软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(Suppl.(2)):162−177

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向查询意图的信息检索技术*

张志强, 彭晴晴, 谢晓芹, 冯晓宁

(哈尔滨工程大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 张志强, E-mail: zqzhang@hrbeu.edu.cn

摘 要: 用户兴趣和行为的多样性使得为不同用户提供更符合其查询意图的搜索结果成为一个具有挑战性的任务.Web 2.0 下的社会标签是用户为他们感兴趣的网页等对象进行标注行为的结果,用户用标签来描述自己感兴趣的话题.这些标签不但代表着用户的兴趣,而且是对网页承载信息的最好揭示.提出了面向用户查询意图的标签推荐方法,旨在把能够体现用户真正查询意图的标签选择出来.标签作为对查询关键词的补充,不仅可以弥补用户短查询的缺陷,而且可以根据标签与网页上曾被标注过的标签间的关系,更准确地判断用户查询意图与网页内容之间的相关度,从而把更符合用户查询兴趣的结果排在靠前的位置上.实验结果表明,该方法比现有的其他方法更有效,这也说明社会标注对更准确地捕捉用户真实查询意图确实有重要作用.

关键词: 查询意图;标签推荐;关键词;社会标签

中文引用格式: 张志强,彭晴晴,谢晓芹,冯晓宁.面向查询意图的信息检索技术.软件学报,2013,24(Suppl.(2)):162−177. http://www.jos.org.cn/1000-9825/13034.htm

英文引用格式: Zhang ZQ, Peng QQ, Xie XQ, Feng XN. Information retrieval techniques based on query intention. Ruan Jian Xue Bao/Journal of Software, 2013,24(Suppl.(2)):162−177 (in Chinese). http://www.jos.org.cn/1000-9825/13034.htm

Information Retrieval Techniques Based on Query Intention

ZHANG Zhi-Qiang, PENG Qing-Qing, XIE Xiao-Qin, FENG Xiao-Ning

(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Corresponding author: ZHANG Zhi-Qiang, E-mail: zqzhang@hrbeu.edu.cn

Abstract: Because of the diversity of users' interests and behavior, it is a great challenge to provide appropriate search results for different users. The social tags in Web 2.0 result from users' annotation behavior for pages that they are interested in. The user uses tags frequently to describe these topics, therefore the tags not only represent the user's query intention, but also reveal the Web information best. This paper presents a research for user-oriented query intention. It aims at adding the tags which can reflect the user's real query intention to the query. Tags as a supplement to the query keywords, not only can make up the defects of short queries, but also accord to the correlation between the query tags and Web content in sorting the related results. If the user just uses the keywords, the query log needs to be used to recommend high-quality tag as a supplement to the query keywords, and return more relevant results. Experimental results show that the proposed method is better in meeting the needs of users than other methods, making the search results more close to user requirements. This means that the social tags can make an important contribution to catching the query intention.

Key words: query intention; tag recommendation; keyword; social tag

互联网已成为全球最大的信息供应站,不断改变着人类的生活、学习和工作环境.在 Web 2.0 的环境下,用户已经开始从被动的接受信息转变为网络信息的组织者和发布者,这一转变进一步加速了网络信息的增长.

* 基金项目: 国家自然科学基金(61202090, 61272184); 教育部新世纪优秀人才支持计划(NCET-11-0829); 黑龙江省自然科学基金(F201130, F201016); 哈尔滨市科技创新人才研究专项基金(RC2010QN010024); 中央高校基本科研业务费专项资金(HEUCF100609, HEUCFT1202)

收稿时间: 2013-03-15; 定稿时间: 2013-07-11

我们现在常用的搜索引擎如 Google 和百度等虽然已经取得了很大的进步,并且已经进入社会化搜索的阶段,但是基本思路一直没有变,那就是需要根据用户提交的查询与信息之间的相关度为用户返回搜索结果,然而在返回的成千上万项结果中,仍然包含了大量用户不需要的信息[1],这些信息所占比例可达 75%[2].用户需要浏览返回结果列表中每一项的标题和摘要来确定返回结果是否符合自己的查询意图[3].因此,如何从海量的网络信息中查找到与用户查询意图匹配的信息仍然是一个需要进一步深入研究的课题,需要提出更高效的检索方法,帮助用户方便快速地获取有用的信息.

有调查显示,超过 80%的用户在使用搜索引擎时不会查看第 3 页以后的搜索结果,其中 41.2%的用户只选择查看第 1 页的搜索结果,25.8%的用户会查看前两页的搜索结果,只有 14.7%的用户会耐心查看完前 3 页的搜索结果,综上所述,总共有 81.7%的搜索引擎用户在查看完前 3 页的搜索结果后,就不会再查阅第 3 页以后的搜索结果[4].以上调查结果表明,搜索引擎需要有良好的排序算法,将与用户查询最相关的结果排在前面.因此,如何对搜索结果进行排序,使其更贴近用户需求,成为一个重要的研究问题.

无论是要将所有与用户查询意图相关的结果都找到,还是要对这些返回结果给出一个良好的排序,前提都是需要我们准确地捕捉用户的真实查询意图,否则很难获得用户满意的结果.

关键词查询表达方式以其简单、易用的特点,已经成为当前最流行的一种查询表达方式.已有实验结果表明,两个不同的用户在使用相同的查询关键词进行搜索时,其查询意图相同的概率小于 20%[5],即不同用户提交相同关键词搜索信息时,大部分情况下,其查询意图是不同的.然而,因为他们提交了相同的查询关键词,搜索引擎则会为他们返回相同的检索结果,这并不是我们想要的,我们希望能够捕捉到用户的查询意图而返回与之匹配的检索结果.这说明仅靠用户提供的查询关键词很难提供足够的信息来准确地描述用户的真实查询意图.

目前,关于捕捉用户查询意图方面的工作,主要有 3 类.第 1 类,查询扩展技术.主要是通过选择一些与用户查询有关的词来扩展到原查询中,使得新查询具有更加明确的语义.第 2 类,标签推荐技术.这类技术最开始并不是用来刻画查询意图,而主要是在一些添加标签的应用中为一些对象,如网页、图片等推荐标签.由于标签对资源对象具有更好的描述能力,因此人们开始将标签作为查询扩展词来源实现查询扩展.第 3 类,查询推荐技术.查询推荐技术推荐的是完整的查询,而不是几个扩展词,其目的是推荐出更符合用户查询意图的新查询.

1 相关工作介绍

1.1 查询扩展技术的研究与进展

查询扩展技术是一种提高查全率的有效方法.其基本思想就是在原来查询的基础上加入与用户查询词相关的词,组成一个更长的,能够更明确用户查询语义的新查询.这种方法可以在一定程度上改善用户查询语义信息不够明确的问题.从扩展词分析范畴角度来看,目前查询扩展技术大致可以概括为两大类:全局分析技术和局部分析技术.

全局分析技术是出现较早的具有实际应用价值的查询扩展方法.全局分析是对词项在全局的范围内进行统计分析.全局分析技术有以下几种:

对关键词进行聚类的查询扩展技术[6],是对文档的全部词语进行集中,根据词的共现次数进行聚类生成不同的簇.对于每一个用户查询,查找包含该用户查询的簇,并把簇中的其他关键词加入到用户查询中实现扩展,但这种方法容易将词分配到不同的类别中,使查询结果更含糊,进而降低查询结果的质量.

基于概念的查询扩展技术[7],其思想是利用关联规则提取查询词之间的关系,并绘制成图的形式,通过分析这些特殊的关系图,挖掘与当前查询相关的概念,并在原始查询中加入这些相关概念.这种方法的计算量比较大,所以效率不高.

分析潜在语义建立索引的查询扩展技术[8],其思想是从语义概念的层面上对查询进行扩展,通过统计关键词的共现信息,来发现关键词之间的重要关联关系,计算出与上下文相似的关键词,实现对原查询的扩展.这种方法查全率比较高,但只能解决部分一词多义的问题.

