DOI:CNKI:11-2746/G2.20121126.2007.001

查询意图研究综述*

陆 伟 周红霞 张晓娟

摘 要 近年来,学界对查询意图进行了广泛探讨,一些重要国际会议如SIGIR,WWW等越来越重视查询意图的相关研究,其研究点主要集中在给定分类体系下的查询意图识别,内容涉及查询意图类目体系构建、特征识别、分类方法以及数据集与评价方法。研究发现当前查询意图研究面临如下问题和挑战:缺乏权威的评测标准,各种分类方法在大规模查询集合上的性能还不确定,有效提取或者获得查询特征的方法仍然值得深入研究,查询意图分类体系的完备性和类别间独立不相关性尚不确定。表2。参考文献71。

关键词 查询意图 查询分类 查询特征

分类号 G353.4

Review of Research on Query Intent

Lu Wei, Zhou Hongxia & Zhang Xiaojuan

ABSTRACT Query intent has recently been deeply studied in academics, in particular when there is a given taxonomy, so more and more important international conferences such as SIGIR and WWW pay much attention to it. The focus of this research is the query intent recognition given the category system, which includes category building, feature identification, classification methods as well as datasets and evaluation methodology. The problems and challenges existing in query intent research are as follows: lack of authoritative evaluation indexes, the uncertainty of classifying performance on mass query sets, how to more effectively extract or obtain query characteristics, the uncertainty of taxonomic completeness and the uncertainty of independent relationship between the classes. 2 tabs. 71 refs.

KEY WORDS Query intent. Query classification. Query feature.

作为网络信息查找的必备工具,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其信息需求,使得返回的结果有时不尽如人意。根据用户输入的查询信息,自动识别出其查询意图(即查询中包含的用户信息需求、查询目标、查询动机等),返回与其信息需求更相关的信息,成为有效把握用户需求、提高搜索引擎检索质量的途径之一。

近年来,学界对查询意图进行了广泛探讨,一些重要国际会议如SIGIR,WWW等越来越重视查询意图的相关研究。从目前来看,查询意图研究主要集中在给定分类体系下的查询意图研究,包括查询意图类目体系构建、查询意图特征识别、查询意图分类方法、数据集与评价方法四个方面,本文以此为基础分别加以介绍和评述。

* 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。

通讯作者:陆伟,Email:reedwhu@gmail.com

1 查询意图类目体系构建

一直以来,学界认为传统信息检索的核心宗旨为:用户内在的信息需求促使其采用相应的信息检索系统并产生相应检索行为,从而将用户查询中所包含的信息需求狭义地界定为信息类信息,即主题类查询[1]。自2002年开始,这一观点受到了质疑,Broder等[2]认为用户执行检索不只是想获取信息类信息,并通过对用户查询及AltaVista日志进行分析将用户查询意图分为三类,即信息类(I)、导航类(N)和事务类(T)。信息类指用户以一种静态方式去查询被认为能在网络上获取的信息,除阅读之外无其他交互信息,查找内容可以是数据、文档、文本或多媒体,信息需求既可以是精确的又可以是模糊的;导航类指用户查找某个特定网站(网页),该网站(网页)可以是个人网站(网页)也可以是组织网站(网页),即用户在执行检索时已在头脑中形成了查找意向,知道或者认为存在某网址可以满足自己的信息需求;事务类指用户通过查找获取一些资源或网络服务,比如购买、下载等。

在上述研究的基础上,Rose等[3]认为Broder的"事务类"不足以概括网上的所有资源,提出以"资源类"将其取代,指出"资源类"不再局限于一般的Web活动,而是包括网页上可获取的任何资源(而非信息类),并在此基础上提出了更细致的层次结构(见表1)。

表1 查询意图层级类目情况

层级一	层级二	层级三
导航类	无	无
信息类	有指导性的(Directed):用户想知道关于某个话题的特定信息	确定的(Closed):用户想找到有关某个问题的唯一没有歧义的答案
		开放的(Open):用户想了解某个开放式或者深度上不受限制的问题
	无指导性的(Undirected):用户想了解关于某个话题的任何信息	无
	建议(Advice):用户想得到一些建议、指南或其他方面的指导	
	位置(Locate):用户想查找某产品或服务的具体地理位置	
	列表(List):用户想得到一组可信的站点或者页面的列表	
资源类	获取(Obtain):用户想得到一个不是必须要通过电脑才能使用的资源	在线的(Online):用户想在线获取资源
		脱机的(Off-line):用户执行额外操作脱机获取资源
	下载(Download):用户想把某个资源下载到本地或者其他的设备上	免费的(Free):所下载资源无需付费
		付费的(Not free):所下载资源可能需要付费
	娱乐(Entertainment):用户想查看网页上的娱乐信息	无
	交互(Interact):用户想用网站上的另一个程序或者服务与资源进行交互	
	结果页(Result page):用户从搜索引擎结果页中打印、保存或阅读资源	链接(Links):用户查找的资源出现在搜索引擎结果页的标题、摘要或链接中
		其他(Other):用户查找的资源出现在搜索引擎结果页的其他地方

