Web查询日志研究综述

付 博[1],赵世奇[1,2],刘 挺[1]
(1.哈尔滨工业大学计算机学院社会计算与信息检索研究中心,黑龙江哈尔滨 150001;2.百度公司,北京 100085)

摘 要: 本文对查询日志在相关领域内的研究现状与进展进行了总结.首先介绍了web查询日志的常用信息和公开的数据集;进而阐述了查询日志在web搜索、信息抽取等方面的相关研究,并对它们进行了细致的介绍和分析;最后指出基于查询日志研究所面临的问题和挑战.重在对基于查询日志研究的主流方法和前沿进展进行概括、比较和分析,以期对后续研究有所助益.

关键词: 查询日志分析; 查询日志挖掘; web搜索; 信息抽取
中图分类号: TP391.2 **文献标识码:** A **文章编号:** 0372-2112 (2013) 09-1800-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.09.021

Research on Analysis and Mining of Web Query Logs

FU Bo[1], ZHAO Shi-qi[1,2], LIU Ting[1]
(1. *Center for Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*;
2. *Baidu Inc., Beijing* 100085, *China*)

Abstract: This paper surveys the state-of-the-art research on query logs analysis. First, the existing corpus of query logs and the information embedded in are summarized and analyzed. Then, important tasks benefiting from query logs are introduced, including web search, information extraction, as well as some closely related topics. Finally, the problems and challenges of current researches are discussed. This paper aims to make a summary, comparison and analysis of the mainstream methods and the latest progress, expecting to be helpful to the future research.

Key words: analysis on query logs; mining on query logs; web search; information extraction

1 引言

由于web技术的发展和普及,互联网逐渐成为了人们获取信息的主要来源,尤其是Google,Baidu等搜索引擎的巨大发展,使得搜索引擎成为最普遍的信息获取工具.一方面,人们通过网络获取信息和服务,如浏览新闻网站了解热点时事,或浏览电子商务网站购买商品;另一方面,人们通过网络共享和发布各种信息,如博客和论坛,所有这些都导致了互联网上网页量的迅速膨胀.因此,为了帮助用户方便地获取信息,迫切需要搜索引擎了解用户的意图或兴趣.而现有的搜索引擎记录了用户查询的关键词及其点击行为(即查询日志),这些日志表达了人们的各种检索意图和兴趣爱好,可以用于改善搜索引擎的性能.鉴于此,基于查询日志的研究成为了近年来信息检索等相关领域的热点问题.

用户查询(query)是指用户用自然语言向搜索引擎提交的词或词串.查询日志(query log)则是用户与搜索引擎交互的记录.表1是查询日志的一个实例,记录了用户查询时间、用户ID、查询、点击的URL链接、所点击链接的位置排序及所在页数等.用户查询方式通常是"查询—点击—浏览",这样一个循环过程,如果搜索引擎返回的结果令人不满意,用户可能修改其查询,以改善搜索结果,如此"查询-浏览结果-修改查询"的行为序列便构成了一个用户会话(session).最初的查询日志分析只是简单的查询词聚类.随着研究的深入,研究者们逐渐从简单的词聚类发展到更为复杂的点击链接与文档分析.按照处理的对象不同,基于查询日志的研究可分为查询与查询、查询与点击文档以及查询与点击链接等几个方面的研究.而按照处理任务的不同,可分为查询分类/聚类、查询扩展/推荐、查询改写以及个性化用户建模等.基于查询日志研究的优点在于:搜索引擎包含成千上万的用户查询及点击,这些丰富的用户信息为

收稿日期:2012-05-30;修回日期:2012-08-28;责任编辑:覃怀银
基金项目:国家自然科学基金 (No.60736044,No.61073126)

用户行为分析提供了保证;此外,查询日志反映了用户关心的查询词和网页内容,可以看作是用户对网页内容和质量的一种隐式评价,因而查询日志为网络信息的质量评估提供了可能.

表 1　查询日志示例

时间	用户 ID	查询	点击	排序	页数
23:59:58	6698	百度	www.baidu.com	1	1
17:32:26	82860	java	www.javaeye.com	4	1
20:19:19	78047	比亚迪 f3	www.bydauto.com.cn	3	2
03:35:28	38746	佟大为	www.zhishiku.com.cn	1	1
…		…	…	…	

查询日志在很多研究方向上都体现出重要的价值.如信息检索中的查询扩展[1~3]、查询推荐[4~6],自然语言处理中的复述[7]、命名实体识别[8~10]等.尽管对查询日志的研究工作开展已久,但没有研究者对该领域进行系统完整地综述.本文综合已有的研究成果,结合自身研究经验与系统分析,将涉及查询日志的研究任务进行归纳总结,如图 1 所示.

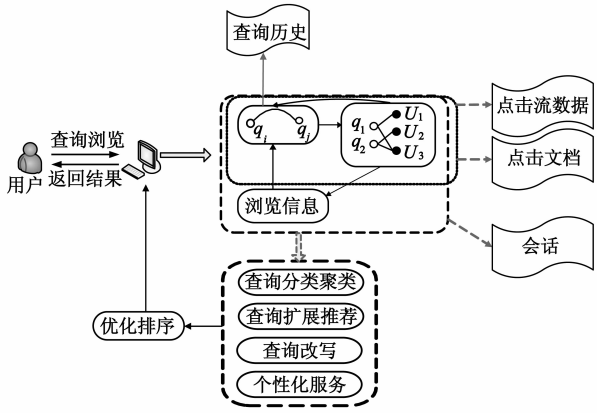


图1　基于查询日志的研究

接下来本文将分别阐述基于查询日志的几个主要任务方向.首先介绍查询日志语料及用户行为信息.然后介绍查询日志在网络搜索、信息抽取,以及其它一些领域的几种应用.最后简要概括本领域目前存在的问题和未来的研究方向.