不同于全局分析,局部分析并不需要计算关键词的全局关系,而是针对用户查询进行两次搜索来实现扩展,

局部分析首先进行初次检索得到与用户查询最相关的前几个返回结果,并从这些返回结果中挖掘合适的词作为扩展词的来源.早在1977年Atter等人的研究[9]中就开始了对局部分析思想的探索.局部分析技术首先要基于一种假设,即进行初次检索得到的返回结果必须是相关的.但是这种方法存在一个很大的问题,当初次检索时得到的返回结果与用户查询的相关度并不高时,局部分析则会把大量不相关的词加入到用户查询中,这无疑会严重降低搜索时的准确度.典型的局部分析技术包括以下几种:

基于相关反馈的查询扩展技术[10],其主要思想是先使用初始查询对结果进行检索,由用户判断哪些结果是相关的,哪些结果是无关的,然后从那些相关的结果中选择有效的词加入到用户查询中,在新形成的查询中增强相关词的权重,降低与查询不相关的词的权重,其缺点是必须由用户提供相关性判断.

基于局部反馈的查询扩展技术[11],也是先要利用初始查询对结果进行检索,但它与相关反馈技术的区别是假设得到的检索结果都是与用户查询相关的,并对这些返回结果中的关键词进行聚类,最后将聚好类的关键词加入到初始查询中,从而对其进行扩展,但是该算法对初始检索的结果非常敏感,若初次检索得到的结果与用户查询相关度很低,则会降低检索性能.

基于局部上下文分析的(LCA)查询扩展技术[12,13],其思想是从初始检索出的结果中选出与原查询共现的词,计算每一个共现词与整个用户查询的相似度并按降序排列,排在前面的词则用作扩展.该算法一定程度上解决了全局分析中计算量大以及局部反馈方法中初始值敏感的问题.

除了上述工作之外,还有一些查询扩展方面的工作.比如,文献[14]提出了根据词之间的语义关系进行扩展和替换的文档重构方法,理论上讲,它将扩展词和被扩展词合并成同一个概念进行检索,通过相关子信息的聚集改进检索的效果,这与传统的仅对用户查询进行扩展的方法不同.文献[15]提出了一种词语-概念相关度的计算方法,该方法利用一些使用本体来标注和组织的文档,从“词语-文档-概念”所属程度和“词语-概念”共现程度两个角度描述“词语-概念”的相关度,以找到与查询语义主旨匹配的概念,进而提高查询效果.文献[16]考虑了 P2P 环境下的查询扩展问题,首先,利用查询与文档的关联关系构建了局部查询扩展;然后,基于查询与文档用词的直接关联,提出了基于历史信息的查询扩展方法.文献[17]则针对XML文档检索中,用户不能很好表达查询意图的问题,提出一种基于相关反馈的查询扩展方法.该方法不仅考虑了传统方法中的语义权重,还通过考虑 XML 文档的结构信息,分析了结构信息对于扩展词选择的影响.最终形成完整的“内容+结构”的查询扩展表达式.

1.2 标签推荐技术的研究与进展

基于用户生成的标签(Tag),为我们提供了一种新型的社会兴趣发现方法.标签可以灵活、有效地为用户的各种资源如图片、网页等添加一个或多个标签.标签与关键词类似,都是对资源信息的一种描述,但不同的是,提供关键词的往往只能是资源的创造者或发布者,而添加标签的可以是任意感兴趣的用户,所以标签可以轻松地帮我们确定社会热点、发现用户兴趣.

社会标签是 Web 2.0 技术的重要应用.它允许用户根据自己对资源的理解自由的选择合适的标签对其进行标注.标签是由用户自身产生的并且可以与他人共享的元数据,能够准确地反映用户对某个资源的理解和需求.

社会标签的推荐功能是标签系统中重要的功能之一.系统使用标签推荐算法,结合用户的信息需求,查询兴趣和行为,从标签集合中选择出合适的标签推荐给用户,用户则可以从推荐的标签中选取自己需要的标签,既方便了用户,又促进了信息资源的共享,对于用户寻找信息具有重要意义.根据推荐原理的不同,标签推荐算法主要分为两种:基于内容的标签推荐算法和基于协同过滤的标签推荐算法.

基于内容的推荐算法的理论依据主要来自于信息检索和信息过滤领域,是标签推荐中的基本方法,主要是根据用户过去的浏览记录来向用户推荐没有接触过的推荐项.这种方法最早的应用可以追溯到 Lang[18]和 Krulwich 等人[19]的文章.推荐方法主要有启发式的方法和基于模型的方法.启发式的方法是根据用户过去的经验确定相关的计算方法,由计算的结果和实际的结果进行比较验证,最后通过不断修改计算方法来实现推荐.基于模型的方法是利用以往的数据建立一个用于推荐的模型.基于内容的推荐是以资源的内容作为标签推荐的来源,一般来说使用的是文本内容.基于内容的标签推荐方法不易受新文档和冷门话题的限制,适用于有充足文本内容的资源,如网页等.

目前研究最多的是基于协同过滤的标签推荐算法.它是一种与基于内容的标签推荐算法基本思想完全不同的标签推荐方法.基于协同过滤的标签推荐算法是通过分析大量用户的兴趣和爱好,根据相似用户的需求推断当前用户最可能需要的信息[20],从而产生推荐.这类算法有多种不同的变种形式.比如,2001 年 Sarwar 等人提出了基于项目的协同过滤推荐算法[21];2007 年 Kuwata 等人提出了“一步到位协同过滤”的算法[22];Breese 等人则从概念的角度实现对标签分析预测,并结合协同过滤方面的研究提出了两种基于模型的协同过滤方法:贝叶斯聚类技术与贝叶斯网络技术[23].与基于内容的推荐相比,基于协同过滤的推荐方法能够推荐具有新颖性的标签,能够为用户发现新的兴趣.但基于协同过滤的推荐也存在着数据稀疏性、可扩展性和冷启动等问题.

除了上述工作之外,文献[24]提出了利用社会标签来实现个性化查询扩展的几种方法,将标签作为查询扩展词的来源取得了较好的效果.但是这类算法需要基于用户添加标签的历史信息来推荐扩展词,对于那些没有历史数据的非系统注册用户来说则无法给出好的推荐.

1.3 查询推荐技术的研究与进展

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[25]使用点击时间在点击行为图上推荐语义相关查询.Zhu 等人[26]通过利用带停止点的流行排序来推荐多样性查询.Guo 等人[27]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[28]提出的查询流图也被用于查询推荐.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 Bordino 等人[29]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[30]利用伪反馈的方法扩展长尾查询的信息,通过 URL 信息将 Clickgraph 和 Skip graph 的推荐结果融合在一起.Pandey 等人[31]对长尾查询中的广告效用进行了分析.Szpektor 等人[32]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[33]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,本文也使用了查询中的单词信息,但我们的方法是对查询意图进行建模,因而能产生更加相关的推荐结果.

文献[34]针对那些频度较低的长尾查询,提出了一种新的关于词项查询图概率混合模型.该模型能够准确地发掘出用户的查询意图.同时还提出了一种融合查询意图的查询推荐方法.该方法利用搜索引擎的查询日志来构建查询流图,进而实现查询推荐,并没有使用社会标注资源,推荐的是查询而不是标签.

2 面向查询意图的检索方法

2.1 面向查询意图的查询方式

传统的搜索引擎需要用户输入关键词集合,并且这些关键词需要出现在信息正文中,但是很多时候仅用简短的关键词进行查询,并不能很好地确定用户真正的查询意图.例如,用户输入关键词“java script”,那么我们无法确定其要查找 java script 方面的书,还是要查找相应开发方面的工具或技术论坛,用户必须对返回结果做进一步的分析.

既然用户可以在查看一个网页或图片之后,对其添加一些标签用来描述这些信息,用词虽然不多但可以很好地刻画用户对其的理解.那么自然也可以采用一些标签信息来描述自己的真实查询意图,毕竟用户对自己的查询意图理解得最准确.这种思路使得我们能够以一种开放式的思维来看待查询意图的描述,而不是原来的封闭式思考方式(查询中的关键词必须出现在原始的信息中).

为此,本文提出面向查询意图的查询方式,查询时用户在输入查询关键词之外,还可以根据他们的理解为其查询添加标签,例如,当用户要输入查询“java script”时,如果在后面添加标签“book”,则说明他要查找的是介绍 java script 的书,而不是相关的软件.因此我们可以将用户的查询形式化表示为

$$q=\{k_1,k_2,\ldots,k_m;t_1,t_2,\ldots,t_n\},$$

其中,$k_i(i=1,\ldots,m)$是传统的搜索关键词,$t_j(j=1,\ldots,n)$表示用户给查询添加的标签.