以上分类体系都是基于学者的经验知识构建的,并非所有学者都同意上述观点,如 Marchionini 等[4]将导航类和事务类归为查找搜索类(lookup search);Kang 和 Kim[1]将查询分为话题查询、主页查询和服务查询;Lee 等[5]将事务类和信息类合并为一个类别,仅包括研究信息类和导航类;Mendoza 等[6]将查询分为信息类、非信息类和歧义类三类,其中,上文所述的导航类和事务类归为非信息类,歧义类是指查询既可以是信息类又可以是非信息类。Waller 等[7]认为搜索引擎除了是获取信息的接口和到达某网站的通道,也是休闲的场所,故将查询意图分类体系扩展为:信息类、导航类、事务类和休闲类,但目前缺乏相关的实证研究。另一些国内学者[8-9]指出上述类目体系存在不合理之处,比如建议子类既可能含有查询意图(例如:如何正确操作使用 X 光机)也可以是简单的叙述(例如:我建议大家一起去游泳,如何),并将查询意图划分为:信息寻找意图、询问意图、下载意图、导航/URL 意图、比较意图和建议意图。除在文本检索中研究用户的查询意图类目外,另一些学者也尝试探讨非文本检索中的查询意图类目体系。如 Lux 等[10]通过研究发现,图像检索很少包含意图结构,并且 Broder 和 Rose 提出的分类体系不适合对图片检索的查询意图进行分类。基于此,Klofer[11]提出了图像检索意图类目体系,主要包含以下四大类:面向知识类(knowledge orientation)、导航类(navigation)、事务类(transaction)和意识图像类(mental image)。Ashkan 等[12]借助赞助搜索,将用户意图分为商业意图和非商业意图;Guo 等[13]又将商业类查询分为商品了解(Research)和商品购买(Purchase)两类。

以上对查询意图的探讨都是围绕用户查询目标展开,另外一些学者尝试从其他维度定位用户查询意图。最初,有学者尝试基于主题构建查询类别,并将图书馆学分类体系应用到查询分类中,但是研究结果表明这种方法适用性不大[14]。经过各阶段的探索,最后建议采用开放式目录(ODP)分类体系作为主题标签[15]。还有一些学者不借助外部类目体系,直接根据使用的数据集生成相应的主题类别[16];另外 Li 等[17]通过产品和工作两个维度来理解用户查询意图;Nguyen 等[18]总结出可以从模糊性、权威敏感度、时间敏感度和空间敏感度四个维度来识别查询意图。在此基础之上,Gonzalez 等[19]认为应从以下维度理解用户的查询意图:信息题材(Genre)、主题(Topic)、任务(Task)、目标(Objective)、专指度(Specificity)、范围(Scope)、权威敏感性(Authority Sensitivity)、空间敏感性(Spatial Sensitivity)、时间敏感性(Time Sensitivity)。在这些维度中,学者对时间和空间属性探讨较多,如 Kanhabua 等[20]将查询时间属性分为时间不敏感型和时间敏感型两类:时间不敏感型是指用户在查询中明确给出时间限定,查询结果不随执行查询时间的变化而变化,如“2008 北京奥运会”;时间敏感型是指用户在检索表达式中没有给出时间限定词,查询结果会因为执行查询的时间不同而不同,也可以认为此查询具有潜在时间意图。Jones 等[21]将潜在时间意图查询分为需求最新型、歧义型和非歧义型三类:最新型是指查找最新信息,非歧义型是指有唯一时间限定,歧义型则包含多个潜在的时间属性。空间敏感型查询是指查询应该考虑用户的地理位置,针对不同地理位置的用户返回不同的查询结果,即 Gravano 等[22]提出的局域(Local)查询,而空间不敏感型则无需考虑用户所在具体位置,不论用户在什么地方执行查询均返回同样的结果,即全局(Global)查询。Ding 等[23]学者进一步将局域查询细分为三个级别:国家级、州(省)级和城市级。在文献[24]中,Jones 等学者对查询的地理属性做了更细致的划分。

上文介绍了多位学者在查询意图类目划分方面所做的研究。虽然每个划分都有其依据和支持,但总体而言,Broder 和 Rose 等人的类目体系最受推崇,所以本文的评述也多将围绕这三个类目展开,并力图兼顾其他方面。

2 查询意图的特征识别

当查询意图类目体系确定后,如何选取分类特征对其进行分类是当前研究的重点。Spink 等[25]

通过对 Excite 搜索引擎日志进行分析,发现查询中包含的词汇数量平均在 2.4 个左右。因此,查询意图特征识别研究需要解决如何从简单的查询中获取充分和足够的特征问题,以便用这些特征来代表查询。张森等[26]将特征获取方法分为两类。第一类为事先方法,这种方法在查询被提交给搜索引擎以前,利用查询本身的特征来表示查询,比如表示特定需求的特征词汇、词与词之间的关系、词性以及词的选择优先性(Selectional Preference)、在语料库中的统计信息等;第二类为事后方法,这种方法利用查询被提交给搜索引擎以后的相关数据来获取查询特征,比如搜索引擎查询日志里相关查询的统计信息、搜索引擎针对该查询返回的检索结果等。本文在文献[26]的基础上,进一步梳理了查询意图分类特征的研究成果,并将其分为如下三类。