2　查询日志语料

2.1　语料

为了便于展开基于查询日志相关工作的研究,本节主要介绍目前该项研究中用到的公共语料资源.由于查询日志涉及用户隐私等原因,除了搜索引擎公司(如微软、雅虎、谷歌等)的实验室,多数大学和研究机构的实验室难以获取用户真实的查询日志.已知的公开的英文查询日志有 AltaVista[11],Excite[12]及 AOL 搜索引擎(http://www.gregsadetsky.com/aol-data/)所提供的文档集作为英文查询实验语料.;而针对中文查询的研究一般是利用搜狗(Sogou)搜索引擎(http://www.sogou.com)提供的公开语料作为中文查询实验语料[13].主要的公开语料资源见表 2.)

2.2　查询日志中的用户行为信息

查询日志中通常包括以下用户行为信息:

(1)查询(query):是指用户用自然语言向搜索引擎提交的词或词串,也称查询串(query string).

(2)会话(session):是用户认知过程在查询日志中的记录,"查询-浏览结果-修改查询"的行为序列便构成了一个用户会话.

(3)词项(term):是指不含分隔符的连续字符序列.其中,分隔符包括逗号,句号,冒号,空格符等.

(4)点击 URL:用户点击网页的结果地址.

(5)结果页面查看:用户习惯点击的结果在搜索引擎返回结果中的大体位置.

3　查询日志在网络搜索中的应用

前人的很多研究致力于从查询日志中挖掘用户的搜索意图,探索提高搜索查全率和查准率的方法,而查询日志主要可应用于查询分类、查询扩展及查询推荐等,并可最终优化搜索性能.

3.1　基于查询日志的查询分类研究

查询是用户和搜索引擎交互的主要渠道,而信息扩张使搜索引擎为用户返回了大量的网页,增加了人们获取信息的复杂性.很自然地,人们首先想到对查询进行分类,进而根据不同的信息类型采取不同的处理方法.与传统的针对文档或网页的分类相比,查询分类处理的对象是用户查询,具有简短、内容信息不丰富的特点,从而挖掘用户真实意图就显得尤为重要.另外,查询日志中包含丰富的外部信息,如点击信息、文档信息等,这些也是传统的文本分类所不具备的.因而研究者从不同的角度提出了各种查询分类方法,代表性的包括基于意图的查询分类和基于主题的查询分类等.

查询意图分类,是指分析查询背后用户的真实信息需求类型.然而到目前为止,关于查询意图并没有一个标准的分类体系,几种典型的查询意图分类方法如表 2 所示.Broder[14]首先提出将查询意图分为导航类、信息类和事务类,随后的分类体系都受其分类标准的影响[15~20].目前查询意图分类方法可以概括为:基于查询内容、基于点击 URL 的方法以及基于外部信息(如锚文本等).在基于查询内容的方法中,主要使用各类查询的模板特征来区分查询,根据查询特点启发式地制定模板[17],但直接针对特定的语言现象制定模板的可扩展性差.为了解决这些问题,有学者[18,19]提出了基于点击 URL 分布的方法,通过计算每个查询的点击分布来区分查询.另外,Li[20]等人则通过构建查询与点击

之间的二分图(click-through bipartite graph),利用少量标注的种子查询类别,采用图的迭代算法来完成查询意图分类.在锚-链接分布(anchor link)的方法中,将锚文本(anchor text)作为用户查询,其链接的 URL 作为用户点击,计算每个锚链接的分布情况[17].此类方法是为了解决查询数据稀疏问题,但锚文本只是对少量真实用户查询的模拟而导致分类效果并不理想[18].以上方法巧妙地利用了查询日志中的用户行为信息,但此类方法通常针对导航类查询意图识别,而对于信息类查询意图区分效果不高.目前,查询意图分类是一个比较有价值的研究点,理解用户的查询意图可以为后续研究提供大量的准备工作,但其仍有许多值得深入研究的问题,一是缺乏权威数据和评价标准[21],这使得各方法之间难以进行相互比较.二是如何定义更为完备的查询意图分类体系,并在其之上进行特征抽取与意图分类是需要进一步研究的方向所在.

表 2 不同地域搜索引擎用户查询行为比较

数据集	Sogou			AOL	AltaVista	Excite
语言	中文			英文	英文	英文
时间	2006 年 8 月	2007 年 3 月	2008 年 6 月	2006 年 3~5 月	2002 年 12 月	2001 年 4 月
查询数	21,426,941	44,165,401	51,537,393	36,389,567	7,175,648	1M
平均查询长度(字)	6.4	7.1	6.8	—	—	—
平均查询长度(词)	3.1	3.5	3.3	2.35	2.35	2.21–2.6
平均会话长度	1.75	—	—	—	2.02	2.3–2.8

注:表中"—"表示没有相关结果

表 3 查询意图框架体系总结

作者	分类体系	特征	方法描述	日志
Broder[14]	导航类/信息类/事务类	–	人工分类:设计调查问卷,人工进行评估	Altavista
Rose[15]	导航类/信息类/资源类	-	人工分类	Yahoo
Baeza-Yates[16]	信息类/非信息类/模糊类		SVM/PLSA 按照查询意图聚类	TodoCL
Jansen[17]	导航类/信息类/事务类	查询内容	各类查询的启发式特征 根据查询词的特点制定模板,判断查询意图.	Dogpilo.com
Lee[18]	信息类/导航类	点击流/锚文本	根据点击分布和锚文本推测查询意图	Google
Liu[19]	导航类/信息类/事务类	点击/会话/nCS/nRS	C4.5 决策树:根据用户会话推测查询意图	Sogou