标签是用户对查询的一个补充描述,用来进一步明确其查询意图.与确定查询关键词的封闭式模式不同(查询关键词应出现在信息的正文中),标签的用词选择完全是开放式的模式,范围更加广泛,不必要求其一定包含在相关信息的正文中.

2.2 面向查询意图的排序方法

随着 Web 2.0 的快速发展,例如 Delicious,Flickr 和 YouTube 等社会标注应用越来越多,为我们提供了大量的社会标注资源.这样使得我们有机会通过标签建立用户查询意图与信息资源之间的桥梁,进一步判断两者的匹配程度.为了能够更好地捕捉用户的查询意图,本文采用了面向查询意图的工作思路,检索时把用户查询中的关键词集合和标签集合分别处理,关键词集合用来对网络中的所有资源进行筛选,用传统的信息检索方法返回匹配关键词集合的所有资源.然后用标签集合去匹配网页上被标注过的标签集合,用集合间的对应关系为首次检索返回的结果做重排序.

用关键词集合返回相关结果的过程,可由传统的搜索引擎完成.下面我们将主要对重排序的过程进行详细介绍.对相关结果进行重排序时我们考虑了两部分的因素,首先是根据关键词集合返回的相关结果列表中网页 P 的位置 s;其次是网页上的标签与用户查询中标签集合间的对应关系,综合这两部分得到网页的权重 $W(P)$,再根据 $W(P)$为这些返回的结果进行重排序.

首先把关键词集合提交给搜索引擎得到返回结果列表.记录每个网页 P 在结果列表中的位置 s,并计算每个网页的位置权重 $W(S)$,如公式(1)所示.

$$W(S)=\frac{1}{s} \tag{1}$$

通过对 Del.icio.us 网站上的标签进行统计,可以得到某个网页 P 上被标注过的次数最多的前 k 个标签,并按降序排列,记作 $T_1=\{t_{11},t_{12},\ldots,t_{1k}\}$.查询集合中的标签是用户自己给出,所以标签的个数是有限的,记作 $T_2=\{t_{21},t_{22},\ldots,t_{2n}\}$.网页 P 的标签权重 $W(T_1)$为 T_1 中所有标签权重的累加和,如公式(2)所示.对于每个 T_1 中的标签 t_{1i},其权重 $w(t_{1i})$计算方式如公式(3)所示.即对于网页 P 上被标注次数最多的前 k 个标签,其中每一个标签 t_{1i} 同时也出现在用户查询的标签集合 T_2 里,则其权重为 i 的倒数,没有出现权重则为 0.然后对网页 P 上这 k 个标签的权重进行累加求和得到网页 P 的标签权重.在本文中我们取 k=10,即我们仅对网页 P 上被标注次数最多的前 10 个标签与用户查询中的标签进行比较.

$$W(T_1)=\sum_{i=1}^{k} w(t_{1i}),\ t_{1i}\in T_1 \tag{2}$$

$$w(t_{1i})=\begin{cases}\frac{1}{i}, & t_{1i}\in T_2\\ 0, & t_{1i}\notin T_2\end{cases} \tag{3}$$

网页 P 的权重 $W(P)$为 P 位置权重和标签权重两部分的叠加,为了能够较好地控制位置权重和标签权重两部分的比重,我们引入系数 C_1 和 C_2,其中 $C_1+C_2=1$.$W(P)$的计算方式如公式(4)所示.

$$W(P)=C_1\cdot W(S)+C_2\cdot W(T_1) \tag{4}$$

根据 Li 等人的研究,发现标签比关键词更贴近于人们对内容的理解,并且资源中的很多关键词与资源真正的内容是毫不相关的[35].所以本文选取了 3 个用户查询,分别是{information retrieval,book},{web design, designer},{mineral composition,education},并给出这 3 个查询的用户根据自身的查询意图计算返回结果的 DCG 值,以便确定公式(4)中的系数.我们对系数取不同的值,并分别完成面向查询意图的检索,发现当系数 C_1=0.2,C_2=0.8 时,所有查询检索后返回结果的平均 DCG 值达到最大值 24.517,在 C_1=0.1,C_2=0.9,C_1=0.3,C_2=0.7,C_1=0.4,C_2=0.6 时的平均 DCG 值分别为 23.558,22.935,22.142,而且之后随着系数 C_1 的增大和 C_2 的减少,所有查询的平均 DCG 值越来越小,并且检索返回的结果越来越趋向于直接把查询关键词提交给搜索引擎时的结果,这说明随着标签在查询中所占比例的降低,它起的作用越来越小,甚至不起作用.因此我们确定公式(4)中的系数 C_1=0.2,C_2=0.8,因为在这种情况下,面向查询意图的检索效果达到最好,即一个网页在首次检索时返回的位置权

重占网页权重的 20%,而网页上的标签权重占网页权重的 80%.把系数 C_1,C_2,公式(1)~公式(3)代入公式(4),最终整理得到公式(5).

$$W(P)=0.2\cdot\frac{1}{s}+0.8\cdot\sum_{i=1}^{k}w(t_{1i}),\ t_{1i}\in T_1 \tag{5}$$

我们希望根据网页的 $W(P)$进行重排后得到的结果,不仅能够满足用户的个性化需求,而且排名越靠前的网页更贴近于用户的查询意图.图 1 给出了基于用户查询意图的重排序算法.

输入:用户提交的查询.
输出:相关结果.
步骤如下:
初始化:
1. *KeyWords*=用户查询中的关键词集合
2. *Tags*=用户查询中的标签集合
3. *Pages*=输入用户查询关键词搜索后返回的结果集合
4. For *i*=0 to *Pages.size* do
5. Begin
6. *S*=第 *i* 个结果的排序位置,*PTag*=第 *i* 个结果页面上的标签集合
7. 根据 *S* 与公式(1)计算结果页面位置权重,根据 *PTag* 和 *Tags*,利用公式(2)和公式(3)计算结果页面标签权重
8. 最后用公式(5)计算第 *i* 个结果页面的权重
9. End For
10. 按照权重从大到小的顺序重新排序
11. 返回结果

Fig.1 The re-ranking algorithm based on the user query intention

图 1 基于用户查询意图的重排序算法

2.3 基于查询日志的标签推荐方法

虽然用户可以根据自己的意愿和理解对查询添加标签,但是有时用户选择标签时也会遇到困难,例如,不知用什么标签来刻画自己的查询需求更合适.为了适应现有的搜索引擎搜索习惯,当用户只给出关键词集合并没有为他们的查询意图打标签时,系统可以为用户推荐标签.

给定一个关键词向量形式的查询,我们如何给其推荐适当的标签呢?我们根据现有的资源信息,提出了下面的概率生成模型来模拟用户给查询分配标注的行为:

1．用户搜索信息时,使用关键词 $w_i^{(q)}$作为搜索关键词的概率是 $P(w_i^{(q)})$.

2．以 $w_i^{(q)}$作为关键词提交查询后,用户选择网页(或图片等)D_k 的概率为 $P(D_k|\ w_i^{(q)})$.

3．当用户遇到网页 D_k 时,选择用标签集合 t 中第 j 个标签来标记 D_k 的概率为 $P(w_j^{(t)}|\ D_k)$.

这个模型实际上模拟了用户的两类行为.第 1 类,用户提交查询后,对返回结果的选择行为,这是通过点击相关结果的链接来体现的.第 2 类,用户对一个信息资源,如网页、文档、图片或视频等对象添加标签的行为.通过结合这两类行为,以间接的方式模拟了用户对一个查询添加标签的行为.

每个搜索引擎都积累了大量的查询日志.查询日志是发生在网站服务器上的所有用户搜索过程的记录.Cui 等人在 2002 年的研究中[36],利用查询日志为用户的查询推荐了合适的关键词做查询扩展,本文则利用查询日志作为关键词与标签间的桥梁,为用户查询推荐高质量的标签.