2.1 基于查询表达式的特征选取

一般而言,查询词是用户经过思考后提交的,是用户查询目标的最好表达,对查询表达式进行分析有助于识别用户的查询意图,Rose 等[3]甚至认为仅仅借助查询词本身就能识别查询中的潜在意图。Bernard 等[27]使用各类(导航、事务、信息)查询的一组启发式特征来区分查询,总结出含有公司、业务、组织、人名等顶级域名的查询为导航类,含有"obtain"、"download""entertainment"等术语的查询为事务类,含有"way to"、"how to"等词汇的查询为信息类查询。Belkin 等[28]通过统计分析得出词长为 2 以下的查询很可能是导航类查询,词长大于 2 的查询属于信息类查询的概率较大。Nguyen 等[18]在分析查询日志后总结出:查询出现频次越大,该查询越可能是导航类查询,是对权威信息的查找。Duan 等[29]将用户意图分为导航类和非导航类,认为与名词共现的动词能表达其意图,再利用动词—名词(verb-noun)之间的依存关系识别非导航类查询中的子类。Truran 等[30]认为如果查询表达中含有价格、购买、出售等字样或直接是对商业类网站比如"淘宝网"的查询,该查询具有商业意图。Chien[31]和 Gruhl[32]对查询的时间敏感度进行了分析,发现了一个普遍存在的现象,即查询总是在小段时间内非常流行,尤其对某类查询特别适用,比如新闻类。Nguyen 等[18]指出如果某查询既可能和时间、空间名称一起出现,又可能单独出现,这种查询属于时间、地理敏感型。Gravano 等[22]发现全局查询通常不包含地名,局域查询一般都包含地名。Jones 等[24]认为参考用户的查询语言可了解查询的地理属性,但是只限于大范围的国家层面。Lee 等[33]认为识别地理属性最简单的方法是在地名字典里面匹配查询中的地理名词,然而这种方法只适用于查询显性包含地名并且地名没有歧义的情况。于是,Nguyen[18]提出借助外部词汇数据库对歧义查询进行消歧。Jones 等[34]从所有样本查询中统计话题的地理距离,对一些常用话题的距离做排名,并认为电影院、日托、医院等相关或相似查询均有距离限制,属于局域查询。Smith 等[35]认为如果查询同时含有人名和地名,比如"中山公园",该查询属于地理查询的范畴,如果仅含有人名,是对人物的查找,不具有地理属性。Lau 等[36]认为用户查询的长度代表了对所查找信息的重视程度,查询越长,所查找的信息也越专业。从以上研究可知,基于查询表达式的特征提取主要关注用户的查询词和查询长度。

2.2 基于检索结果的特征选取

通常情况下,查询无法提供足够的特征信息,并且不一定能真实反映用户的信息需求,所以仅仅依靠查询表达对查询意图进行分类效果并不理想,为了解决这个问题,多位学者提出借助外部知识,尤其是检索结果进行查询分类。该方法基于如下假设:搜索引擎针对特定查询检索出的最靠前的一系列检索结果与查询相关。Kang 等[1]提出根据查询词在检索结果中文档、标题和锚文本中的出现方式来识别用户的查询类型,并提出了锚使用率、查询词分布和词间依赖三个特征,认为导航类查询的检索词在结果页的锚文本中出现概率较大,在网站主页中出现次数较多,词与词间的依赖性较强。在此文的基础上,文献[37]对其进行了扩展,利用锚文本链接类型来识别用户意图,其中包括事务类意图。虽然以上两种方法假设合理,但是可操作性并

不强,因而不能有效识别用户的查询类型。一般而言,导航类查询的结果页是少量权威网页,大多数用户会选择链接到相同网站,因而结果页中锚文本分布的偏斜度(skewness)较大。基于此,Lee[5]和Yuan[38]做了相关工作。Lee 提出了锚—链接(Anchor-link)分布特征,并绘制了导航类和信息类的锚—链接分布图形,证实了上述假设。Yuan 引入了链接熵(Link Entropy)和网站熵(Site Entropy)来定量计算查询属于导航类和信息类的概率,当用户使用锚文本作为查询时,可能是想查找权威网页,为导航类查询,两个熵值都较小,当熵值较大时,查询很有可能是信息类查询。Dai 等[39]提出根据结果页中广告的多少来判断查询是否具有商业意图,并指出商业类查询的结果页比较稳定。

