

浅谈“查全/查准率”

□ 吴广印 / 中国科学技术信息研究所 北京 100038

摘要: “查全/查准率”是情报检索系统评估的第一要素,目前我们评估这一指标主要是从软件技术和检索语言方面考虑的,而真正影响这一指标的是系统管理数据的质量、加工深度、知识库等因素。文章从“知识获取五要素”的角度对学科、人物、机构、主题、基金等关键数据项的加工深度与“查全/查准率”关系进行了详细论述。

关键词: 情报检索系统, 查全率, 查准率, 五要素, 后控检索, 前端导航

DOI: 10.3772/j.issn.1673-2286.2010.07.007

引言

“查全/查准率”对情报检索而言是一个“古老”而热门的话题,它基本上随着情报检索这一学科的诞生而提出。而今信息检索技术的快速发展,似乎已经从传统的情报检索跨越到“智能检索”、“知识服务”等高级检索服务的新阶段。从被大众接受的Google、百度等搜索引擎,到提供专业情报服务的万方数据知识服务系统、中国知网、维普资讯,这些系统的“查全/查准率”到底如何?几乎所有的检索服务系统都评价自己的系统具有较高的“查全/查准率”。本人多年来一直从事情报检索系统的开发和研究工作,尤其近几年一直在从事基于文献内容信息服务的研究与管理工作,始终不忘“查全/查准率”这一检索系统的终极评价目标。本文的主要内容是谈谈自己在多年工作经验的基础上对“查全/查准率”的一些认识和看法,希望能够抛砖引玉,引起同行和服务商对这“古老”指标的高度重视。

1 概述

1955年,美国的佩里(J. W. Perry)和肯特(A. Kent)最先提出了查全率和查准率的概念。自从英国学者克里维顿(C. M. Cleverdon)在他著名的Cranfield I试验中首次将查全率(Recall)和查准率(Precision)

作为信息检索系统效率的评价指标,并从试验中得出查全率与查准率存在互逆关系的结论以后,这两个指标就一直成为对信息系统进行评价和试验的重要指标^[1]。查全率用来衡量在一次检索中检出相关文献的能力;查准率用来衡量系统拒绝不相关文献的能力。

查全率和查准率是评价信息检索系统效率的两个重要指标。查全率和查准率反映了系统的过滤能力,即系统让所需要的文献通过,并阻止不需要的文献。只有将二者相结合,才能对一个系统作出全面的评价,单独使用任何一个都不能说明检索效果的优劣。此外查全率和查准率受多种因素的影响,例如,情报检索语言的影响、标引的影响、检索系统的影响、用户对检索目标主观因素的影响及查全率与查准率相关性的制约等,这些影响因素在系统评估工作中必须考虑到。对于一个检索系统来说,最理想的情况是相关文献完全被检出,这时查全率和查准率达到最大,意味着标引、词表、检索策略以及用户/系统交互都完全是理想的。查全率和查准率作为信息检索系统对用户信息需求的响应能力的评价指标,通过该系统的检索指令和标引标识组成的检索提问式在该系统进行匹配而得到的结果中反映出来,是检索系统质量和检索策略质量的综合体现^[2]。到目前为止,学术界对上述影响因素讨论最多的是检索系统的算法和交互能力,其次是检索语言;对于系统的数据加工、标引深度和质量产生的重要影响的相关讨论很少。

用户到底通过哪些途径（检索项）来获取信息和知识呢？2007年作者在多年研究成果的基础上提出了“知识获取五要素”的检索服务体系。“五要素”指的是学科、人物、机构、主题和基金（国家或组织提供专项经费支持的事业），用户通常获取知识的检索途径主要是这五个检索项（关于“知识获取五要素”的详细论述见参考文献[3]）。本文主要讨论基于“知识获取五要素”文献质量对检索系统“查全/查准率”的影响。

