情 报 学 报 ISSN 1000 - 0135
第31卷第12期1252 - 1258,2012年12月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 31 No. 12, 1252 - 1258 December 2012

doi:10.3772/j.issn.1000-0135.2012.12.003

基于主题与用户偏好分析的查询推荐研究[1)]

陆 伟 张晓娟
(武汉大学信息资源研究中心 信息检索与知识挖掘研究所,武汉 430072)

摘要 查询日志分析作为近年来常用的查询推荐方法,常采用基于词共现的上下文来生成查询推荐。本文利用AOL日志,在词上下文分析基础之上,采用主题分析,再结合用户偏好,进行查询推荐建模,实验结果表明:采用主题分析可以显著提升查询推荐的精确度,进一步考虑用户偏好后,推荐效果又有了进一步的提升。

关键词 查询 查询推荐 查询替换 查询主题 用户偏好

Study on Query Recommendation Based on the Analysis of Topic and User Personalization

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

Abstract As a common method used in query recommendation in recent years, query log analysis often recommends queries by using the contextual information based on co-occurrence of words. On the basis of analyzing the context of words, this paper employs the topic analysis and combines the user personalization analysis to model the query recommendation by using AOL log. The final results show that the adoption of topic analysis improves the accuracy of the query recommendation significantly and the combination of user personalization analysis further improves the accuracy of the results.

Keywords query, query recommendation, query substitution, query topic, user personalization

1 引 言

作为用户查找网络信息的必备工具之一,搜索引擎在一定程度上降低了用户查找信息的难度。但因搜索引擎大多基于关键词组合来搜索信息,从而导致用户提交给搜索引擎的有限关键词常常不能完整地表达其信息需求,研究表明,查询只能表达用户意图的冰山一角[1]。鉴于此,搜索引擎服务提供商如Google等努力尝试采用多种方法去探测用户的查询意图(即查询所包含的用户信息需求、目标、动机),并将生成查询推荐作为其重要的环节。查询推荐向用户推荐若干与用户输入查询相关的查询,能帮助用户生成更加符合其搜索意图的查询关键字,引导用户的搜索行为,优化搜索结果。查询推荐除被应用到搜索引擎的查询重构,也被广泛应用到其他领域如广告计算、商品推荐、拼写检查、问答系统、探索式搜索等。由于有着巨大的应用需求,查询推荐成为近几年来的研究热点。

本文将在分析查询推荐研究现状的基础上,采用主题分析,结合用户偏好,进行查询推荐建模。文章的结构如下,第二节介绍了查询推荐相关研究现

收稿日期:2012年4月26日

作者简介:**陆伟**,男,1974年生,博士生导师,教授,主要研究方向:信息检索与智能挖掘、数字图书馆、知识管理等。E-mail:reedwhu@gmail.com。**张晓娟**,女,1987年生,武汉大学信息管理学院2011级情报学博士研究生,主要研究方向:查询处理。

1) 本文系教育部人文社科基地重大项目"面向细粒度的网络信息检索模型及框架构建研究"(项目编号:10JJD630014)和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)的研究成果之一。

状,第三节论述了本文采用的查询推荐方法和模型,并在第四节进行了实验验证与评价,第五节做了总结分析。

2 相关研究

文献[2]在对大量相关文献综述的基础之上,根据查询推荐所依赖的数据差异将其方法分为基于文档和基于日志方法。前者是较传统的查询推荐方法,其主要思想:通过处理包含查询或查询词的文档来分析查询,并从相关文档中找出与输入查询相关的词或短语,以此生成候选查询,如文献[3]。此类方法主要包括查询扩展、查询构建、伪相关反馈、隐性语义索引等。此外,一些采用词典(如WordNet)[4]、人工编辑语料(如 Wikipedia、Open Directory Project)[5,6]或其他相关资源产生查询相关词的研究也属于该方法之列。基于文档的方法虽能解决查询稀疏等问题,但存在推荐大量无关噪声词以及人工标注需花费大量人力等缺陷,且最大的难题是如何将生成的相关词或者短语合成查询。