在对查询主题进行分类方面,现在主流的研究方法是先对查询结果分类,再将查询划分到这些类目中。Broder 等[40]认为根据某篇检索出的文档可以确定查询主题,但这仅仅适用于非模糊查询。为了解决这一问题,Song 等[41]提出借助文档可同时属于多个类别的思想识别模糊查询:如果某查询的检索结果涉及多个话题,该查询很有可能是模糊查询。Chang 等[42]借助查询片段,使用概率推理模型,识别用户可能的查询意图,认为在查询片段中出现概率最大的目标为用户的查询意图。Nguyen 等[18]提出根据目标答案在结果页中出现的频率识别权威敏感查询。He 等[43]根据查询结果中的示意动词、URL 信息和标题等来识别用户的意图。Radlinski 等[44]基于 TREC Web Track 查询集,根据查询日志中每个查询的修改和点击情况来识别查询意图,但是该方法只能识别出与用户意图相关的词,未能定位到真正的查询意图。Vallet 等[45]通过对查询结果中的实体类别进行排序来识别用户意图。Dai 等[39]认为专业搜索引擎检索结果能满足用户的特定信息需求,可以通过对来自不同搜索引擎的查询结果进行过滤、选择和排序来识别用户意图。

2.3 基于用户行为的特征选取

结果页是系统自动呈现给用户的,表达的是系统设计人员对查询的理解,并不能真正代表用户的查询意图,因而用户行为成为分析查询意图最有力的助手。用户行为是指用户检索过程中表现出来的行为,是用户检索目标的显性表达,是最能体现用户查询意图的特征。学界对用户行为的研究主要集中在三个方面:用户交互行为、用户点击行为和语境变化。用户交互行为旨在捕获用户在结果页中的行为事件,是识别用户查询意图和个性化检索的重要途径。早在 2008 年,Buscher 等[46]就验证了目光追踪识别查询意图的有效性,但是这种方式需要昂贵的设备投入,没能在学界引起共鸣。近年来,学者们开始注重用户鼠标活动及目光停留时间的研究,停留时间越长表明用户对查询结果越满意[47],Guo 等[13]进一步量化了满意度时间,认为停留时间超过 30 秒为满意,不足 15 秒为不满意。Mendoza 等[48]发现用户花在导航类查询上的时间比信息类少。Muller 等[49]对鼠标活动做了详细研究,认为如果用户在做第一次选择时迟疑很久,那么第二次选择和第一次选择会非常相关;如果用户将鼠标移动到空白区域,可能对查询结果不满意或者对结果比较犹豫。语境变化旨在捕获查询提交情况。一些学者在这方面进行了探索,如 Jansen 等[50-51]在对大量查询日志分析的基础上发现,在执行导航类查询时,用户只浏览第一页的查询结果,只进行一次会话,无视相关查询;在执行信息类查询时则会频繁提交查询,会话次数较多,较关注系统建议的相关查询。Huang 等[52]研究发现用户在不满意查找结果时会提交更多查询,或者选择使用高级检索,在结果页上花费的时间增加,倾向提交更复杂的检索表达式。Huntingon 等[47]发现如果用户使用单个检索词,仅执行一次查询,那么用户并没有明确的查找意图,只希望寻找一些简单的相关内容,若用户使用较长的检索表达,则他很有可能会执行额外的查询以获得更精确的信息。

用户点击行为旨在捕获查询结果点击的类型和属性,是导航类和信息类查询分类的重要参考依据。一般意义上,用户使用导航类查询是想寻找少量权威网页,因而这些网页被点击的概率很大,点击分布图形坡度较大,而信息类查询却正好相反。

基于这个前提,多位学者做了研究工作。Lee 等[5]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[53]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,并提出了 N 个点击满意度(nCS)和前 N 个结果满意度(nRS)指标。Yuan 等[38]基于同样的假设提出了点击熵(click entropy)和域名点击熵(domain click entropy)。Ashkan 等[12]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。Brenes 等[54]通过分析查询日志中的点击数据提出了三类导航类查询特征,cPopular(被点次数最多的 URL 占该查询所有被点 URL 的比例)、cDistinct(被点的不同 URL 个数占所有被点 URL 个数的比例)和 cSession(只包括查询 q 的 Session 占所有包含查询 q 的 Session 的比例)。Mendoza 等[6]同样认为查询(或文档)中的词及点击次数是分析用户意图的有力工具。

虽然有大量学者对查询意图分类特征进行了研究,但是人工分类主要使用查询表达式方面的特征,自动分类则多借助查询结果中的锚文本链接和用户的点击行为。

3 查询意图分类方法

对查询意图分类的探讨始于人工分类。最初,Broder[2]就是通过用户调查和日志分析将查询人工划分为三类,Rose 等[3]使用日志分析人为地扩展了 Broder 的思想。Steven 等[55]让 AOL 编辑人员将查询分为 18 个主题类目。Law 等[56]考虑到对查询意图进行人工标注费时费力,于是采用人工计算方法,构建一个大众喜欢的在线游戏,让用户以一种玩游戏的方式来完成数据标注,其标注方式不是从传统的根据查询标注其意图类别,而是给定意图类别,标注出可能包含的查询。