从查询日志提供的大量相关反馈数据中,我们可以知道,对于某个查询用户选择点击的资源都有哪些.由于用户数量巨大且相互独立,因此我们有理由认为:对于相同的查询,用户经常点击的文档集合所对应的标签集合最能体现用户的查询意图.这是基于一个假设,即用户点击的文档与用户查询是相关的.事实上,用户并不是随机地选择点击的文档,而且用户的选择确实表明了一个确定的相关度.这一点在关于查询日志方面的研究工作中得到了印证,而且进一步证明了基于文档点阅的方法比相关性反馈的方法更有效[37].

用户的查询日志为查询词和标签之间建立了一个桥梁.可以利用查询日志得到用户查询关键词和资源标签间的一个对应关系,当用户提交查询关键词时,根据查询日志中查询关键词与标签间的对应关系推荐高质量的标签.用户查询关键词、查询日志和资源标签之间的关系如图 2 所示.

通过分析这些大量的连接关系,利用概率可以计算得到查询关键词和标签之间的对应关系,如图 3 所示.

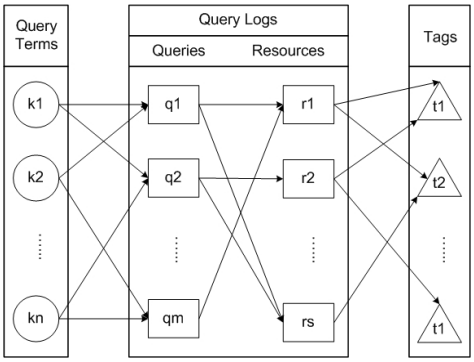


Fig.2 The relationship among query terms, query logs and the tags

图 2 用户查询词/查询日志/资源标签的关系

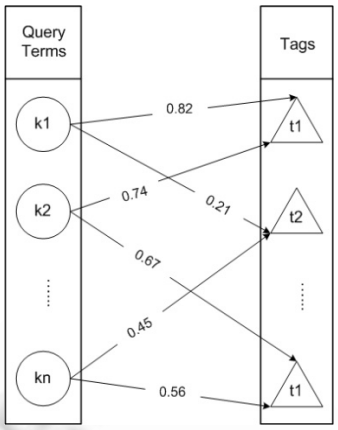


Fig.3 The probability relationship between query terms and tags

图 3 用户查询词和资源标签间的概率对应关系

查询关键词和标签之间的对应关系由概率来度量,通过查询日志可以计算得到用户提交的某个查询词和标签集合中的某个标签相对应的概率,即 $P(w_j^{(t)}|w_i^{(q)})$.$w_i^{(q)}$是用户提交的查询 q 中第 i 个查询关键词,$w_j^{(t)}$是标签集合 t 中第 j 个标签,所以概率 $P(w_j^{(t)}|w_i^{(q)})$可以用来度量当查询关键词 $w_i^{(q)}$被提交的情况下,标签 $w_j^{(t)}$被推荐的概率,概率 $P(w_j^{(t)}|w_i^{(q)})$可由公式(6)来计算:

$$P(w_j^{(t)}\mid w_i^{(q)})=\frac{P(w_j^{(t)},w_i^{(q)})}{P(w_i^{(q)})}=\frac{\sum_{\forall D_k\in S}P((w_j^{(t)},w_i^{(q)})\mid D_k)\cdot P(D_k)}{P(w_i^{(q)})}=\frac{\sum_{\forall D_k\in S}P(w_j^{(t)},w_i^{(q)},D_k)}{P(w_i^{(q)})}=\frac{\sum_{\forall D_k\in S}P(w_j^{(t)}\mid(w_i^{(q)},D_k))\cdot P(w_i^{(q)},D_k)}{P(w_i^{(q)})} \tag{6}$$

其中,D_k 是查询日志中用户点击的某个资源.S 是查询 q 被用户提交时,用户点击的资源集合.当 $D_k\in S$ 时,D_k 则代表用户提交查询 q 时,用户点击的一个资源.概率 $P(w_j^{(t)}|(w_i^{(q)},D_k))$其实是表示,当用户用 $w_i^{(q)}$作为查询关键词,并且点击了资源 D_k 的前提下,用户选择标记 $w_j^{(t)}$标记查询的概率.事实上,当前用户看到资源 D_k 并添加标签的行为,与关键词 $w_i^{(q)}$几乎没有关系,而是根据他对资源内容的整体认知和理解来添加的.由于查询词 $w_i^{(q)}$的发生与标签 $w_j^{(t)}$之间没有直接的关系,他们之间是通过 D_k 的发生来间接关联,同时又由于我们采用的传统 IR 技术要求只有那些包含关键词 $w_i^{(q)}$的资源 D_k 才会被返回,所以我们可以认为 $P(w_j^{(t)}|(w_i^{(q)},D_k))=P(w_j^{(t)}|D_k)$,这样,公式(6)可以继续推导为公式(7).

$$P(w_j^{(t)}\mid w_i^{(q)})=\frac{\sum_{\forall D_k\in S}P(w_j^{(t)}\mid D_k)\cdot P(w_i^{(q)},D_k)}{P(w_i^{(q)})}=\frac{\sum_{\forall D_k\in S}P(w_j^{(t)}\mid D_k)\cdot P(D_k\mid w_i^{(q)})\cdot P(w_i^{(q)})}{P(w_i^{(q)})}=\sum_{\forall D_k\in S}P(w_j^{(t)}\mid D_k)\cdot P(D_k\mid w_i^{(q)}) \tag{7}$$

条件概率 $P(D_k|w_i^{(q)})$是查询词 $w_i^{(q)}$出现在用户查询 q 中时,资源 D_k 被点击的概率,条件概率 $P(w_j^{(t)}|D_k)$是资源 D_k 被选择时,选择标签 $w_j^{(t)}$用来标注资源 D_k 的概率.概率 $P(D_k|w_i^{(q)})$和 $P(w_j^{(t)}|\ D_k)$可分别由用户查询日志和 Del.icio.us 标签网站上的统计数据得到,计算公式如公式(8)和公式(9)所示.

$$P(D_k\mid w_i^{(q)})=\frac{f_{ik}^{(l)}(w_i^{(q)},D_k)}{f^{(l)}(w_i^{(q)})} \tag{8}$$

$$P(w_j^{(t)} \mid D_k) = \frac{f_{jk}^{(d)}(w_j^{(t)}, D_k)}{f^{(d)}(D_k)} \tag{9}$$

公式(8)中,$f_{ik}^{(l)}(w_i^{(q)},D_k)$是在查询日志中查询词 $w_i^{(q)}$和资源 D_k 同时出现的次数,$f^{(l)}(w_i^{(q)})$是查询日志中包含查询词 $w_i^{(q)}$的所有查询的个数.

公式(9)中,$f_{jk}^{(d)}(w_j^{(t)},D_k)$是在 Del.icio.us 标签网站中,标签 $w_j^{(t)}$用来对资源 D_k 进行标注的次数,$f^{(d)}(D_k)$是 Del.icio.us 标签网站中,资源 D_k 被标注过的次数.

把公式(8)和公式(9)代入公式(7)中可得到公式(10).

$$P(w_j^{(t)} \mid w_i^{(q)}) = \sum_{\forall D_k \in S} \frac{f_{jk}^{(d)}(w_j^{(t)}, D_k)}{f^{(d)}(D_k)} \cdot \frac{f_{ik}^{(l)}(w_i^{(q)}, D_k)}{f^{(l)}(w_i^{(q)})} \tag{10}$$

当查询 q 被提交时,对于查询 q 中的每一个查询关键词 $w_i^{(q)}$,通过公式(10)返回所有相关的资源标签 $w_j^{(t)}$,联合所有查询关键词之后,标签 $w_j^{(t)}$对于查询 q 的权重可以由公式(11)计算得到.

$$W(w_j^{(t)}) = \sum_{\forall w_i^{(q)} \in q} P(w_j^{(t)} \mid w_i^{(q)}) \tag{11}$$

因此,对于每一个查询 q,我们可以得到一系列的候选标签.根据公式(11)计算的标签权重对候选标签进行排序,选择权重最大的 Top-k 标签对查询 q 进行推荐,图 4 给出了标签推荐算法.

