Information Retrieval* (*SPIRE'09*). Berlin, Heidelberg: Springer, 2009: 131 – 142.

[12] Shi X, Yang C C. Mining Related Queries from Web Search Engine Query Logs Using an Improved Association Rule Mining Model [J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(12): 1871 – 1883.

[13] Jones R, Rey B, Madani O, et al. Generating Query Substitutions [C]. In: *Proceedings of the* 15*th International Conference on World Wide Web*. New York, NY, USA: ACM, 2006: 387 – 396.

[14] Wen J R, Nie J Y, Zhang H J. Query Clustering Using User Logs [J]. *ACM Transactions on Information Systems*, 2002, 20(1): 59 – 81.

[15] Hosseini M, Abolhassani H. Hierarchical Co – clustering for Web Queries and Selected URLs [C]. In: *Proceedings of the* 8*th International Conference on Web Information Systems Engineering* (*WISE'07*). Berlin, Heidelberg: Springer – Verlag, 2007: 653 – 662.

[16] Yi J, Maghoul F. Query Clustering Using Click – Through Graph [C]. In: *Proceedings of the* 18*th International Conference on World Wide Web*. New York, NY, USA: ACM, 2009: 1055 – 1056.

[17] Chan W S, Leung W T, Lee D L. Clustering Search Engine Query Log Containing Noisy Clickthroughs [C]. In: *Proceedings of the* 2004 *International Symposium on Applications and the Internet*. 2004: 305 – 308

[18] Baeza – Yates R, Hurtado C, Mendoza M. Improving Search Engines by Query Clustering [J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(12): 1793 – 1804.

[19] Huang X, Du Y, Ren Y. Query Clustering Based on User Feed Back [J]. *Journal of Computational Information Systems*, 2011, 7(13): 4871 – 4879.

[20] Jarvelin A, Jarvelin K. S – grams: Defining Generalized N – grams for Information Retrieval [J]. *Information Processing & Management*, 2007, 43(4): 1005 – 1019.

[21] Jones R, Klinkner K L. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs [C]. In: *Proceedings of the* 17*th ACM Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2008: 699 – 708.

[22] Sadikov E, Madhavan J, Wang L, et al. Clustering Query Refinements by User Intent [C]. In: *Proceedings of the* 19*th International Conference on World Wide Web* (*WWW'10*). New York, NY, USA: ACM, 2010: 841 – 850.

[23] AOL [EB/OL]. [2012 – 12 – 14]. http://www.gregsadetsky.com/aol – data/.

[24] He D Q, Goker A. Detecting Session Boundaries from Web User Logs [C]. In: *Proceedings of the* 22*nd Annual Colloquium on Information*. 2000.

[25] Berry K J, Mielke P W. A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters [J]. *Educational and Psychological Measurement*, 1998, 48(4): 921 – 933.

(**作者** E – mail: zhangxiaojuan624@gmail.com)