2 基于“知识获取五要素”的“查全/查准”探讨

2.1 学科

学科是我们在一个专业概念上对一个文献源的分类，专业人员一般是按中图分类法进行分类的，也是用户获取检索文献的专业限定条件。目前只有专业情报检索服务系统提供这一限定检索支持，而Google、百度这样的面向大众的检索系统都不支持。学科分类是分级的，也就是说，你的分类越细，系统的检索准确度越高，其分类加工的难度也越大。之所以一般搜索引擎不支持分类检索，主要是因为分类是一项很专业的工作，一般需要专业人员真正理解了文献所阐述的内容后，才能给出正确的分类。为了节省数据加工成本，计算机自动分类系统就成了市场需要的一个热门产品；由于文献分类的专业性强、涉及技术复杂，到目前为止市场上还没有比较成功的文献自动分类系统，其主要原因是分类需要计算机较好地理解文献的内容（有待自然语言理解这一学科的进一步发展），计算机要做到这一点，还需要很长时间。另外目前国内有许多分类体系，都是针对不同领域的用户和应用而制定的；作为一个文献系统，所面对的用户群体是多样化的，尤其基于互联网的面向大众的信息检索系统，而对一个系统同时做多种分类也是不现实的。一般而言，一个服务系统，按一个分类标准对数据进行分类，在服务展示时转换成不同的对应分类给不同的用户。比如：万方数据资源服务系统内部数据分类采用的是中图分类体系，而一般用户并不知道中图分类法是如何分类的，在最终用户那里展示的是一个简单的分类提示。学科分类是文献检索的一个基础性很强但又很重要的检索条件，一个系统分类的质量直接

影响系统的“查全/查准率”。

2.2 人物

对于一篇文献而言，人物可能指的是作者、成果完成人、著作权人、专利拥有者等被描述的人物对象。要确定一个文献对象，人物是我们检索的一个非常重要的途径，特别对于专业人员，同行的名字还是比较明确的，尤其是行业专家；我们获取知识，首先想到的专家、教授、行业领头人。比如：情报界很多人知道我是搞情报检索的，这个专业的人要想获取文献，可能会先想到我的名字，看看我有什么新观点、新成果等。按人物“吴广印”进行检索时，可是文献检索结果中发现了医学和生物学的文章，再仔细一看是另外两个同名学者的相关文献。对于这样的检索系统，如果收集的文献是完整的，只能说“查全率”高，肯定不能说查准率好，其实也很难说查得全，因为我还有在国外发表的文章，用的是拼音拼写的名字。其实针对这样一个人物重名的问题，目前国内外哪家搜索系统能够说具有很好的“查全/查准率”呢？我们看看目前万方数据提供的“科技创新辅助决策系统”（以下简称STADS）是如何解决这一问题的。



图1 万方数据提供的“科技创新辅助决策系统”

如图1，在人物框内输入“吴广印”后，系统随后提示吴广印发表文章时所在单位的名称，用户如果了解我，自然就可以选出我的所在单位，用户起码应该知道他需要的相关文献出自我所在的那一个单位。一旦用户做出了选择，STADS系统在后面的统计分析过程中会自动把所选定的人物单位进行限定，这样最终

的检索结果也就不会出现另外两位“吴广印”的文献了。另外，如果我的一些文献在加工时没有标识作者单位，也就自然会漏检，所谓查全不够。显然如果把姓名的拼音英文名字也扩展进去，系统的查全率会更高一些。在举一个人物的例子：我们要查找有关“毛泽东”的相关文献，系统能检出“毛主席”、“毛润芝”、“李得胜”等相关的文献吗？当然也有可能某个老百姓和主席同名。显然要保证人物的“查全/查准率”需要做大量的必需的辅助工作，构建一个人物知识库是必需的。

2.3 机构

和人物一样，机构也是人们获取知识的一个关键要素，“某某单位是这个学科的权威”就是用户获取知识很好的一个线索，如果和学科、人物组配检索，想必产生的结果是很精确的。我们先举一个机构检索的例子：我们输入“中国科技信息研究所”作为导师单位，在学位论文文献里检索时，命中文献为空，因为中国科技信息研究所是业界的“俗称”，规范名称是“中国科学技术信息研究所”，而学位论文的单位名称都采用的是单位的规范名称。即使你使用了规范名称“中国科技信息研究所”，这个单位以前还有“中国科技情报研究所”这个曾用名呢。我们还是先看一下STADS系统对机构处理的输出结果：



图2 STADS系统对机构处理的输出结果

如图2所示，我们只输入了中国科技信息研究所，而系统同自动扩展了对应单位的简称、曾用名和对应的英文名称，这样可以说系统的查全率提高了。实际

上STADS系统内部使用了一个有关机构的知识库（机构词表）。但是如果作者在写文章时使用了“中国科学技术信息研究所文献中心”，又该如何？如果系统只用来做检索，通过全文检索也许部分能够解决这个问题；如果STADS系统需要对机构进行统计分析，甚至要做排名推荐，那又该如何？我的建议是：只有对文献数据库进行深加工，也就是在作者给出的机构名称后面注释一个规范名称，统计、排名时使用这个规范名称。当然这样的深加工需要大量的人力和物力，一个高质量的二次文献库，绝不是只靠简单采集或扫描识别就可以做到的。