输入:用户提交的查询关键词.
输出:推荐标签.
步骤如下:
1. 初始化:
   List⟨*String*⟩*Keywords*=用户输入的原始查询 *q*
2. For *i*=0 to *Keywords.size* do
   Begin
3. 用公式(8)~公式(10)计算得到与第 *i* 个查询关键词相关的标签及其被推荐的概率
   End For
4. 用 WordNet 处理所有标签,并将标签及其概率放入到 *listTags* 中
   Map⟨*String*,*Integer*⟩*listTags*=与查询 *q* 相关的标签和标签的概率
5. 利用公式(11)将所有相关标签的概率进行累加求和,即得到标签被推荐的概率,
   根据概率对标签进行降序排列,从而进行标签推荐.
6. 返回结果

Fig.4 Tag recommendation algorithm based on user query intention

图 4 面向用户查询意图的标签推荐算法

实际当中对于给定的一个查询,如何选择 Top-k 中的合适的 k 值呢?下面我们通过一个例子来介绍本文采用的确定 k 值的方法.例如,对于查询“java script”,按照前面介绍的方法可以得到一组被推荐标签的概率,其中我们发现推荐的标签与用户提交的关键字之间存在重复现象,这些与关键词重复的标签在推荐时并不会推荐给用户,但是在重排序时还会得到应用,因为这些出现在网页正文中的词,又被用户选中作为标签而添加到网页上,说明这些词更能刻画网页的主要内容和用户的观点看法.因此,这些重复的标签起到了一种类似查询扩展的作用,同时又以标签的身份进一步被用来评估网页与查询意图的匹配程度.对于查询“java script”,去除掉与用户查询关键字重叠的标签后剩余的标签见表 1.

为了更直观地查看被推荐的标签间的概率关系,我们把表 1 中的数据绘制成曲线图,如图 5 所示.从图 5 中我们可以清楚地看到,在第 5 个标签“tool”后,折线的形状有一个明显的变化,被推荐的标签概率慢慢趋于平缓,并且概率值都较小,这说明标签“tool”以后的所有相关的标签将不具有明显的区分性,即不如前面的标签适合用于推荐.我们认为概率比标签“tool”大的标签是与用户的查询意图强烈相关的,而概率值小于等于“tool”的标签则是不适合用于推荐的,所以我们需要得到概率值大于“tool”概率的所有标签.

我们知道,曲线二次导数的几何意义可以体现函数的凹凸性,当二次导数大于 0,表示图形是向下凹的,值越大,说明图形变化趋势越明显.二次导数小于 0,表示图形是向上凸的,值越小表明图像变化越明显.即二次导数的

波峰所对应的点在图形上是曲线的极小值,所以我们选择用求二次导数的方法确定曲线变化趋势较明显的点,二次导数曲线的波峰则对应是曲线上有较强变化的转折点.

Table 1 Probability of tags recommended by algorithm for query “java script”

表 1 为查询“java script”推荐的标签及其概率

序号	Tag	Tag_*P*	序号	Tag	Tag_*P*
1	web	0.003 662 128	8	html	0.001 410 013
2	programming	0.003 627 341	9	design	0.001 127 254
3	tutorial	0.002 875 274	10	code	0.000 846 007
4	development	0.002 058 739	11	ajax	0.000 803 392
5	tool	0.001 601 931	12	resource	0.000 681 395
6	software	0.001 583 217	13	computer	0.000 661 135
7	css	0.001 437 799	…	…	…

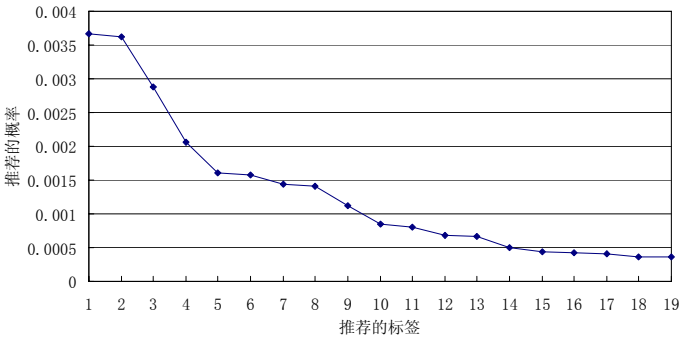


Fig.5 The tags’ probability curve of query “java script”

图 5 查询“java script”被推荐标签概率曲线

对图 5 中的曲线利用 Matlab 工具分别进行一次求导和二次求导后得到如图 6 所示的曲线.从中可以明显看到,二次导数曲线在第 5 个标签处达到最大,正是图 5 中我们要求的点,所以本节中选取的 Top-*k* 标签的方法就可以用对标签概率曲线求二次导数的方法实现.

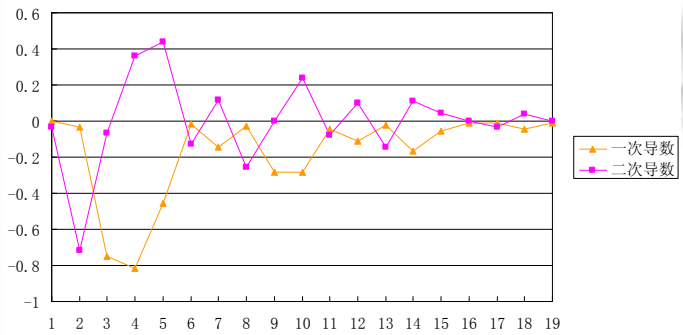


Fig.6 The first derivative and second derivative of probability curve of query “java script”

图 6 查询“java script”被推荐标签概率曲线的一次导数和二次导数曲线

3 实验方案

3.1 数据集的获取

本文的查询集合分为两部分,一是用户能够根据自己的查询意图给出合适的标签,二是用户只给出了关键字,没有为自己的查询意图打标签.在第 1 种情况下,我们从 TREC 数据集[38]的不同领域中选取了不同的查询主

题,分别由20个用户选择感兴趣的主题,并根据自己的查询意图给出查询关键词和标签,见表2.在第2种情况下,我们从 AOL 搜索日志中抽选出十个不同的查询关键词,并用本文提出的方法为这些查询推荐标签,然后让 20 个用户根据已有的关键词确定自己的查询意图,以此来评估本文算法推荐标签的质量,见表 3.表 3 中的第 3 列表示,20 个用户提交第 1 列中查询关键词时,其查询意图的统计结果,例如查询关键词“free music”,有 10 个用户的意图是下载免费音乐,6 个用户的意图是下载免费 MP3 格式的音乐.这表明本文方法推荐的标签“MP3”和“download”可以覆盖80%的用户查询意图.对于表中的9个样例查询,本文方法推荐的标签可以覆盖70%以上的用户查询意图.对于查询“animal”来说,20 个用户的查询意图分歧比较大,有的是查找动物学方面的信息,有的是查找宠物方面的信息,有的是查找词义等,本文算法选择推荐了一个标签“science”,可以覆盖 45%的用户,如果推荐更多的标签,则可以获得更好的覆盖率.这也从用户的角度证明了本文算法推荐的标签具有较好的可靠性和质量.

Table 2 Query set with tags annotated by user

表 2 用户给出标签的查询集

User ID	User’s query intention	Keywords	Tags
User 1	查找信息检索方面的书	information retrieval	book
User 2	查找信息检索研究领域里,涉及科学方面的资源	information retrieval	science
User 3	查找信息检索领域里可以用来做参考的资料,而不仅仅是参考书、参考题之类	information retrieval	reference
User 4	查找介绍信息检索的文献或者论文	Information retrieval	paper
User 5	查找可以作为教材的 Web design 方面的资料	web design	tutorial
User 6	查找 Web design 领域设计师的个人信息	web design	designer
User 7	查找有关 Web design 方面的机构	web design	agency
User 8	查找矿物组成研究领域中涉及科学研究的知识	mineral composition	science
User 9	查找可以用作教育的矿物组成方面的资料	mineral composition	education
User 10	查找有关小提琴的介绍信息	violin	introduction
User 11	查找小提琴演奏的古典音乐	violin	classical
User 12	查找纽约市就业信息里有关 IT 行业的信息	New York employment information	IT
User 13	查找政府机关发布的有关纽约市的就业信息	New York employment information	government
User 14	查找纽约市就业信息里有关教师的信息	New York employment information	teacher
User 15	查找股票行业里和经济相关的信息	stock	economy
User 16	找着股票行业里和金融界相关的信息	stock	finance
User 17	查找免费的 PPT 模板的资源	PPT template	free
User 18	查找可以下载的 PPT 模板资源	PPT template	download
User 19	查找 PPT 模板里用到的工具	PPT template	tool
User 20	查找办公专用的 PPT 模板资源	PPT template	office