查询主题(topic)分类,主要指将查询映射到预定义的主题中[22],如"政治","经济","体育","财经"等.然而这里有两个问题有待解决,一是如何构建查询主题分类体系,二是如何丰富查询特征或是增加训练数据来训练分类器,来为查询日志中未出现过的查询进行分类.目前,对于查询主题分类也没有一个标准的分类体系,通常是采用人工构建或是自动挖掘的方法.在人工构建的方法中,很自然想到针对高频的查询或根据查询主题分布构建类别,然后再根据查询特征丰富查询类别[23],但由于查询具有长尾特性,导致此种方法定义类别覆盖率很低(仅占 16%).在自动挖掘方法中,通常直接利用网站目录作为分类体系(如 Yahoo 和 ODP (http://www.dmoz.org/)等).此外,与查询意图分类缺乏公共评测不同,KDD CUP 2005 首次提出了针对英文查询主题分类的评测[24].但由于语言的差异性,这种类别体系并不适合于中文查询的研究中.纵观目前的工作,查询主题分类可分为两种研究思路:基于查询内容的方法以及基于特征分类的方法.前者主要通过扩充查询,丰富查询的特征表示.后者主要是使用机器学习的方法,选取大量有意义的特征来完成分类任务.这两种研究思路有很多代表性的研究工作.文献[25~27]首先定义查询上下文可以是单个查询,或是同一会话中的查询,然后统计查询词 x(如"阿凡达")的上下词 y(如"下载","上映","导演"等),并进而确定其类别为 u(如"娱乐类"),若一个新的查询词 x'的上下文与上述上下文 y 很相似的话,则可判断 x'的类别也应该是 u.这种方法的重点一般都放在的相似查询的抽取和查询上下文判断的研究上.在基于特征分类的方法中,Beitzel 等人[26]首次将机器学习的方法应用于查询主题分类任务中.他们尝试使用 n-gram 词语特征,并对比了精确匹配、感知器和选择偏好(selection preference)这 3 种分类模型,发现组合后的分类器要好于每个单独的分类器.Cao[27]等人除了考察传统的 n-gram 之外,还引入了会话(即上下文)特征和点击特征.类似于查询意图分类任务,基于特征的方法的研究重点在于有效特征的发现以及特征选择和特征融合等问题的研究.

3.2 基于查询日志的查询扩展研究

查询扩展(query expansion)是搜索引擎在接收到用户查询之后,自动地使用与用户查询相关的词来对查询进行扩展,以构造内容更丰富且更易检索到结果的新查询[28].早期方法的扩展词来源于文档,主要分为全局分析法(global analysis)和局部分析法(local analysis)两

种方法[3].但传统的扩展方法通常会引入噪声,甚至出现“转义(drift)”现象,即查询扩展后的主题偏离了用户的真正意图.由于前人研究证明查询扩展技术可以有效改进检索性能,因而这类技术一直是信息检索领域的研究热点.

查询扩展技术的关键在于扩展词的来源以及如何估计扩展词权重的问题.这里主要讨论扩展词的来源问题,可分为两种思路:基于查询内容的扩展和基于点击流的扩展等方法.在基于查询内容的研究中,主要利用查询间的关联关系.如有人使用词聚类的方法获取扩展词[29~31],即首先计算各查询之间的相似度,然后在此基础上进行聚类.这里,查询相似度的计算可以基于词重叠率,其前提假设是词语之间是完全独立的.也可以基于词的共现性[32],即在语料库中经常共同出现的词语往往相关度很大.分析共现性时,可以采用词语粒度、短语粒度等.此类方法的优点在于查询日志中查询词的数量众多,因此使用该方法可以构建相当规模的候选扩展查询.这种方法的缺点在于文本聚类、词聚类等产生的错误级联,使得获取的扩展词往往含有大量噪声,准确率较低.为了避免词语的多义性会引入噪声,有学者利用点击信息获取候选查询词.因而在基于点击流的方法中,主要是利用点击信息获取候选查询词.此方法的主要思想是在用户查询与其点击文档之间构建图,利用图挖掘查询之间的相似性,进而获取候选扩展词[1,3,33],其最大的优点是可以方便地解决查询表达方式多样性的问题.基于点击图的方法又可细分为以下几类:(1)若两个查询倾向于点击相同的URL,则这两个查询的意思相近.相应地,此类方法通过计算两个查询所点击的URL的相似度作为查询本身的相似度,利用余弦相似度、编辑距离及Jaccard相似度算法计算查询相似度,如图2(a)所示.(2)若两个查询点击的网页内容相似,则这两个查询的意思便相似.此类方法使用查询所点击的文档来获取候选扩展词,需要首先抽取每个查询的点击文档的特征,以作为该查询的特征,如图2(b)所示.

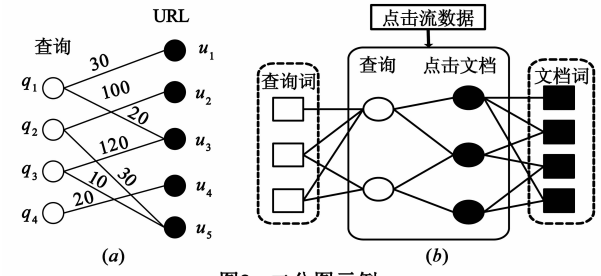


图2　二分图示例

查询日志作为一种辅助的资源,其优势在于把用户的历史记录作为一种隐式相关反馈.目前查询扩展技术仍有几方面值得深入思考,一是扩展词的个数确定是一个尚未解决的问题;二是如何进一步的研究语义查询扩展问题.