因查询日志中记录了用户构造的各种真实查询,通过分析查询日志更容易找出并推荐合适的查询,于是,基于查询日志的推荐方法逐渐成为近年来常用的方法。该方法的主要思想是:通过查询日志分析寻找过去出现过的相似查询,并根据一定算法排序后择优推荐给用户。其常用方法是利用同一个 Session 中临近的或者同时发生的查询来计算查询之间语义相似性,如 Jones 等[7]利用根据 Session 中查询间的共现信息,利用互信息度量查询间相似性,以此生成候选查询;Shi 和 Yang[8,9]提出了一种基于关联规则的模型来挖掘 Session 中的相关查询;李亚楠等[10]从 Session 中挖掘查询之间的间接联系建立查询关系图,并在图结构相似算法 SimRank 的基础上提出了加权 SimRank 用于查询推荐。朱小飞等[11]基于一个大规模商业搜索引擎查询日志,利用查询数据内在的全局流行度来获得查询之间的相关性,并提出了一种基于流行度排序的查询推荐方法。除以上方法外,有研究者尝试利用查询日志中查询共有相同点击 URL 数生成候选查询,如文献[12]~[13]。还有研究者指出相似查询的搜索频率在时间分布上应该是相似的,于是提出了基于查询频率在时间分布情况来获得相似查询,以此生成查询推荐,如文献[14]~[15]。

以上的研究都是基于查询级别,其研究的单元是整个查询。目前,一些学者尝试采用基于词语级别的方法,将词作为推荐单元,如 Jones 等[7]利用查询对独立假设的可能性比识别用户 Session 中的具有关联性的查询短语或者词对;Wang 等[16]通过分析历史查询中词的上下文相似性提出了一种上下文转移模型,即若某词与原查询中查询词具有相似的上下文分布,则该词有可能替换此查询词,从而生成候选查询推荐。此外,如何越过词字面形式上的相似性,基于主题推荐相关查询也引起了人们的关注。如 Wan 等[17]针对视频检索,提出了一种基于文档主题的查询推荐方法,即采用 LDA 算法来获得数据集中文档的潜在语义结构,进而从文档中获得与查询词主题相关的词,生成候选查询。

考虑到上述查询推荐方法只能定位与原查询相关的普遍用户意图,而不能根据个体用户独特的查询目标生成个性化查询推荐,有学者尝试将用户偏好分析纳入到查询推荐中。例如,梅翔等[18]提出了一种基于用户偏好分析的查询优化方法,该方法将用户对网页的偏好转化为对本体知识库中实例的偏好,并分析本体实例之间的语义关联,发现隐含的用户偏好,在综合用户偏好历史的基础上,建立用户当前状态下偏好的数学模型,以预测用户对网页的关注程度。Sugiyama 等[19]通过分析用户浏览的网页文本,将词语作为记录用户兴趣的基本单位。Mirco 等[20]将网页归类到具有特定意义的概念,用概念权重来表示用户偏好。Liu 等[21]提出了个性化搜索中用户查询映射到分类类目的映射方法,此方法通常推荐前 3 个权值最高的类别,并要求在提交查询请求之前用户需要进行类别选择。

综上可以看出,当前利用主题和用户偏好分析进行查询推荐的探讨都基于文档方法,据笔者所知,基于日志法采用主题和用户偏好分析进行查询推荐的相关研究几乎没有。因此,本文将基于 AOL 查询日志,探讨如何从主题级别进行查询推荐,并在此基础上将用户偏好考虑其中。需要说明的是,通常查询推荐研究可以分为查询替换、查询添加和查询删除几种,而本文主要探讨的是查询替换,且是相同词数的查询替换。

3 相关模型

本文基于词级别生成查询推荐,即通过替换查询中的词来生成候选查询,其任务界定为:给定一个原查询 $q = w_1,\cdots,w_{i-1}w_iw_{i+1},\cdots,w_n$,则查询替换是

将其中一个词 w_i 替换为与其语义相关的词 s，从而形成一个新的更能满足用户查询意图的候选查询 $q' = w_1, \cdots, w_{i-1} s w_{i+1}, \cdots, w_n$。本文主要从词的上下文分析及其词的潜在主题分析两方面来探讨查询词之间的替换模型。

3.1 上下文分析

本小节主要参照文献[17]中的思想：即通过分析查询日志中词的上下文分布，构建词的上下文转移模型，以此来确定在查询中词之间的相似度。其主要内容如下。

3.1.1 上下文的定义

给定一查询词 w，与其共现于同一查询中的词称为词 w 的上下文。其中，w 左边第 i 个上下文 L_i 表示位于词 w 左边且距离为 i 的词构成的集合；同理，查询词 w 右边第 i 个上下文 R_i 的定义与 L_i 类似。例如，某一查询为“Gainesville third party cdl testing”，则词“third”和“Gainesville”分别包含于词“party”的左边第一个上下文 L_1 与左边第二个上下文 L_2 中；“cdl”与“testing”分别包含于词“party”右边第一个上下文 R_1 与右边第二个上下文 R_2 中 。