Table 3 Query set with tags recommended by algorithm

表 3 算法推荐标签的查询集

Keywords	Recommended tags	User’s query intention
free music	mp3 download	50%的用户要下载免费音乐,30%的用户要下载 mp3 格式的免费音乐
elearning	technology online	30%的用户在进行网络学习时是学习与技术相关的知识,40%的用户进行在线学习
design	inspiration art architecture	35%的用户要查找设计灵感,45%的用户查找与艺术相关的设计,例如建筑设计,婚纱设计等
movie reviews	film entertainment	80%的用户搜索的是娱乐网站的影评
phone book	directory	85%的用户在搜索电话簿时是想查找电话目录
social security	government	70%的用户在查找社会保障时是想查找政府颁布的条例
sports club	fitness health	75%的用户在搜索运动俱乐部时是想查找能帮助自己健身的资料
animal	science	45%的用户在搜索动物科学方面的资料
virtual worlds	games	90%的用户在搜索虚拟世界时是为了查找游戏
java script	web 2.0 programming tutorial development	55%的用户搜索 java 脚本是为了编程,30%的用户是想查找用于教学的 java 脚本资料

查询日志的数据来自于 2006 年 AOL 提供的 AOL 搜索日志[39],数据收集于 2006 年 3 月 1 日~2006 年 5 月

31 日.标签数据均为从 Del.icio.us 网站上抓取下来的,实验数据收集于 2010 年,数据量接近 5.07G,表 4 给出了数据集的基本信息.其中,|*D*|代表网页总数,|*W*|代表网页中词的总数,|*M*|代表网页中去重的词的总数,|*N*|代表标签总数,|*T*|代表去重的标签总数,|*U*|代表用户总数.

Table 4 Basic information of data set

表 4 数据集的基本信息

Data set	|*D*|	|*W*|	|*M*|	|*N*|	|*T*|	|*U*|
Numbers	260 703	2 914 424	149 908	1 090 230	83 407	2 936

3.2 评估度量方法

本文采用了两种比较典型的搜索系统评价方法对实验中的数据进行评估,分别是 11 点 P-R 曲线方法[40]和 DCG 的评价方法[41].

(1) 11 点 P-R 曲线方法

在 *N*-point P-R 曲线[40]的评估方法中,*N* 的取值通常为 11,即 11-point P-R 曲线.使用方法是对于每一个给定的查询,设定 11 个召回率的点作为横坐标,分别记作:0.0,0.1,0.2,…,1.0,然后在每一个召回率的点分别计算检索结果在该点的准确率.

$$P=\frac{\text{返回的相关文档数}}{\text{返回的所有文档数}},\ R=\frac{\text{返回的相关文档数}}{\text{所有相关文档数}} \tag{12}$$

对多个查询计算其准确率的平均值[40],则可以得到多个查询的平均 11-point P-R 曲线,11-point P-R 曲线的特点是越靠上则说明该查询算法的性能越好.

(2) DCG 评价方法

DCG(discounted cumulative gain)评价方法是由 Järvelin 和 Kekäläinen 在 2000 年提出的一种多级制相关性评价方法[41].该方法的主要思想是:对于每个搜索结果,按照其与用户查询意图的相关程度赋予一个[0,*k*]区间的分值,*k* 可以视具体的情况事先设定.例如在本文中,*k* 的取值可以取 3,返回结果的分值是人为判断给出.

DCG 方法考虑了排序位置对相关度的影响,每个结果的排序位置都会影响它的分值,对其进行一定的打折,这就符合了搜索时,排序越靠前的结果则越符合用户查询需求的思想,其计算公式如式(13)所示:

$$DCG_q=\sum_{j=1}^{m}\frac{2^{r(j)}-1}{\ln(1+j)} \tag{13}$$

其中,*j* 代表返回结果的位置,*m* 表示搜索时返回的结果数,*q* 是用户搜索时提交的查询,*r*(*j*)表示在位置 *j* 处的结果和用户查询的相关度值.对于不同的搜索方法,DCG 值越大,则说明该搜索方法的返回结果越好.

3.3 实验结果分析

表 2 中的用户不但给出了自己搜索时用到的查询关键词,还为自己的查询意图添加了标签.表 3 中当用户只给出搜索用到的查询关键词而没有为自己的查询意图打标签时,通过标签推荐方法为用户推荐了标签.对于表 2 和表 3 中的所有查询集,我们都可以得到一个查询关键词集合和一个标签集合.为了验证此标签集合是对关键词集合的一个强有力的补充,并且可以有效弥补用户输入短查询词语义不明的缺点.我们考察了 3 种方式:第 1 种,直接将关键词提交给 Google;第 2 种,把查询集中的关键词集合和标签集合合并到一起提交给 Google;第 3 种,本文面向查询意图的方法.最后比较 3 种方式返回结果的平均 11 点 P-R 曲线.然后再采用 DCG 评价方法对本文方法与文献[42]的方法以及 Google 进行了对比.

3.3.1 11 点 P-R 曲线方法

要使用 11 点 P-R 曲线判断搜索时返回结果集的质量,首先要判断返回的每条结果与用户查询的相关性,因此根据它们之间的相关性,请专家为每个结果打分,分值从 0 到 3,具体的打分标准如下:

- 0,结果内容与用户查询意图完全不相关.
- 1,结果内容与用户查询意图基本不相关,但包含少量的相关内容.

• 2,结果内容与用户查询意图基本相关,但包含少量不相关的内容.

• 3,结果内容与用户查询意图完全相关.

当搜索结果的相关度分数大于等于 2 时,则认为是相关的,当小于 2 时则认为是不相关的.我们取第 30 个相关结果作召回率为 1.0 的点,也就是前 3 个相关的结果判别为召回率为 0.1 的点,然后计算在该点的平均准确率.对应表 2 得到了如图 7 所示的平均 11-point P-R 曲线,对应表 3 得到了如图 8 所示的平均 11-point P-R 曲线.

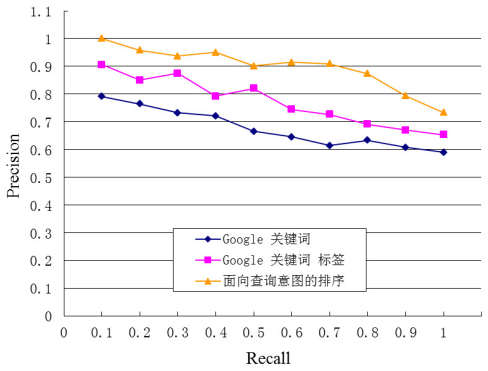


Fig.7 11-Point P-R curve of queries from Table 2

图 7 表 2 中查询集的平均 11-point P-R 曲线

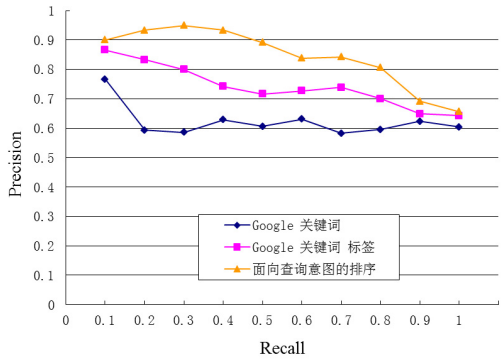


Fig.8 11-Point P-R curve of queries from Table 3

图 8 表 3 中查询集的平均 11-point P-R 曲线

从图 7 中可以明显地看出,当把用户标签和用户查询关键词一起合并提交给 Google 搜索引擎时,搜索的查全率和查准率都比仅用关键词查询要好一些,此时用户标签可以看作是对查询关键词的查询扩展,这说明用户标签中包含能够覆盖用户查询意图的词,可以起到弥补用户输入短查询词语义不明的缺点.而当用本文提出的面向查询意图的排序方式为用户返回查询结果时,发现面向查询意图的排序方式比另外两种方式还要好一些,这说明用网页上被标注的标签为网页排序要比仅仅用标签做查询扩展更容易满足用户的查询需求.