3.3　基于查询日志的查询推荐研究

查询推荐(query suggestion)旨在自动为用户推荐相关查询,通过推荐更合适的查询来提高搜索引擎的性能以及用户的搜索体验.针对大量的查询,查询推荐旨在统计与查询相关的高频词后,自动分析和归纳整理出与查询主题相关的词或短语,以节省用户查找相关文档的时间.查询推荐可以看作是比查询扩展更直接、更方便与用户交互的手段,因而查询推荐一直是搜索引擎研究中的一个热点问题.图3展示的便是百度搜索引擎的查询推荐效果.查询推荐并不局限于通用搜索引擎中,还广泛运用到广告推送、电子商务中的商品推荐等.

图3　百度查询推荐实例

基于查询日志的查询推荐方法与查询扩展方法类似,大多是将其定义为查询与候选查询之间的相似度计算问题.研究者一般会在多个维度上计算两者之间的相似度,如查询词相似度,查询点击相似度以及在同一会话中查询相似度[6,34,35],依此对候选查询进行排序,并为用户推荐前 k 个相关查询.然而仅对查询本身计算相似度很困难,为了计算查询的语义特征,两个查询是否相似应取决于查询所处的上下文.基于这一考虑,有人提出了基于会话的方法,该方法首先找到包含查询 q_1 的若干会话,然后在每个会话中挖掘与 q_1 相关的查询对(q_1,q_2),此时 q_2 即为用户所推荐的查询.在排序候选查询时,可以利用关联规则[34]、互信息[35]的方法,通过计算会话中所有查询间的关联度来挖掘相似查询.此方法有助于明确用户的查询意图,使查询推荐效果更加准确.但基于会话的方法需要对会话进行准确的划分.

还有一部分研究采用了基于图的方法来进行查询推荐.此方法是将查询相似度问题转化为图中节点相似度的问题,根据不同的理解和需要定义图中的节点

和边,并赋予不同的权值,继而采用各种基于图的迭代算法来完成查询推荐.如有学者考察查询与点击文档而构建点击图,继而将查询表示为所点击链接的向量形式[36],以此计算不同查询间的相似度.但此类方法也存在几方面的问题:(1)点击数据稀疏;(2)空间维数较高.为了解决点击数据稀疏的问题,Antonellis[37]利用SimRank算法将节点间的相似度转化为邻居节点间相似度;Craswell 和 Szummer[38]采用随机游走模型,依据节点间的连通性和转移概率传递节点间的关联关系.为了将向量空间维数降低,Baeza-Yates[4]利用 k-means 聚类算法将用户点击查询进行聚类;Cao 等人[5]在点击图中融合会话信息对查询进行聚类.此外,最近有学者采用查询流图(query-flow graph)[39,40]的方法来进行查询推荐,通过考察会话中的查询间的关系而构建图.基于图的方法是一种新颖的方法,它可以灵活的将词语间的各种联系作为特征融入图中,继而进行迭代计算.然而,寻找更有效的词语特征以及如何选取图算法是值得深入研究的问题.

4 查询日志在信息抽取中的应用

信息抽取(Information Extraction,IE)是指从文本中抽取出特定的事实信息,并以结构化的形式对其描述[41].对于搜索应用而言,应用信息抽取技术的主要用途是识别与挖掘命名实体和语义关系等.由于命名实体是查询信息的主要载体,并且包含了重要的查询信息,因而,有效识别命名实体及其语义关系将对提高搜索引擎排序效果有着很深刻的影响.与命名实体类似的研究包括新词发现,例如"给力","跑酷"等.由于新词发现的研究与命名实体工作的研究有重合之处,同时大部分的新词也是命名实体,因而以下主要介绍基于查询日志的命名实体识别,以及语义关系抽取方法.

4.1 基于查询日志的命名实体挖掘与识别

命名实体(Named Entity,NE)挖掘,主要是指从文本中挖掘出人名、地名、机构名等[42].研究 NE 对于自然语言处理的各项应用都很重要.早期 NE 的研究方法是基于人工规则的算法.近年来,一些机器学习的方法被应用到 NE 的研究之中.利用查询日志来挖掘或识别 NE 的工作开展时间较晚,然而有研究表明,用户查询涉及到 NE 的占总查询的 70%左右[8],从而为研究 NE 提供了更大的资源.由于查询具有简短、语义模糊、语法结构残缺的特点,使得直接利用文本挖掘或识别技术处理查询中的 NE 可行性不高,因此迫切需要对查询日志中的 NE 研究提出新方法和新策略.

目前从查询日志中挖掘 NE 的方法,主要包括以下几个步骤:(1)从查询日志中选取一组 NE 作为种子,并标注其类别(如:"诺基亚 5230"属于"数码类");(2)基于种子实例从查询日志中抽取每一类种子的上下文信息组成这一类别的特征向量,以训练模型;(3)利用上述模型挖掘每一类别的新 NE.一部分学者对步骤(2)采用基于分布假设的方法,其基本思想是那些倾向于出现在相似的上下文中的词意思相近,基于此计算词义相似度并实现同义词自动聚类[8].也有学者基于主题模型(topic model)展开对步骤(2)的研究.具体地,Guo 等人[9]针对每个种子 NE 获取相应描述文档,然后利用种子 NE 来学习一个弱指导主题模型(WS-LDA).具体来说,对每个种子 NE,他们通过遍历查询日志获取所有包含此 NE 的用户查询,将查询中除 NE 以外的其他词作为查询模板.如表 4 中所示.此类方法仅限于当前查询可以从查询日志中抽取到实体和上下文,否则无法对其进行识别.针对这种情况,Du 等人[10]同时利用会话(在该文章只考虑了相邻查询)和点击行为,采用了类别特征和词重叠特征识别 NE,说明查询会话和点击的方法要优于仅利用查询文本的方法.