3.1.2 上下文分布

设定 C 表示特定类型的上下文（可能是 L_i 或 R_i），$C(w)$ 表示词 w 特定上下文 C 中的词集合，且 $count_w(c_i)$ 表示词 c_i 出现在词 w 特定上下文词集 $C(w)$ 的频次。则词 w 的上下文分布概率如公式(1)所示：

$$P_C(c_i \mid w) = \frac{count_w(c_i)}{\sum_{c_j \in C(w)} count_w(c_j)} \quad (1)$$

为避免零概率问题，公式(2)用于对其进行平滑。其中，$P(c_i \mid \theta)$ 表示词 c_i 在整个数据集中出现的概率，μ 表示狄利克雷先验参数，将其设定为3000。本文采用 $P_C(. \mid w)$ 和 $\tilde{P}_C(. \mid w)$ 分别表示词 w 的未平滑和平滑过的上下文概率分布。

$$\tilde{P}_C(c_i \mid w) = \frac{count_w(c_i) + uP(c_i \mid \theta)}{\sum_{c_j \in C(w)} count_w(c_j) + u} \quad (2)$$

3.1.3 上下文转移模型

基于如下假设：出现在相似上下文中的词彼此之间是相关的，能替换彼此。例如词“auto”和“car”有许多共有的上下文词如“sales”与“insurance”等，则二者可相互替代。其中，公式(3)利用KL距离来计算词 w 与 s 的上下文差异性。

$$D(P_C(. \mid w) \parallel \tilde{P}_C(. \mid s)) = \sum_{c \in C(w)} P(c \mid w) \log \frac{p(c \mid w)}{\tilde{P}_C(c \mid s))} \quad (3)$$

其中，$D(P_C(. \mid w) \parallel \tilde{P}_C(. \mid s))$ 表示词 w 上下文与词 s 上下文之间的KL距离，KL距离越大，则二者的上下文相似度越小，反之，越大。于是，在公式(3)的基础之上，采用公式(4)计算词在特定上下文中 s 与 w 之间的转移概率。

$$t_C(s \mid w) = \frac{e^{-D(P_C(. \mid w) \parallel \tilde{P}_C(. \mid s))}}{\sum_u e^{-D(P_C(. \mid w) \parallel \tilde{P}_C(. \mid u))}} \quad (4)$$

以上公式可用于各种类型的上下文模型。其中，u 为本文实验选取的高频词；文献[3]采用上下文 L_1 与 R_1 来计算两词之间的转移概率，见公式(5)。其中，$|L_1|$ 与 $|R_1|$ 分别表示出现在词 w 的 L_1 与 R_1 上下文中的词总数。

$$t(s \mid w) = \frac{|L_1(w)| \times t_{L_1}(s \mid w) + |R_1(w)| \times t_{R_1}(s \mid w)}{|L_1(w)| + |R_1(w)|} \quad (5)$$

3.2 潜在主题分析

因上述上下文分析方法是基于词共现方法，只从字面级别探讨词之间的相关性，则可能存在所推荐候选查询偏离原查询主题的缺陷，如某用户键入查询“apple iPhone”想查找与苹果相关电子产品，而根据上下文分析方法，可能推荐偏离原查询主题的候选查询如“apple salad”。针对此类问题，本小节探讨查询替换中词之间的潜在主题相关性。其中，如何构建伪文档是对查询日志进行潜在语义分析的一个主要问题，一种直接方法是将点击的URL地址作为一个单独文档单元。考虑到大多数的URL地址包含少数查询，存在着数据稀疏问题，本文尝试将所有与某用户相关的查询用于构造用户伪文档，并采用PLSA算法[22]对用户伪文档进行潜在语义分析，使用期望最大化（EM）算法进行参数训练，得到表示“用户（文档）-潜在语义-关键词”三者之间关系的概率模型。用Z表示PLSA生成的一系列潜在主题，每个主题 z 是由一些词的多项分布构成的，从而可获得每个词下的主题分布 $p(z \mid w)$ 以及某一用户（文档）的相关主题分布 $p(z \mid u)$ ，本文即在此基础之上探讨了查询词的主题转移模型与用户偏好性