图 8 是用户没有给自己的查询添加标签,采用本文的标签推荐算法为查询推荐了标签.从图中我们可以看出,推荐标签后,把这些标签和关键词一起放在 Google 搜索引擎上搜索效果要比没有标签时好,而且面向查询意图的排序方式更能满足用户的查询需求,比另外两种方法的查全率和查准率都要高,这说明我们的推荐算法已经在一定程度上满足了更多用户的查询需求.

3.3.2 DCG 评价方法

由于实验中需要对传统的搜索系统返回的结果进行重排序,所以这里我们取返回结果的前十个结果,为其进行网页的权重计算,并根据权重对其进行重排序,即 m=10,实验中要对前 10 个结果打分并进行评价.对应表 2 得到了如表 5 所示的实验结果,对应表 3 得到了如表 6 所示的实验结果.图 9 和图 10 是实验结果的柱状图比较.

从表 5 中可以看出,当 Google 返回结果的 DCG 值相对较高时,根据标签重排序后会得到更高的 DCG 值,这是因为面向查询意图的排序方法把相关的结果排在了更靠前的位置上,从而计算得到的 DCG 值则会更高一些.当 Google 首次返回结果的 DCG 值很低时,面向查询意图的排序的值也会较低,这是因为面向查询意图的排序是基于 Google 返回结果的,所以 Google 返回的结果质量影响了面向查询意图排序的结果质量.但是通过实验我们可以看出,无论首次返回结果质量如何,通过标签对返回结果进行重排序的方法,每个查询的 DCG 值都会有较大程度的提高,这说明本文面向查询意图的排序方法把质量高的结果排在了更靠前的位置上.而且几乎所有的情况,本文的方法也要好于文献[42]中提出的排序方法.这更说明标签的位置这一因素的重要性,结合查询关键词与标签为用户返回相关结果,是很容易捕捉用户的查询意图的,并且返回结果的质量也较为理想.

表 6 中推荐的标签大部分都可以帮助用户返回更贴近用户需求的结果,并将其排在靠前的位置上.当推荐的标签合适时,本文方法的 DCG 值会明显高于只用关键词搜索时的 DCG 值.虽然文献[42]中提出的排序方法也在一定程度上提高了排序质量,但是由于文献[42]中提出的排序方法并没有考虑标签的位置因素,所以排序结

果并没有本文提出的面向查询意图的排序结果效果好.

Table 5 DCG value of queries from Table 2

表 5 表 2 中查询的 DCG 结果

User ID	Keywords	Google 的 DCG 值	Users' Tag	文献[42]方法的 DCG 值	本文方法的 DCG 值
1	information retrieval	20.191	Book	25.307	25.834
2	information retrieval	6.660	Science	9.1243	10.267
3	information retrieval	17.887	Reference	21.057	21.257
4	Information retrieval	9.420	Paper	15.170	15.170
5	web design	12.393	Tutorial	13.681	19.353
6	web design	10.560	Designer	12.853	12.306
7	web design	15.341	Agency	23.517	23.839
8	mineral composition	18.902	Science	22.686	25.215
9	mineral composition	11.546	Education	14.277	15.262
10	violin	19.650	Introduction	19.683	20.762
11	violin	8.644	Classical	11.811	16.742
12	New York employment information	15.569	IT	16.960	21.467
13	New York employment information	15.236	Government	15.411	16.032
14	New York employment information	14.828	Teacher	17.715	18.123
15	Stock	11.594	Economy	15.810	17.649
16	Stock	29.509	Finance	35.208	36.181
17	PPT template	17.103	Free	21.615	22.142
18	PPT template	13.501	Download	16.740	17.747
19	PPT template	18.870	Tool	21.662	23.284
20	PPT template	19.986	Office	20.145	20.762
Average	–	15.370	–	18.522	19.970

Table 6 DCG value of queries from Table 3

表 6 表 3 中查询的 DCG 结果

Keywords	Google 检索的 DCG 值	推荐的标签	文献[42]方法的 DCG 值	本文方法的 DCG 值
free music	31.339	mp3 download	32.142	32.761
elearning	18.257	technology online	18.552	21.381
design	17.967	inspiration art architecture	19.712	25.326
movie reviews	29.159	film entertainment	27.545	31.893
phone book	24.319	directory	25.137	29.085
social security	16.673	government	18.736	19.821
sports club	17.224	fitness health	19.813	17.982
animal	18.571	science	14.317	16.317
virtual worlds	20.436	games	25.619	26.712
java script	29.221	web 2.0 programming tutorial development	30.187	33.216
Average	22.317	–	23.176	25.449

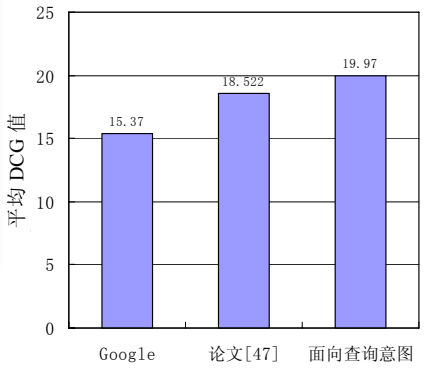


Fig.9 DCG histogram of Table 5

图 9 表 5 计算的平均 DCG 值对比

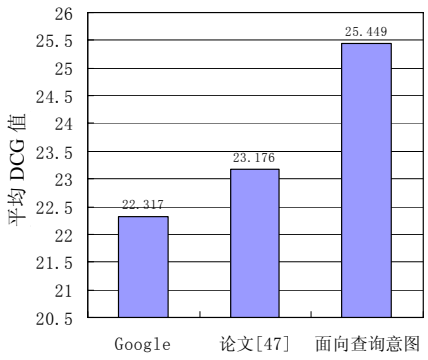


Fig.10 DCG histogram of Table 6

图 10 表 6 计算的平均 DCG 值对比

图 9 和图 10 分别是对表 5 和表 6 中最终实验结果的柱状图表示.它们可以更鲜明地展示本文提出的面向

查询意图的排序方法要优于文献[42]中提出的排序方法,而且 DCG 值比 Google 要高.并且通过用户查询日志为用户推荐的标签也能较好地满足大部分用户的查询需求,帮助用户返回最贴近其查询意图的结果.通过以上各种实验的比较,表明了面向查询意图排序方法的可用性和有效性,也表明了标签对信息检索的重要性.

4 结 论

针对现有关键词查询方式刻画用户查询意图不足的问题,我们将用户查询与网页、图片等同看待,采用了通过对查询添加标签来明确查询意图的研究思路,提出了根据用户查询日志为用户推荐标签的方法,以及利用查询的标签信息重新排序的算法.

与现有的查询扩展工作相比,本文推荐词的来源是已有的社会标签,推荐的将不再是几个同义词或关联词,而是一类标签,这类标签不是与某个领域有关,而是与多个领域有关,这些领域不仅是原 Web 信息内容所能够涵盖的领域,而且还包括大众对该 Web 信息的认识与理解,这些理解不一定被该信息的内容所涵盖.与现有的标签推荐工作相比,本文方法是对用户查询意图推荐标签,用来提高搜索引擎的质量,而不是在标签服务系统中为资源和图片等推荐标签.与现有的查询推荐工作相比,推荐的对象和目的都不同,查询推荐是通过推荐其他查询来代替原始查询来明确查询意图,而本文方法的目的则是通过推荐标签来捕捉用户的查询意图.

在后续工作中,我们将探讨社会标注在查询推荐中的应用.现有查询推荐还仅仅依赖于搜索引擎的日志来进行,尚未考虑利用已有的社会标注资源来实现推荐.毕竟,用户对一个网页添加标签的行为与选用一组关键词来搜索这个网页,两者之间既有关联又有不同,因为两者的出发点是不同的.

References:

[1] Zhang D, Dong Y. Semantic, hierarchical, online clustering of Web search results. In: Proc. of the 6th Asia Pacific Web Conf. Berlin, Heidelberg: Springer-Verlag, 2004. 69−78.

[2] Selberg E, Etzioni O. Multi-Service search and comparison using the MetaCrawler. In: Proc. of the Fourth Int'l WWW Conf. Boston. 1995.