基于查询日志的 NE 识别与挖掘研究作为信息抽取、自然语言处理等任务的基础工作,近年来广受关注.此项工作仍有许多值得深入研究的问题,一是目前的研究工作主要是面向限定领域的,需要进一步深入研究开放领域的、细分类的 NE 识别方法.二是目前与 NE 相关的研究工作多是针对英文查询,中文查询的相关研究刚刚起步,如何结合中文处理的特点,将一些成熟的技术和资源应用到中文 NE 识别领域,是一个值得我们积极探索的任务.

表 4 基于查询模板的 NE 识别方法

NE Sample	Template		type
	Context	Context,click-through	
Harry Potter	# book	# book,amazon.com	Book
	# walkthrough	# walkthrough,cheat.com	Movie
Titanic	# DVD	# DVD,amazon.com	Game
	# trailer	# trailer,appel.com	Movie

4.2 基于查询日志的语义关系抽取

由于目前的信息检索技术多是采用词串匹配的方法,只有查询词出现在文档中,才有可能被检索到.人们在现实生活中描述同样的对象或事件时用词存在着多样性,如:"电脑"和"笔记本"都属于计算机这一概念范畴.因此许多学者着手研究从查询日志中抽取语义关系[43,44]并构建 Ontology,研究查询间的语义关系可细分为两类,一类是基于概念(concept)的方法 (以下简记为方法 1),另一类则基于分布假设的方法(记为方法 2).

方法 1 主要利用查询日志挖掘新概念或是概念间所隐含的关系.Fonseca[2]利用关联规则的方法挖掘相关查询,即利用包含了相同查询词(如:"jaguar")的会话构

建关系图,找到有强关系的查询词构成查询子集,将其称为概念(如包含上例的子集为:概念 1:lion,tiger;概念 2:cars;概念 3:atari 等),用概念来描述查询主旨,找到与查询语义相关的概念对查询进行扩展(如:jaguar 扩展为 jaguar AND [lion OR tiger]).此外,Li 等人[45]着重分析用户查询中的名词短语,将查询结构划分为中心词和修饰词,如:"阿凡达-导演"中"导演"为中心词,"阿凡达"为修饰语.方法 1 主要用于语义查询扩展[2]、语义查询推荐[46]等研究中.

与方法 1 不同,方法 2 需要输入种子和定义类别体系,再根据每个查询的上下文构建模板.如 Sekinei[47]利用少量查询词作为种子,将其在查询日志中的上下文(即种子左、右各 n 个词)作为模板,进而使用这些模板抽取更多同类型查询词.比如,对于查询词"阿凡达",查询中经常与之共现的词有"电影","导演","下载","影评"等.这些高频的共现词即可视为模板,用以学习更多的电影类查询词.Pasca[48]利用网络文档制定模板(如:A such as B,认为 A 是实例类别,B 为类别属性),进而使用这些模板从查询日志中抽取实例类别与类别属性对.方法 2 可以扩大每个类别下的实例个数,但在计算过程中需要存储大量模板的上下文信息,因此计算时间和空间复杂度较高.

5　查询日志在其它方面的应用

除上面介绍的在搜索引擎的应用之外,查询日志还可以应用到其它研究中.

5.1　查询日志应用于个性化搜索

基于查询日志的个性化搜索,是指通过追踪和分析用户的历史检索记录,挖掘出用户的个性化信息,并依此预测用户偏好,满足不同背景、不同目的及不同时期的查询需求.实现个性化搜索的关键在于要准确描述用户的兴趣和行为[49],并且跟踪其变化[50].目前,个性化检索的研究比较复杂,可以利用多种资源帮助用户进行个性化检索[51].有些用户行为(如浏览、下载和评价等)不包含在查询日志中,因而个性化搜索的研究中,查询日志仅是一种辅助资源[52].尽管查询日志的信息不够全面,但还是可以从中发现许多有意义的信息,比如通过用户行为特征推断是否为个性化查询;通过创建或更新用户描述文件(user profile),以获得页面的点击次数、页面停留时间和页面访问顺序等信息,可以分析出用户在每个资源上所花费的时间,从而可以为用户推荐合适的信息[53].但从另一方面讲,个性化搜索想要发挥作用,还需要提出有效保护用户隐私的机制[54,55].

5.2　查询日志应用于在线广告

在线广告的研究中,查询日志的作用体现在两个方面.一方面,查询日志可用于竞价排名广告(sponsored search)中.竞价排名是搜索引擎关键词广告的一种形式,根据企业购买关键词的价格对网站进行排名,通常以赞助商链接的推广形式出现在搜索结果页的顶端或侧边栏(如:Baidu,Google 等),企业用户可自由投放广告的关键词,同时按照用户点击次数收取相应费用.另一方面,查询日志还可用于在线商业意图(online commercial intent)识别.利用查询词[56]、查询点击广告页面等用户行为[57~59]挖掘用户对某类产品的兴趣,并以此提高广告投放的有效性,改进广告检索效果,还可以面向用户推送相关广告以获取更多利润.