分析模型。

3.2.1 主题转移模型

基于如下事实:用户在构造查询时其查询意图是单一的,即查询所包含的主题具有唯一性,则推荐查询应与原查询保持主题相似性。给定一原始查询 $q = w_1,\cdots,w_{i-1}w_iw_{i+1},\cdots,w_n$,当词 t 替换查询 q 中某一词时,词 t 的潜在主题与 q 的主题越相关,则生成的候选查询与原查询主题越能相似。本文通过计算词 t 与 q 中各个查询词之间主题相似性以此得到该词与 q 之间的主题相似性。公式(6)用于衡量两词的主题相似性。

$$p_{topic}(t \mid w) = \frac{sim_{topic}(t,w)}{\sum_{s \in W} sim_{topic}(s,w)} \tag{6}$$

其中,W 在本文表示所选取的高频词集合,$sim_{topic}(t,w)$ 表示主题之间的相似性,本文采用 KL 距离对其进行测度,具体计算方法参见公式(7)和公式(8)。其中,$p(.\mid w)$ 与 $p(.\mid t)$ 均为通过 PLSA 算法迭代所生成词的潜在主题分布情况。

$$sim_{topic}(t,w) = e^{-D(p(.\mid t)\,\|\,p(.\mid w))} \tag{7}$$

$$D(p(.\mid t)\,\|\,p(.\mid w)) = \sum_{z_i \in Z} p(z_i \mid t)\log\frac{p(z_i \mid t)}{p(z_i \mid w)} \tag{8}$$

本文假设查询 q 中查询词是相互独立的,则公式(9)表示词 s 与原查询 q 之间的主题相似性 $p_{topic}(q \mid s)$ 等于 q 中每个查询词与该词主题相似概率的乘积。

$$p_{topic}(q \mid s) = \sqrt{\prod_{t_i \in q} p_{topic}(t_i \mid s)} \tag{9}$$

3.2.2 用户偏好分析模型

基于如下事实:具有不同应用背景、偏好的用户,输入的查询词能体现出各自不同的信息需求,即不同用户有不同的个性化偏好,如一些用户偏好查询一些与体育相关主题,另一些用户偏好查询一些与电脑游戏相关主题等。针对特定用户,进行查询替换时,替换词越与该用户的偏好相关,则生成的候选查询越能定位到该用户的查询意图。如上文所示,本实验为用户建立相关的查询个人描述,通过 PLSA 算法的 EM 步骤迭代运算后,得到用户的相关主题分布情况 $p(z \mid u)$,其不同的权值表示用户对该主题的偏好程度。基于如下思想:某词的主题分布与某用户主题的分布越相似则该词越能满足此用户偏好,其中,公式(10)采用 $sim(u,w)$ 衡量替换词 w 的主题与特定用户 u 偏好性之间的相关性,其值越大,二者主题相关性越大。

$$sim(u,w) = \frac{\sum_{z_i} p(z_i \mid u) \times p(z_i \mid w)}{\sqrt{\sum_{z_i} p(z_i \mid u)^2} \times \sqrt{\sum_{z_i} p(z_i \mid w)^2}} \tag{10}$$

3.3 候选查询生成模型

给定一个原查询 $q = w_1,\cdots,w_{i-1}w_iw_{i+1},\cdots,w_n$,遍历该查询中的所有查询词并将其替换,以此生成与原查询只存在一个词不同的候选查询,且候选查询根据候选查询词与原查询词之间的替换概率进行排序。其中,该模型的关键为如何生成候选查询词,本文主要分为以下两步:首先计算候选查询词语原查询之间的替换概率,并将其概率降序排列,然后根据排名前 N(本文取值为 15)的候选查询词与原查询词之间的互信息对其进行筛选,进一步确定满足相关要求的候选查询词。

3.3.1 候选查询词替换模型

文献[3]中的查询替换模型参见公式(11),$p(w_i \to s \mid q)$ 表示在查询 q 中,将词 s 替换为词 w_i 的概率。

$$P(w_i \to s \mid q) \propto t(s \mid w_i) \times p(w_1,\cdots, w_{i-1}\_w_{i+1},\cdots,w_n \mid s) \tag{11}$$