因为人工分类在低频查询面前分类效果欠佳,于是有学者提出了自动分类的思想,在一定程度上理解大规模查询日志的属性,提高系统在单个查询上的有效性。对查询意图自动分类的探讨,始于 Kang 等[1],他们基于主页类查询和话题类查询在各种特征上的分布差异提出了分布差异算法,并验证了该方法的有效性。Lee 等[5]认为 Kang 等的分类特征有效性不够,选取了其他特征验证了分布差异算法在分类上的有效性。Liu 等[53]则使用典型决策树算法将 nCS、nRS 和点击分布三种特征结合起来执行分类任务,得到了比 Lee 更好的分类效果。另一方面,Mendoza[48]首次尝试基于用户日志,分别利用 SVM 与 PLSA 对查询意图进行归类;Ashkan 等[12]使用 Matlab 建立 SVM 和核方法执行自动分类功能;Yuan[38]和吴[8]同样借助 SVM 验证了新分类特征的有效性。Gravano 等[22]使用机器学习识别查询的地理属性,该研究表明数据的稀疏性会严重影响分类效果,要想获得满意的分类结果应该借助外部资源,比如用户反馈和辅助数据库去拓展查询特征。除查询外,Nettleton 等[57]利用自组织图将用户 Session 分为信息、导航和事务类。

在信息类即查询主题的自动分类方面,国内外也有一些相关成就,如 2005 KDD Cup 参赛者 Shen[58]、Kardkovacs[59]和 Vogel[60]等使用不同算法证明了映射传递的有效性,即先将查询映射到中间类目,然后再将查询从中间类目映射到目标类目。但是该方法存在两个潜在缺陷:第一,只要目标类目的结构发生变化,第二次映射的分类器就需要再训练,而在实际应用中,目标类目取决于服务供应商的需求和网络内容的分布,因而该方法不够灵活;第二,使用 ODP 作为中间类目造价很高。于是,Shen 等[61]对该方法做了改进,引入查询分类算法,先建立离线模式的中间分类器,然后在线使用该分类器通过中间分类法将查询映射到目标类目中。Broder 等[40]提出基于伪相关反馈对查询主题进行分类,根据检索结果的类目决定查询类目。Steven[55]利用查询简短的特点,将选择优先性用于查询主题分类。选择优先性原本是语言学中的方法,描述词语在句子中的搭配情况,比如动词“吃”后面经常跟的是食物的名字。这种方法相对于把查询作为文档分类的方法来说,更倾向于去理解查

询的语言学结果在统计学上的意义。使用选择优先性可以确定查询里的未知类别和歧义词汇可能的类别。He 等[43]也对查询主题分类进行了探讨,用查询结果中的某些词项代表查询构建高维空间,并借助点击信息将查询间的语义关系作为回归因素运用到学习系统中,而实验结果也证明了该方法的有效性。

另外一些学者尝试在不给定分类类目的情况下,通过借用外部资源来自动识别用户意图。如 Hu 等[62]利用外部资源即 Wikipedia 来识别用户的意图,该方法较其他方法的不同之处在于:不需要大量的人工标注集来训练分类器,减少了劳动力的投入。该方法的核心思想为:先人工标注每个意图的种子查询,然后通过挖掘 Wikipedia 的结构为 article 和 category 生成一定的意图概率,再将输入的查询映射到 Wikipedia 的概念中,根据一定算法来识别该查询的意图。Yoon 等[63]认为可以使用与查询相关的提问识别用户意图,提出借用外部资源 Yahoo Answers 获得与查询相关的提问和类别来执行分类任务。Zaragoza[64]使用聚簇手段从查询中产生类目,该方法虽然能保证充足的查询量,但是单个查询的特征仍然不足。于是,Beeferman 等[65]用"会话数据"聚类代替查询聚类。

还有很多学者基于图论研究自动分类方法,这些研究均基于一个共同的假设:具有相似点击模式的查询可能属于同一类目,利用已知类别的查询可以推导新查询的类别。比如,Li 等[66]使用二维图(bipartite graph)、Szummer[67]使用马尔科夫随机行走(markov random walks)、Zhu 等[68]使用标签传播(label propagation)、Zhou 等[69]使用局部和全局一致性学习(learning with local and global consistency)、Belkin 等[70]使用流行正规化(manifold regularization)对查询分类做了方法论研究。

一般而言人工分类的准确率较高,却面临投资大的问题,自动分类借助机器学习虽然可以处理大规模数据,但是分类的准确性较差,如果仅使用一种方法,分类结果可能不理想,因此有学者尝试对各种分类方法进行组合,来得到新的分类方法获取各自的优点。比如,Beitzel 等[71]讨论了人工分类、监督分类和规则分类三种独立方法在查询主题识别上的效果,结果表明:将三种方法结合起来会得到更佳的分类效果。