5.3　查询日志应用于舆情分析

互联网的开放性和虚拟性,决定了网络舆情具有突发性和随意性的特点.越来越多的互联网用户通过网络这一载体表达观点,因而,网络舆情对社会稳定的影响与日俱增,如果网络舆情事件处理不当,有可能诱发民众的不良情绪,进而对社会稳定形成严重威胁.而查询日志以时间顺序记录着用户大量的查询,对一段时间内用户检索的关键词进行统计,根据被搜索次数来说明词的热门程度,从海量信息中找到热点、敏感话题,并对其趋势变化进行追踪[60~62].因而可以依靠监测查询词等方式自动地对舆情进行监控.虽然目前这一应用点的研究成果还不是很多,但不影响其成为一个有价值的应用点.

除了上面介绍的 3 个主要应用领域以外,查询日志在其它领域也扮演着重要的角色.例如,查询会受到时间、突发事件等因素影响,将查询日志作为事件检测的资源,挖掘出相关事件[63].此外,查询日志中包含着丰富的同义词信息,进而与复述相结合,能更好的抽取具有歧义的复述短语资源[7].广泛应用查询日志的研究在很大程度上源于人们改进检索性能的愿望.查询日志在以上众多研究领域的应用使其成为了一个非常重要的资源.

6　结束语

本文在充分调研和深入研究的基础上对查询日志的应用研究进展进行了综述.其中重点介绍了基于查询日志的几个主要研究方向,包括查询日志在网络搜索中的应用、在信息抽取中的应用及个性化搜索研究中的应用等.近年来,基于查询日志的研究在相关领域内广泛使用,其中也融合了包括自然语言处理、网络挖掘和机器学习等各项技术.然而,由于查询日志自身的特殊性和复杂性,对其的应用研究尚有许多值得深入探索的问题.在本文的最后,我们提出一些值得进一步研究和探讨的方面.

问题 1:虽然人们已经提出了多种方法用于大规模

数据的处理.然而总的来看,处理的方法计算成本高,时效性低,可扩展性不好.因而,如何结合查询日志的特点(如用户查询的重复率高,具有长尾特性等),提出更加成熟的算法应用到查询日志这种大规模动态的数据中,是一个亟待解决的问题.另外,针对这种大规模的查询日志,大部分工作都集中在用户行为数据的分析任务中,对数据的模型化还不够.因此需要提出一个好的模型,准确的表达用户行为(如:用户搜索、点击、浏览等),体现不同用户其背景、特点和需求不同,也非常值得深入且重点加入研究.

问题 2:相对于大规模查询日志的处理,资源获取方面也存在一定的困难,其中最重要的原因是涉及到用户隐私问题.为了有效保护用户隐私的机制,这方面的研究工作还处于探索阶段.因此,接下来的一个主要目标就是如何找到一种有效的方法,对用户查询日志的隐私信息进行屏蔽,使得搜索引擎公司可以释放查询日志数据,特别是供个性化搜索研究使用.

问题 3:目前来看,搜索引擎呈现检索结果的方式主要是返回与查询词匹配的链接排序.然而更有意义的呈现方式是面向查询直接给出简洁有效的准确答案,这种基于任务驱动的研究也引起了国外一些学者的关注.但总的来看,由于研究得比较粗糙,准确率并不是很高,无法达到实用.因此在未来的工作中,需要我们进一步将工作细化,如确定检索结果的形式,以及如何构建庞大的知识库等都是值得深入研究的问题.

问题 4:如前所述,搜索引擎所面临的问题是深入理解用户查询,因而查询意图分析在众多相关研究中都具有很重要的作用.虽然人们已经投入了很大的精力,但总的来看,目前用户意图分析结果的准确率并不是很高.其中一个关键原因是:查询表达多样性常常给基于关键词的检索带来许多语义理解错误.因此,在今后的工作中,需要我们进一步将工作细化,如深入挖掘用户查询中包含的语义信息,研究用户意图与 web 文档作者意图匹配等,都是值得我们积极探索的任务.

致谢 在此我们向对本文的研究工作提供帮助的老师和同学表示感谢.

参考文献

[1] Cui H, Wen JR, Nie JY, Ma WY. Query expansion by mining user logs[J]. IEEE Transaction on Knowledge and Data Engineering, 2003, 15(4): 829 – 839.

[2] Fonseca BM. Concept-based interactive query expansion[A]. Proceeding of the CIKM[C]. New York: ACM Press, 2005. 696 – 703.

[3] 崔航,文继荣,李敏强.基于用户日志的查询扩展统计模型[J].软件学报,2003,14(9):1593 – 1599.
Cui Hang, Wen Ji-rong, Li Min-qiang. A statistical query expansion model based on query logs[J]. Journal of Software, 2003, 14(9): 1593 – 1599. (in Chinese).

[4] Baeza-Yates, Hurtado C, Mendoza M. Query recommendation using query logs in search engines[A]. Proceeding of the EDBT Workshop[C]. Berlin, Heidelberg: Springer-Verlag, 2004. 588 – 596.

[5] Cao HH, Jiang DX, Pei J, He Q, Liao Z, Chen EH, Li H. Context-aware query suggestion by mining click-through and session data[A]. Proceeding of the KDD[C]. New York: ACM Press, 2008. 875 – 883.

[6] Huang C, et al. Relevant term suggestion in interactive web search based on contextual information in query session logs[J]. Journal of the American Society for Information Science and Technology, 2003. 54(7): 638 – 649.

[7] Zhao SQ, Wang HF, Liu T. Paraphrasing with search engine query logs[A]. Proceeding of the COLING[C]. Morristown: ACL, 2010. 1317 – 1325.

[8] Pasca M. Weakly-supervised discovery of named entities using web search queries[A]. Proceeding of the CIKM[C]. New York: ACM Press, 2007. 683 – 690.

[9] Guo JF, Xu G, Cheng XQ, Li H. 2009. Named entity recognition in query[A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2009. 267 – 274.