其中,$t(s \mid w_i)$ 表示词 s 与词 w_i 的上下文转移概率[参见公式(5)]。$p(w_1,\cdots,w_{i-1}\_w_{i+1},\cdots,w_n \mid s)$ 表示在由 q 确定的上下文中,词 s 出现的频次,本文在此只考虑与词 w_i 距离为 2 范围内的上下文词,计算方法参见公式(12)。

$$p(w_1,\cdots,w_{i-1}\_w_{i+1},\cdots,w_n \mid s) \propto \tilde{P}_{L_2}(w_{i-2} \mid s) \times \tilde{P}_{L_1}(w_{i-1} \mid s) \times \tilde{P}_{R_1}(w_{i+1} \mid s) \times \tilde{P}_{R_2}(w_{i+2} \mid s) \tag{12}$$

公式(13)在公式(12)的基础之上将替换词与查询之间的主题相关性考虑其中,$p_{topic}(q \mid s)$ 用于衡量 s 与查询 q 主题的相关性,参见公式(9)。其中,α 与 β 的权值之和为 1,α、β 的权值分别设定为 0.6、0.4。

$$P(w_i \to s \mid q) \propto \alpha \times t(s \mid w_i) \times p(w_1,\cdots, w_{i-1}\_w_{i+1},\cdots,w_n \mid s) + \beta \times p_{topic}(q \mid s) \times 10^6 \tag{13}$$

在公式(13)的基础之上,公式(14)将用户偏好考虑其中,$sim(u,s)$ 的计算参见公式(10)。其中,α、β 与 γ 的权值之和为 1,本文将 α、β、γ 的权值分别设定为 0.6、0.2、0.2。

$$P(w_i \rightarrow s \mid q,u) \propto \alpha \times t(s \mid w_i) \times p(w_1,\cdots,w_{i-1}\_w_{i+1},\cdots,w_n \mid s) + \beta \times p_{topic}(q \mid s) \times 10^6 + \gamma \times sim(u,s) \quad (14)$$

3.3.2 候选查询词筛选模型

为进一步确保候选查询生成的准确性,本文根据文献[17]的思想,利用两词之间在查询日志用户Session中互信息进一步确定原查询词与候选词之间的相关关系,参见公式(15),其中,X_s 与 X_w 分别表示某Session是否包含词 s 或 w 的二元值(0没出现,1出现),$P(X_s)$ 表示包含或者不包含词 s 的Session数与总Session数的比值,如 $P(X_s = 1)$ 表示包含词 s 的Session数与总Session数的比值,$P(X_w = 0)$ 表示不包含词 s 的Session数与总Session数的比值。$P(X_w)$ 的意义与 $P(X_s)$ 相同;$P(X_s,X_w)$ 表示词 s 与 w 在Session中的联合分布概率,如 $P(X_s = 1,X_w = 1)$ 表示同时包含词 s 与 w 的Session数占整个Session数的概率,$P(X_s = 0,X_w = 1)$ 表示不包含词 s 但包含词 w 的Session所占的比例。

$$I(s,w) = \sum_{X_s,X_w \in \{0,1\}} P(X_s,X_w)\log\frac{P(X_s,X_w)}{P(X_s)P(X_w)} \quad (15)$$

公式(16)用于对公式(15)的互信息计算方法进行规范化,其中,$NMI(s,w)$ 的取值区间为[0,1],$NMI(w,w) = 1$。本实验首先通过候选查询生成模型,生成相关候选查询词,然后通过设定阈值 κ 筛选出候选词与原查询不相关词,即当 $NMI(s,w)$ 值小于该阈值时,则剔除相关候选查询词,本文将 κ 值设为0.0015。

$$NMI(s,w) = \frac{I(s,w)}{I(w,w)} \quad (16)$$

4 实验及其结果分析

4.1 数据处理

本文采用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1所示,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在结果列表中的排序和点击的URL地址。因原始数据集中包含中许多噪音,本实验首先对其进行清理:如剔除导航类查询与包含色情词查询、停用词处理等。本文将数据集分为历史数据集和测试数据集两部分。其中,历史数据集包含前两个月的查询数据而测试集包含后一个月的数据。本文对历史查询中出现频次在前100,000的词进行上下文转移模型运算。笔者通过"15分钟划分法"[24]识别Session边界,剔除无任何用户点击的Session,并将具有相同初始查询的Session进行归并。为了生成用户描述文档,本文从历史数据集中随机选择了至少有100个Session的1000个用户,将每个用户在历史数据中的查询用户构建用户的描述文档,且在PLSA潜在主题分析时,设置潜在主题数为30,迭代次数为150次。