4 数据集与评价方法

数据集是查询意图分类研究尤其是自动分类

表 2 数据集来源及优缺点

类别	数据来源	优 点	缺 点
搜索引擎提供查询日志	AltaVista(http://www.altavista.com)、Excite(http://www.excite.com)、Dogpile.com (http://www.dogpile.com)、Sogou(http://www.sogou.com/labs/dl/q.html) AOL(http://www.gregsadetsky.com/aol-data/) tianwang.com、hao123.com 等其他搜索引擎提供的查询日志	查询数据样本大且具有多样性;真实的用户交互环境,贮存了用户的真实信息需求;便于统计分析,探测相关现象与规律。	数据具有稀疏性:即单个查询仅有少量特征;存在大量噪声数据;涉及用户隐私,多数查询日志较难获取。
测评会议提供数据集	TREC、KDD Cup 等会议查询集	数据集易于获取;对每个查询特征的表述内容量大且全。	数据样本少;不能完全反映真实环境,与真实网络环境存在差异。
研究者自建数据集	研究工作中利用某些方式收集到的数据:如根据查询获取与之相关的文档或网页等	对搜索引擎查询日志的补充;能获取除查询外的其他分类特征。	操作困难;用户易受外界影响,数据缺乏真实性;容易产生噪声数据。

研究的重要基础。目前,在查询意图分类领域还没有权威的数据集,各学者在进行研究时往往根据自身需要选取不同的数据集。根据数据集来源情况,笔者将其分为三类,并对各类数据集的优缺点做了简要对比分析(见表2)。

目前,查询意图研究领域没有特定的评价方法,常用的评价方法通常来自文本分类,包括正确率、召回率、精确度、错误率和F值等指标。需要注意的是,正确率、召回率和F值都只针对单个类别进行评价,代表局部性能。因此,需要引入能够在全局(所有类别)上对分类器效果进行评价的方法。目前主要有宏平均(macro averaging)和微平均(micro averaging)两种评价方法来度量在所有类别上的分类性能,微平均评价指标容易受到大类分类性能的影响,而宏平均评价指标容易受到小类分类性能的影响。

5 结语

本文详细介绍了Web查询的类目体系、分类特征、分类方法、数据集及评价方法。从上文可以看出,经过多年的努力,查询意图研究取得了很大进展。但是,查询意图领域依然存在如下问题和挑战:缺乏权威的评测标准;各种分类方法在大规模查询集合上的性能还不确定;有效提取或者获得查询特征的方法仍然值得深入;查询意图分类体系的完备性和类目间的独立不相关性尚不确定等。此外,当前查询意图分类研究大多只考虑把一个查询分到一个类目的情况,如何将分面分类应用到查询意图研究中也是一个值得探索的问题。而如何将这些识别出的意图应用在搜索引擎优化中,将是今后查询意图研究的一大趋势。

参考文献:

[1] Kang I,Kim G. Query type classification for Web document retrieval[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003: 64 - 71.

[2] Broder A. Ataxonomy of Web search[J]. SIGIR Forum, 2002, 36(2): 3 - 10.

[3] Rose D E,Levinson D. Understanding user goals in Web search[C]// WWW 2004: Proceedings of the 13th International Conference on World Wide Web, 2004: 13 - 19.

[4] Marchionini G. Exploratorysearch: From finding to understanding[J]. Communications of the ACM, 2006, 49(4): 41 - 46.

[5] Lee U,Liu Z,Cho J. Automatic identification of user goals in Web search [C]// WWW 2005: Proceedings of the 14th International Conference on World Wide Web, 2005: 391 - 401.

[6] Mendoza M,Ricardo Baeza-Yates. A Web search analysis considering the intention behind queries[C]//LA-WEB 2008: Proceedings of the Latin American Web Conference, 2008: 66 - 74.

[7] Waller V. Not just information: Who searches for what on the search engine Google?[J]. Journal of the American Society for Information Science and Technology, 2011, 62(4): 761 - 775.

[8] 吴晓晖,宋萍萍,张荣欣. 有无查询意图的分类与实现架构模型研究[J]. 情报科学, 2009, 27(12): 1830 - 1833. (Wu Xiaohui,Song Pingping,Zhang Rongxin. Research on implementation framework model and classification based on query intention and non-query intention[J]. Information Science, 2009, 27(12): 1830 - 1833.)

[9] 杨艺,周元. 基于用户查询意图识别的Web搜索优化模型[J]. 计算机科学, 2012, 39(1): 264 - 267. (Yang Yi, Zhou Yuan. Web retrieval optimization model based on user's query intention identification[J]. Computer Science, 2012, 39 (1): 264 - 267.)

[10] Lux M,Kofler C,Marques O. A classification scheme for user intentions in image search[C]//Proceedings of the 28th

International Conference Extended Abstracts on Human Factors in Computing Systems, 2010: 3913 - 3918.

[11] Kofler C. An exploratory study on the explicitness of user intentions in digital photo retrieval[C]//Proceedings of the 9th I-KNOW and I-SEMANTICS, 2009: 208 - 214.

[12] Ashkan A, Clarke C L A, Agichtein E. Classifying and characterizing query intent[C]//Proceedings of the 31th Annual European Conference on Information Retrieval Research (ECIR' 09). Berlin, Heidelberg, 2009: 578 - 586.

[13] Guo Q, Agichtein E. Ready to buy or just browsing? Detecting Web searcher goals from interaction data[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010: 130 - 137.