[3] Waxman BM. Routing of multipoint connections. IEEE Journal on Selected Areas in Communications, 1988,6(9):1617−1622.

[4] http://soft.yesky.com/tools/25/2062525.shtml

[5] Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in human-system communication. Communications of the ACM, 1987,30(11):964−971.

[6] Jones KS. Automatic Keyword Classification for Information Retrieval. London: Butterworths, 1971.

[7] Qiu Y, Frei HP. Concept based query expansion. In: Proc. of the 16th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1993.160−169.

[8] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990,41(6):391−407.

[9] Attar R, Fraenkel AS. Local feedback in full-text retrieval systems. Journal of the ACM, 1977,24(3):397−417.

[10] Salton G. The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, 1971.

[11] Salton G, Buckley C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 1990,41(4):288−297.

[12] Buckley C, Salton G, Allan J, Singhal A. Automatic query expansion using SMART: TREC 3. In: Harman DK, ed. The 3rd Text Retrieval Conf. Department of Commerce, 1995. 69−80.

[13] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: ACM press, 1999.

[14] Zhang M, Song RH, Ma SP. Document refinement based on semantic query expansion. Chinese Journal of Computers, 2004, 27(10):1395−1401 (in Chinese with English abstract).

[15] Tian X, Du XY, Li HH. Computing term-concept association in semantic-based query expansion. Ruan Jian Xue Bao/Journal of Software, 2008,19(8):2043−2053 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2043.htm [doi: 10.3724/SP.J.1001.2008.02043]

[16] Zhang Q, Zhang X, Liu JR, Sun Y, Wen XZ, Liu Z. Query expansion and its search algorithm in hybrid peer-to-peer networks. Ruan Jian Xue Bao/Journal of Software, 2006,17(4):782−793 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/782.htm [doi: 10.1360/jos170782]

[17] Wan CX, Lu Y. Structural query expansion based on weighted query term for XML documents. Ruan Jian Xue Bao/Journal of Software, 2008,19(10):2611−2619 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2611.htm [doi: 10.3724/SP.J.1001.2008.02611]

[18] Lang K. Newsweeder: Learning to filter netnews. In: Proc. of the 12th Int'l Conf. on Machine Learning. 1995. 331−339.

[19] Krulwich B, Burkey C. Learning user information interests through extraction of semantically significant phrases. In: Proc. of the AAAI Spring Symp. on Machine Learning in Information Access. 1996. 100−112.

[20] Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews. In: Proc. of the 1994 ACM Conf. on Computer Supported Cooperative Work. New York: ACM Press, 1994. 175−186.

[21] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proc. of the 10th Int'l Conf. on World Wide Web. New York: ACM Press, 2001. 285-295.

[22] Kuwata S, Ueda N. One-Shot collaborative filtering. In: Proc. of the IEEE Symp. on Computational Intelligence and Data Mining, 2007 (CIDM 2007). 2007. 300−307.

[23] Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, Inc., 1998. 43−52.

[24] Zhang ZQ, Meng QH, Xie XQ. Research on personalized query expansion techniques based on social bookmarks. Journal of Frontiers of Computer Science and Technology, 2010,4(9):812−829 (in Chinese with English abstract).

[25] Mei QZ, Zhou DY, Church K. Query suggestion using hitting time. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 469−478.

[26] Zhu XF, Guo JF, Cheng XQ, Du P, Shen HW. A unified framework for recommending diverse and relevant queries. In: Proc. of the 20th Int'l Conf. on World Wide Web. New York: ACM Press, 2011. 37−46.

[27] Guo JF, Cheng XQ, Xu G, Shen HW. A structured approach to query recommendation with social annotation data. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2010. 619−628.

[28] Boldi P, Bonchi F, Castillo C, Donato D, Gionis A, Vigna S. The query-flow graph: Model and applications. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 609−618.

[29] Bordino I, Castillo C, Donato D, Gionis A. Query similarity by projecting the query-flow graph. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2010. 515−522.

[30] Song Y, He LW. Optimal rare query suggestion with implicit user feedback. In: Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 901−910.

[31] Pandey S, Punera K, Fontoura M, Josifovski V. Estimating advertisability of tail queries for sponsored search. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2010. 563−570.

[32] Szpektor I, Gionis A, Maarek Y. Improving recommendation for long-tail queries via templates. In: Proc. of the 20th Int'l Conf. on World Wide Web. New York: ACM Press, 2011. 47−56.

[33] Bonchi F, Perego R, Perego R, Silvestri F. Recommendations for the long tail by term-query graph. In: Proc. of the 20th Int'l Conf. Companion on World Wide Web. New York: ACM Press, 2011. 15−16.

[34] Bai L, Guo JF, Cao L, Cheng XQ. Long tail query recommendation based on query intent. Chinese Journal of Computers, 2013, 36(3):636−642 (in Chinese with English abstract).

[35] Li X, Guo L, Zhao YH. Tag-Based social interest discovery. In: Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 675−684.

[36] Cui H, Wen JR, Ma WY. Probabilistic query expansion using query logs. In: Proc. of the 11th Int'l Conf. on World Wide Web. New York: ACM Press, 2002. 325−332.

[37] Wen JR, Nie JY, Zhang HJ. Clustering user queries of a search engine. In: Proc. of the 10th Int'l Conf. on World Wide Web. New York: ACM Press, 2001. 162−168.

[38] Bailey P, Agrawal D, Kumar A. TREC 2007 enterprise track at CSIRO. In: Proc. of the TREC 2007. Gaithersburg, 2007. 205−210.

[39] Arrington M. AOL proudly releases massive amounts of private data. TechCrunch, 2006. http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

[40] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

[41] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents. In: Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2000. 41−48.

[42] Zhang ZQ, Liang TT, Xie XQ. An user-label based search result ranking algorithm. Journal of Computer Research and Development, 2009,46(Suppl.):351−358 (in Chinese with English abstract).

附中文参考文献:

[14] 张敏,宋睿华,马少平.基于语义关系查询扩展的文档重构方法.计算机学报,2004,27(10):1395−1401.

[15] 田萱,杜小勇,李海华.语义查询扩展中词语-概念相关度的计算.软件学报,2008,19(8):2043−2053. http://www.jos.org.cn/1000-9825/19/2043.htm [doi: 10.3724/SP.J.1001.2008.02043]

[16] 张骞,张霞,刘积仁,孙雨,文学志,刘铮.混合 P2P 环境下有效的查询扩展及其搜索算法.软件学报,2006,17(4):782−793. http://www.jos.org.cn/1000-9825/17/782.htm [doi: 10.1360/jos170782]

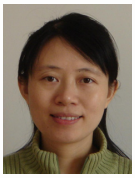
[17] 万常选,鲁远.基于权重查询词的 XML 结构查询扩展.软件学报,2008,19(10):2611−2619. http://www.jos.org.cn/1000-9825/19/2611.htm [doi: 10.3724/SP.J.1001.2008.02611]

[24] 张志强,孟庆海,谢晓芹.基于社会书签的个性化查询词扩展技术研究.计算机科学与探索,2010,4(9):812−829.

[34] 白露,郭嘉丰,曹雷,程学旗.基于查询意图的长尾查询推荐.计算机学报,2013,36(3):636−642.

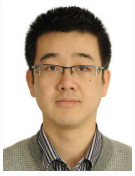
[42] 张志强,梁婷婷,谢晓芹.一种基于用户标记的搜索结果排序算法.计算机研究与发展,2009,46(增刊):351−358.

张志强(1973－),男,黑龙江哈尔滨人,博士,教授,CCF 高级会员,主要研究领域为信息检索,智能信息处理.
E-mail: zqzhang@hrbeu.edu.cn



谢晓芹(1973－),女,博士,副教授,CCF 高级会员,主要研究领域为社会网络分析与挖掘,服务计算,智能信息处理.
E-mail: xiexiaoqin@hrbeu.edu.cn

彭晴晴(1987－),女,硕士,主要研究领域为数据库,信息检索.
E-mail: pqq6name1987@126.com



冯晓宁(1976－),男,博士,副教授,CCF 会员,主要研究领域为数据库与知识库,分布式系统仿真,软件建模,无限传感器网络.
E-mail: fengxiaoning@hrbeu.edu.cn