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1 AOL数据集格式

4.2 实验结果评测

基于如下事实:在每个Session中,当用户对当前查询不满意时,则会修改查询继续进行搜索,直到构建出能表达其信息需求的查询,则本实验将Session中查询分为两类,即满意查询和不满意查询[25]。其中,位于Session的结束,且该查询被提交后,用户至少点击了一个URL地址,则将该查询视为满意查询,而将位于Session开始的查询视为不满意查询。本文在测试数据集中从上文所述1000个用户的Session中选取一些不满意查询进行实验,即将Session中的第一个查询视为不满意查询,且所选取不满意查询满足以下条件:① 至少含有三个词;② 至少含有三个与该查询存在一词(除停用词外)不同的满意查询。笔者共获取245个满足以上条件的不满意查询,以及1 292个满意查询,平均为每个不满意查询选取了约5个满意查询,并将这些满意查询作为实验的基准答案。其中,所选取满意查询与不满意查询的分布情况参见表1所示。

本实验进行评测的主要方法为:对于测试集中

出现的查询 Q,找出用户在同一 Session 中使用 Q 后又构造的其他查询$\{Q_1,Q_2,Q_3,\cdots,Q_i\}$,对一种待评价推荐方法 R,如果 R 的推荐结果中包括$\{Q_1,Q_2,Q_3,\cdots,Q_i\}$,则认为推荐成功[2,13]。本文将实验返回的前 N 个结果进行评测,若基准答案出现在前 N 个候选查询中,则该查询被认为是成功的。本文把所选取的不满意查询作为实验输入,且采用的评测指标为 P@ N(Precision at N)和 Recip_Rank。其中,P@ N指标评价前 N 个推荐结果中相关查询所占的比例,鉴于平均每个不满意查询平均约有 5 个满意查询,本文设 N 的值为 1,3,5;Recip_Rank 指标衡量返回第一个满意查询的能力。笔者对三种查询替换方法进行了比较,如表 2 所示。其中,"baseline"是利用公式(11)(即文献[17]采用的方法)生成候选查询词的方法,"+topic"是在公式(11)基础上增加了主题分析的方法[参见公式(13)],"+topic+personalization"是在公式(13)的基础上增加了用户偏好分析的方法[参见公式(14)]。

表 1　所选取不满意查询与满意查询的分布情况

每个不满意查询包含的满意查询数	满足条件的不满意查询数	所占比例(≈)
3	70	30%
4	65	27%
5	47	20%
6	20	10%
7	16	6%
8	6	3%
9	2	1%
>9	5	3%

从表 2 可以看出,另两种推荐方法的准确度相对"baseline"方法最大提升比例为 28%。其中,"+topic"方法的准确度相对于"baseline"方法,最大提升比例为 20%,因"+topic"方法除考虑了词之间的上下文相似性,也考虑了词之间的主题相似性,说明在上下文相似的基础上考虑到与查询主题的相似性,对于提高查询替换的准确度有着一定意义;相对于"+topic"方法,"+topic+personalization"方法在"baseline"方法基础上准确度提高的幅度更大,最大提升比例为 28%,且在所有指标上都优于 +topic 方法。由此可见,在"+topic"方法的基础上,结合用户偏好分析,可进一步提高查询替换的准确性。实验结果同时显示,当 $N=1$ 和 $N=3$ 时,"+topic"和"+topic+personalization"方法在准确度提升方面效果显著,表明这两种方法能更早地返回满意查询,从 Recip_Rank 的指标也可以看出这一点。

5　总　结

本文基于 AOL 查询日志,主要探讨的是查询推荐中的查询替换,且是候选查询与原查询只存在一词不同的查询替换。在文献[17]所提出的上下文转移模型基础之上,本文除考虑词之间的上下文相似性,也采用了主题分析,并结合用户偏好,以此来生成候选查询。其实验结果表明,该方法获得较文献[17]中方法较好的实验效果。但该实验仍存在一些不足之处,笔者还需在未来工作中做以下几个方面的深入研究:① 该类方法的扩展,如何将此类方法用于多词查询替换以及查询推荐中的另外方面如查询添加、查询删除等;② 考虑词之间的相互依存性,即在计算词之间主题相似性时,如何考虑词之间的非独立性;③ 采用更精确的 Session 切分方法;④ 在各种方法组合时,进一步探索参数改变下的效果情况;⑤ 建立更有效的实验结果评测方法,即在进行实验评测时,综合考虑人工评价的效果,如设计相应界面,借用众包思想对实验结果进行评测。