[10] Du J, Zhang ZM, Yan J, Cui Y, Cheng Z. Using search session context for named entity recognition in query[A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2010. 765 – 766.

[11] Jansen BJ, Spink A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs[J]. Information Processing and Management, 2006, 42(1): 248 – 263.

[12] Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web[J]. Information Processing and Management, 2000, 36(2): 207 – 227.

[13] Wu DY, Zhang Y, Liu T. Analysis of named entity queries in web search logs[J]. Journal of Computational Information Systems, 2011, 7(16): 5837 – 5844.

[14] Broder A. A taxonomy of web search[J]. SIGIR Forum, 2002, 36(2): 3 – 10.

[15] Rose DE, Levinson D. Understanding user goals in Web search[A]. Proceeding of the WWW[C]. New York: ACM Press, 2004. 13 – 19.

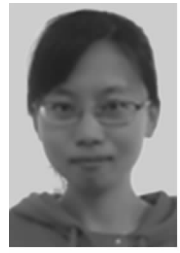
[16] Baeza-Yates R, Calderron-Benavides L, Gonzalez-Caro C. The intention behind web queries[J]. Lecture Notes in Computer Science 4209, 2006. 98 – 109.

[17] Jansen BJ, Booth DL, Spink A. Determining the user intent of web search engine queries[A]. Proceeding of the WWW[C].

New York:ACM Press,2007.1149 – 1150.
[18] Lee U,Liu Z,Cho J. Automatic identification of user goals in web search[A]. Proceeding of the WWW[C]. Japan: ACM Press,2005.391 – 400.
[19] Liu YQ,Zhang M,Ru LY,Ma SP. Automatic query type identification based on click through information[A]. Proceeding of the AIRS[C]. Berlin, Heidelberg: Springer-Verlag, 2006.593 – 600.
[20] Li X,Wang YY,Acero A. Learning query intent from regularized click graphs[A]. Proceeding of the SIGIR[C]. New York:ACM Press,2008.339 – 346.
[21] 张森,王斌. Web 检索查询意图分类技术综述[J]. 中文信息学报,2008,22(4):75 – 82.
Zhang Sen,Wang Bin. A survey of web search query intention classification[J]. Journal of Chinese Information Processing, 2008,22(4):75 – 82(in Chinese).
[22] Brenes DJ,Gayo-Avello D,Perez-Gonzalez K. Survey and evaluation of query intent detection methods[A]. Proceedings of the 2009 Workshop on Web Search Click Data[C]. New York:ACM Press,2009.1 – 7.
[23] Beitzel SM,et al. Hourly analysis of a very large topically categorized web query log[A]. Proceeding of the SIGIR[C]. New York:ACM Press,2004.321 – 328.
[24] Li Y, Zheng ZJ, Dai HH. KDD CUP-2005 report: Facing a great challenge[J]. SIGKDD Explorations Newsletter,2005,7(2):91 – 99.
[25] Hu J,Wang G,Lochovsky F,Sun JT,Chen Z. Understanding user's query intent with Wikipedia[A]. Proceeding of the WWW[C]. New York:ACM Press,2009.471 – 480.
[26] Beitzel SM,et al. Automatic classification of web queries using very large unlabeled query log[J]. ACM Transactions on Information Systems,2007,25(2):1 – 29.
[27] Cao HH. Context-aware query classification[A]. Proceeding of the SIGIR[C]. New York:ACM Press,2009.3 – 10.
[28] Carpineto C,Romano G. A survey of automatic query expansion in information retrieval[J]. ACM Computing Surveys (CSUR),2012,44(1):1 – 56.
[29] Beeferman D,Berger A. Agglomerative clustering of a search engine query log[A]. Proceeding of the KDD[C]. New York:ACM Press,2000.407 – 416.
[30] Chan WS,Leung WT,Lee DL. Clustering search engine query log containing noisy clickthroughs[A]. Proceeding of SAINT[C]. Tokyo:IEEE Computer Press,2004.305 – 308.
[31] Wen JR,Nie JY,Zhang HJ. Clustering user queries of a search engine[A]. Proceeding of the WWW[C]. New York: ACM Press,2001.162 – 168.
[32] Shen X, Tan B, Zhai CX. Context-sensitive information retrieval using implicit feedback[A]. Proceeding of the SIGIR[C]. New York:ACM Press,2005.43 – 50.
[33] Cui H,Wen JR,Nie JY,Ma WY. Probabilistic query expansion using query logs[A]. Proceeding of the WWW[C]. New York:ACM Press,2002.325 – 332.
[34] Fonseca BM,Golgher PB,Moura ES,Ziviani N. Using association rules to discover search engines related queries[A]. Proceeding of the LAWEB[C]. Santiago: Citeseer, 2003.66 – 71.
[35] Jones RR,et al. Generating query substitutions[A]. Proceeding of the WWW[C]. New York:ACM Press,2006.387 – 396.
[36] 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.
Li Ya-nan,Wang Bin,Li Jin-tao. A survey of query in search engine[J]. Journal of Chinese Information Processing, 2010, 24(6):75 – 84(in Chinese).
[37] Antonellis I, Molina HG, Chang CC. Simrank + +: query rewriting through link analysis of the click graph[A]. Proceeding of the VLDB[C]. Auckland, New Zealand,2008.408 – 421.
[38] Craswell N, Szummer M. Random walks on the click graph [A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2007.239 – 246.
[39] Boldi P,Bonchi F,Castillo C,Donato D,Vigna S. Query suggestions using query-flow graphs[A]. Proceeding of the WSCD Workshop[C]. New York:ACM Press,2009.56 – 63.
[40] Bordino I,Castillo C,Donato D,Gionis A. Query similarity by projecting the query-flow graph[A]. Proceeding of the SIGIR [C]. New York:ACM Press,2011.
[41] Banko M,Cafarella M,Soderland S. Open information extraction from the web[J]. Communications of the ACM,2008,51(12):68 – 74.
[42] Nadeau D,Sekine S. A survey of named entity recognition and classification[J]. Linguisticae Investigationes,2007,30(1):3 – 26.
[43] 蔡怡峰,彭鑫,钱乐秋. 面向语义构件检索的交互式查询方案生成[J]. 电子学报,2008,36(8):1631 – 1636.
Cai Yi-feng,Peng Xin,Qian Le-qiu. An Interactive query generation method for semantics-based component retrieval[J]. Acta Electronica Sinica,2008,36(8):1631 – 1636(in Chinese).
[44] 乔亚男,齐勇. 查询语义图辅助的信息检索性能预测模型[J]. 电子学报,2011,39(A03):158 – 162.
Qiao Ya-nan, Qi Yong. Predicting query performance using smantic chart[J]. Acta Electronica Sinica,2011,39(A03):158 – 162 (in Chinese).
[45] Li X. Understanding the semantic structure of noun phrase queries[A]. Proceeding of the ACL[C]. Morristown: Association for Computational Linguistics,2010.1337 – 1345.
[46] Meij E, Bron M, Hollink L, Huurnink B, Rijke M. Learning semantic query suggestions[A]. Proceeding of the ISWC[C].