[14] Park S. Analysis of characteristics and trends of Web queries submitted to NAVER, a major Korean search engine[J]. Library and Information Science Research, 2009, 31(2): 126 - 133.

[15] Segev E, Ahituv N. Popular searches in Google and Yahoo!: A "Digital Divide" in information uses?[J]. The Information Society, 2010, 26(1): 17 - 37.

[16] Ross N C M, Wolfram D. End user searching on the Internet: An analysis of term pair topics submitted to the excite search engine[J]. Journal of the American Society for Information Science, 2000, 51(10): 949 - 958.

[17] Li X, Wang Y Y, Acero A. Learning query intent from regularized click graphs[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008: 339 - 346.

[18] Nguyen V B, Kan M Y. Functional faceted Web query analysis[C]//WWW 2007: Proceedings of the 16th International Conference on World Wide Web, 2007: 32 - 39.

[19] Gonzalez C, Beaza-Yates R. A multi-faceted approach to query intent classification[C]//Springer Berlin / Heidelberg, 2011: 368 - 379.

[20] Kanhabua N, Nørvåg K. Determining time of queries for re-ranking search results[C]//Springer Berlin / Heidelberg, 2010: 261 - 272.

[21] Jones R, Diaz F. Temporal profiles of queries[J]. ACM Transactions on Information Systems (TOIS), 2007, 25(3): 1 - 31.

[22] Gravano L, Hatzivassiloglou V, Lichtenstein R. Categorizing Web queries according to geographical locality[C]// Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM), 2003: 325 - 333.

[23] Ding J, Gravano L. Computing geographical scopes of Web resources[C]// Proceedings of the 26th International Conference on Very Large Databases (VLDB'00), 2000: 326 - 348.

[24] Jones C, Abdelmoty A, Fu G. Maintaining ontologies for geographical information retrieval on the Web[C]//Springer Berlin / Heidelberg, 2003: 934 - 951.

[25] Spink A, Wolfram D, Jansen M B J. Searching the Web: The public and their queries[J]. Journal of the American Society for Information Science and Technology, 2001, 52(3): 226 - 234.

[26] 张森,王斌. Web 信息查询意图分类技术综述[J]. 中文信息学报, 2008(22): 75 - 82. (Zhang Sen, Wang Bin. A survey of Web search query intention classification[J]. Journal of Chinese Information Processing, 2008(22): 75 - 82.)

[27] Bernard J, Jansen D, Amanda S. Determining the user intent of Web search engine queries[C]// WWW 2007: Proceedings of the 16th International Conference on World Wide Web, 2004: 1149 - 1150.

[28] Belkin N J, Kelly D, Kim G. Query length in interactive information retrieval[C]// Proceedings of the 26th Annual International ACM Conference on Research and Development in Information Retrieval, 2003: 205 - 212.

[29] Duan R, Wang X, Hu R. Dependency relation based detection of lexicalized user goals[J]. Ubiquitous Intelligence and Computing Lecture Notes in Computer Science, 2010(6406): 167 - 178.

[30] Truran M,Schmakeit J,Ashman H. The effect of user intent on the stability of search engine results[J]. Journal of the American Society for Information Science and Technology, 2011, 62(7): 1276 - 1287.

[31] Chien S,Immorlica N. Semantic similarity between search engine queries using temporal correlation[C]// WWW 2005: Proceedings of the 14th Conference on World Wide Web, 2005: 2 - 11.

[32] Gruhl D,Guha R,Nowell D. Information diffusion through blog space[C]// WWW 2004: Proceedings of the 13th International Conference on World Wide Web, 2004: 491 - 501.

[33] Lee W,Wang C,Xie X. Detectingdominant locations from search queries[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005: 424 - 431.

[34] Jones R,Zhang W,Rey B. Geographicintention and modification in Web search[J]. International Journal of Geographical Information Science, 2008, 22(3): 1 - 20.

[35] Smith D,Crane G. Disambiguating geographic names in a historical digital library[C]//Springer Berlin / Heidelberg, 2001: 127 - 136.

[36] Lau T,Horvitz E. Patterns of search: Analyzing and modeling Web query refinement[C]// Proceedings of User Modeling, 1999: 119 - 128.

[37] Kang I H. Transactional query identification in Web search[C]// Proceedings of the 2nd Asia Conference on Asia Information Retrieval Technology, 2005: 221 - 232.

[38] Yuan X-J,Dou Z-C,Zhang L. Automatic user goals identification based on anchor text and click-through data[J]. Wuhan University Journal of Natural Sciences, 2008, 13(4): 495 - 502.

[39] Dai H,Nie Z,Wang L. Detecting online commercial intention (OCI)[C]//WWW 2006: Proceedings of the 15th International Conference on World Wide Web, 2006: 829 - 837.

[40] Broder A,Fontoura M,Gabrilovich E. Robust classification of rare queries using Web knowledge[C]//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007: 231 - 238.

[41] Song R,Luo Z,Wen J. Identifying ambiguous queries in Web search[C]//WWW 2007: Proceedings of the 16th International Conference on World Wide Web, 2007: 1169 - 1170.