表 2　实验结果评测

评测指标 / 相关方法	p@1	p@3	p@5	Recip_Rank
baseline	0.25	0.49	0.67	0.46
+topic	0.30(+20%)	0.54(+10.2%)	0.70(+4.5%)	0.51(+10.8%)
+topic+personalization	0.32(+28%)	0.56(+14.3%)	0.72(+7.5%)	0.53(+15.2%)

参 考 文 献

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

[2] 李亚楠, 王斌, 李锦涛. 搜索引擎查询推荐技术综述[J]. 中文信息学报, 2010, 24(6): 75-84.

[3] Sahami M, Heilman T D. A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets[C]// Proceedings of the 15th International Conference on World Wide Web, 2006: 377-386.

[4] Stoica E, Hearst M, Richardson M. Automating Creation of Hierarchical Faceted Metadata Structures [C]// Proceedings of NAACL HLT 2007, 2007: 244-251.

[5] Chen Y, Xue G R, Yu Y. Advertising Keyword Suggestion Based on Concept Hierarchy [C]//Proceedings of the International Conference on Web search and web data mining, 2008: 251-260.

[6] Xu Y, Jones G J F, Wang B. Query Dependent Pseudo-Relevance Feedback Based on Wikipedia [C]// Proceedings of SIGIR2009, 2009: 59-66.

[7] Jones R, Rey B, Madani O, et al. Generating Query Substitutions [C]//Proceedings of the 15th International Conference on World Wide Web, 2006: 87-396.

[8] 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.

[9] Shi X, Yang C C. Mining related queries from search engine query logs[C]//Proceedings of 15th International Conference on World Wide Web, 2006: 943-944.

[10] 李亚楠, 许晟, 王斌. 基于加权 SimRank 的中文查询推荐研究[J]. 中文信息学报, 2010, 24(3): 3-10.

[11] 朱小飞, 郭嘉丰, 程学旗, 等. 基于流形排序的查询推荐方法[J]. 中文信息学报, 2011, 25(2): 38-43.

[12] 王继民, 彭波. 搜索引擎用户点击行为分析[J]. 情报学报, 2006, 25(2): 154-162.

[13] Antonellis I, Garacia-Molina H, Chang C C. SimRank++: query rewriting through link analysis of the click graph[C]//Proceedings of VLDB, 2008: 408-421.

[14] Zhang W, Yan J, Sh.-Ch. et al. Termporal query substitution for ad search[C]//SIGIR'09. New York: ACM, 2009: 798-799.

[15] Chien S, Immorlica N. Semantic Similarity Between Search Engine Queries Using Temporal Corrlation[C]// Proceedings of International Conference on the World Wide Web, 2005: 2-11.

[16] Wang X, Zhai C. Mining Term Association Patterns from Search Logs for Effective Query Reformulation [C]// Proceeding of the 17th ACM Conference on Information and Knowledge Management, 2008: 479-488.

[17] Wan K W, Tan A H, Lim J H, et al. Topic Based Query Suggestions for Video Search [J]. LNCS, 2012, 7131(1): 288-299.

[18] 梅翔, 孟祥武. 一种基于用户偏好分析的查询优化方法[J]. 电子与信息学报, 2008, 30(1): 33-37.

[19] Sugiyama K. Adaptive Web Search Based on User Profile Constructed Without Any Effort From Users [C]// Proceedings of the 13th international conference on World Wide, 2004: 675-684.

[20] Mirco S, Susan G. Personalized Search Based on User Search Histories [C]//Proceeding of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005: 622-628.

[21] Liu F, Yu C, Meng M. Personalized Web Search for Improving Retrieval Effectiveness[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(1): 28-40.

[22] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis[J]. Machine Learning, 2001, 42(1): 177-196.

[23] AOL 日志主页[EB/OL][2012-03-14]. http://www.gregsadetsky.com/aol-data/.

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

[25] Li B, Wai L. Investigation of Web Query Refinement via Topic Analysis and Learning with Personalization[C]// Proceedings of SIGIR 2011, 2011: 9-12.

(责任编辑 马 兰)