New York: Springer-Verlag, 2009. 424 – 440.
[47] Sekinei S, Suzuki H. Acquiring ontological knowledge from query logs[A]. Proceeding of the WWW[C]. New York: ACM Press, 2007. 1223 – 1224.
[48] Pasca M. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs [A]. Proceeding of the ACL[C]. Morristown: Association for Computational Linguistics, 2008. 19 – 27.
[49] Weding S, Madani O. A large-scale analysis of query logs for assessing personalization opportunities[A]. Proceeding of the KDD[C]. New York: ACM Press, 2006. 742 – 747.
[50] 韩立新,陈贵海,谢立. 一种面向 Internet 的个性化信息检索系统模型[J]. 电子学报, 2002, 30(2): 240 – 244.
Han Li-xin, Chen Gui-hai, Xie L. A model of personalized information retrieval systems for internet applications[J]. Acta Electronica Sinica, 2002, 30(2): 240 – 244(in Chinese).
[51] Liu F, Yu C, Meng W. Personalized web search for improving retrieval effectiveness[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(1): 28 – 40.
[52] 曾春,刑春晓,周立柱. 个性服务技术综述[J]. 软件学报, 2002, 13(10): 1952 – 1961.
Zeng Chun, Xing Xiao-chun, Zhou Li-zhu. Survey of personalization technology[J]. Journal of Software, 2002, 13(10): 1952 – 1961. (in Chinese)
[53] 谢海涛,孟祥武. 适应用户需求进化的个性化信息服务模型[J]. 电子学报, 2011, 39(3): 643 – 648.
Xie Hai-tao, Meng Xiang-wu. A personalized information service model adapting to user requirement evolution[J]. Acta Electronica Sinica, 2011, 39(3): 643 – 648(in Chinese).
[54] Weerkamp W, Berendsen R, Kovachev B, Meij E, Balog K, Rijke M. People searching for people: Analysis of a people search engine log[A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2011. 45 – 54.
[55] Feild HA, Allan J, Glatt J. CrowdLogging: Distributed, private, and anonymous search logging[A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2011. 375 – 384.
[56] Ashkan A, Clarke C. Term-based commercial intent analysis [A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2009. 800 – 801.
[57] Dai HH, Zhao LZ, Nie ZQ, Wen JR, Wang L, Li Y. Detecting online commercial intention[A]. Proceeding of the WWW [C]. New York: ACM Press, 2006. 829 – 837.
[58] Singh G, Parikh N, Sundaresn N. User behavior in zero-recall ecommerce queries[A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2011. 75 – 84.
[59] Zhang W, Yan J, Yan SC, Liu N, Chen Z. Temporal query substitution for ad search[A]. Proceeding of the SIGIR[C]. New York: ACM Press, 2009. 798 – 799.
[60] Cartright MA, White RW, Horvitz E. Intentions and attention in exploratory health search[A]. Proceeding of the SIGIR [C]. New York: ACM Press, 2011. 65 – 74.
[61] Brilliant L. Detecting influenza epidemics using search engine query data[J]. Nature. 2009, 457(7232): 1 – 5.
[62] Zhang Y, Sun M, Zhang Y. Chinese new word detection from query logs[A]. Proceeding of the ADMA[C]. Berlin, Heidelberg: Springer-Verlag, 2010. 233 – 243.
[63] Zhao QK, Liu TY, Bhowmick SS, Ma WY. Event detection from evolution of click-through data[A]. Proceeding of the KDD[C]. New York: ACM Press, 2006. 484 – 493.

作者简介



付 博 女,1983 年 10 月出生于黑龙江海伦. 哈尔滨工业大学计算机科学与技术学院博士研究生. 主要研究方向为信息检索和社会计算.
E-mail: bfu@ir.hit.edu.cn

赵世奇 男,1981 年 6 月出生于辽宁抚顺. 博士,CCF 学生会员,主要研究方向为自然语言处理和知识挖掘.
E-mail: zhaosq@ir.hit.edu.cn

刘 挺 男,1972 年 2 月出生于黑龙江哈尔滨. 现为哈尔滨工业大学计算机科学与技术学院教授、博士生导师. 主要研究方向为自然语言处理、信息检索和社会计算.
E-mail: tliu@ir.hit.edu.cn