[42] Chang Y,He K,Yu S. Identifying user goals from Web search results[C]//Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006: 1038 - 1041.

[43] He K. Improving identification of latent user goals through search-result snippet classification [C]//Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence, 2007: 683 - 686.

[44] Radlinski F,Szummer M,Craswell N. Inferring query intent from reformulations and clicks[C]//WWW 2010: Proceedings of the 19th International Conference on World Wide Web, 2010: 1171 - 1172.

[45] Vallet D,Zaragoza H. Inferring the most important types of a query: A semantic approach[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008: 857 - 858.

[46] Buscher G,Dengel A,Elst L. Query expansion using gaze-based feedback on the subdocument level[C]//Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008: 387 - 394.

[47] Huntington P,Nicholas D,Jamali H. Employing log metrics to evaluate search behavior and success: Case study BBC search engine[J]. Journal of Information Science, 2007, 33(5): 584 - 597.

[48] Mendoza M,Zamora J. Identifying the intent of a user query using support vector machines[C]//Springer Berlin / Heidelberg, 2009: 131 - 142.

[49] Muller F,Lockerd A. Cheese: Tracking mouse movement activity on websites,a tool for user modeling[C]//Proceedings of CHI Extended Abstracts on Human Factors in Computing Systems, 2001: 279 - 280.

[50] Jansen B J. Seeking and implementing automated assistance during the search process[J]. Information Processing & Management, 2005, 41(4): 909 - 928.

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

[52] Huang J,Efthimiadis E N. Analyzing andevaluating query reformulation strategies in Web search logs[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009: 77 - 86.

[53] Liu Yiqun,et al. Automatic query type identification based on click through information[C]//Springer Berlin / Heidelberg, 2006: 593 - 600.

[54] Brenes D,Gayo-Avello D. Automatic detection of navigational queries according to behavioral characteristics[C]//Proceedings of Special Interest Group Information Retrieval, 2008: 41 - 48.

[55] Steven M,Jensen E C,Lewis D D. Automatic classification of Web queries using very large unlabeled query logs[J]. ACM Transaction on Information Systems (TOIS), 2007, 25(2): 1 - 29.

[56] Law E,Mityagin A,Chickering M. Intentions: A game for classifying search query intent[C]//Proceedings of the 27th International Conference Extended ABSTRACT on Human Factors in Computing Systems, 2009: 3805 - 3810.

[57] Nettleton D,Calderon L,Baeza-Yates R. Analysis of Web search engine query sessions[C]//Proceedings of WEBKDD, 2006: 1 - 14.

[58] Shen D,Pan R,Sun J. Our winning solution to query classification in KDD Cup[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 100 - 110.

[59] Kardkovacs Z,Tikk D,Bansaghi Z. The ferrety algorithm for the KDD Cup 2005 problem[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 111 - 116.

[60] Vogel D, Bickel S, Haider P. Classifying search engine queries using the Web as background knowledge[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 117 - 122.

[61] Shen D,Sun J T,Yang Q. Building bridges for Web query classification[C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006: 131 - 138.

[62] Hu J,Wang G,Lochovsky F. Understanding user's query intent with Wikipedia[C]//WWW 2009: Proceedings of the 18th International Conference on World Wide Web, 2009: 471 - 480.

[63] Yoon S,Jatowt A,Tanaka K. Intent-based categorization of search results using questions from Web Q&A corpus[J]. Web Information System Engineering Lecture Notes in Computer Science, 2009(5802): 145 - 158.

[64] Zaragoza H. Information retrieval: Algorithms and heuristics[J]. Information Retrieval, 2002, 5(2): 271 - 274.

[65] Beeferman D,Berger A. Agglomerative clustering of a search query log[C]//Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000: 407 - 416.

[66] Li X,Wang Y,Shen D. Learning with click graph for query intent classification[J]. ACM Transactions on Information Systems, 2010, 28(3): 121 - 140.

[67] Szummer M,Jaakkola T. Partially labeled classification with Markov random walks[J]. Advances in Neutral Information Processing Systems, 2001, 14(2): 945 - 952.

[68] Zhu X J,Chahramani Z B. Learning from labeled and unlabeled data with label propagation. CMU - CALD - 02 - 107 [R]. Pittsburgh: Carnegie Mellon University, 2002.

[69] Zhou D,Bousquet O,Lal T. Learning with local and global consistency[J]. Advances in Neutral Information Processing

Systems, 2003, 16(2): 321 -328.

[70] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning form labeled and unlabeled examples[J]. The Journal of Machine Learning Research, 2006, 7(1): 2399 -2434.

[71] Beitzel S, Jensen E, Frieder O. Automatic Web query classification using labeled and unlabeled training data [C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005: 581 -582.

陆 伟 武汉大学信息资源研究中心教授,博士生导师。通讯地址:武汉市珞珈山。邮编:430072。

周红霞 武汉大学信息管理学院 2011 级情报学硕士研究生。通讯地址同上。

张晓娟 武汉大学信息管理学院 2011 级情报学博士研究生。通讯地址同上。

(收稿日期:2012 -06 -13)