2.4 主题

“主题”是所有检索系统都支持的检索途径，一般而言的“查全/查准率”也更强调的是主题检索，主题检索是所有用户都必需的。我们先看一下现在互联网上的检索系统功能如何。目前最好的搜索引擎是Google，Google的特点是简单易用，也有简单的人物、机构识别，对于前面提到的人物、机构检索存在的问题，在Google系统里也是明显存在的。Google所支持的是一种简单的全文检索，比如你输入了“吴广印”，只要文献里有这个字符串就行了，它没有题目、作者、主题词、摘要等字段内容的限制。Google也提供精确检索和模糊检索，如果把检索用英文引号括起来就是精确检索，即检索命中的结果中一定包含完整的检索词，否则只要有一个字出现都是命中目标。在这一点上，专业检索系统要专业一些，起码它可以限定要检索的字段。我们举一个主题检索的例子，检索词为“移动终端”，现在好的检索系统也就是能正确地命中包含“移动终端”字符串的文献，有些系统可能还会把“移动终端计算”也检出来，这也是多年情报检索领域一直争论的问题，理论上“移动终端”和“移动终端计算”不是一回事。那么“手持设备”、“手机”等等是不是“移动终端”呢？这个问题是很显然的，如果这问题得不到解决，也就很难说系统的“查全/查准率”如何了。这类问题的实际上是同义词和相关词的问题，系统如果自动把主题词按同义词、相关词、对应外文词等进行扩展检索，系统的查全率就会得到很大程度的提高。从技术层面上看，这是一个基于词表的后控检索的问题，也是多年前大家就提出来的问题。中国科技信息研究所多年前就开始了这

方面的研究和产品开发^[4],但目前市场还没有一个较为成功的系统存在,其原因是没有一个好的可用于后控检索的词表系统。上个世纪中国科学技术情报研究所编写了一个“汉语主题词表”,可多年来一直再也没有做过修订,也无法机读;最近他们正在组织一个课题要重建一个新的工程词表,如果这个词表能够尽快建成并投入应用,一定会促进这一问题的进一步解决。关于主题词表的建立、应用的重要性,本人深有体会,在此不作细论。

2.5 基金

基金项目的科研产出是和学科、前沿课题、国家急需技术等紧密相关的高质量的科研产出,无论是科研前沿的研发人员还是政府科研管理、投资部门都需要掌握某一基金项目的科研产出情况。因此在科技文献里“项目来源”(出自某一基金项目)是很重要的一个数据项,这一数据项理论上和“查全/查准率”关系不大。但由于目前绝大部分作者对基金项目的描

述不规范,而文献加工时又没有作进一步规范,看看STADS系统处理基金项目的导航结果就可想而知了。

用户输入了863导航关键词,系统显示了可选择的基金计划名称,显然选择框里的提示都可能是用户输入的基金名称,我们要对863计划的科研产出进行统计,看来也只有选择这些所有的名称,让系统进行扩展检索了。基金项目的问题比较简单,只要在文献加工过程中进行基金项目规范即可。

3 结语

综上所述,本文针对“知识获取五要素”对检索系统的“查全/查准率”所存在的问题进行了详细论述,显然导致目前检索系统普遍存在的“查全/查准”问题的原因主要在二次文件的加工深度和质量方面,这也是国内信息内容服务商普遍存在的问题。近几年大家都把信息内容的服务放在加工全文的数量上面,而忽略了二次文献数据质量的建设,今天我们重谈二次文献质量问题,几乎已经是一个很难解决的问题了。以万方数据的期刊数据为例,目前已经具备1500多万的记录的数量,可谓“海量数据”,要想重新对原有文献进行深度加工,其工作量和需要的人力、物力之大可想而知。针对这一问题,万方数据在2006年初就明确提出了建立“知识获取五要素”知识库,利用前端导航技术提高系统“查全/查准率”的解决方案。经过近四年的努力工作,在STADS系统里已经实现了“基于知识获取五要素”的前端导航服务,这对于提高系统的“查全/查准率”而言,其效果是明显的。本文虽然主要讨论的是二次文献的加工质量和深度对检索系统“查全/查准”的影响,并旨在引起文献内容服务商的高度重视,检索系统的成败很大程度上取决于二次文献的质量(文献质量并不仅限于“五要素”字段),这尤其对未来的知识服务系统更为重要,但正如本文开篇所言,影响“查全/查准”的因素还有软件系统,检索语言等重要因素。



图3 STADS系统处理基金项目的导航结果

参考文献

- [1] 范晴,徐建华,宋震. 查全率与查准率关系初探[J]. 情报杂志,2002,21(9).
- [2] 余丹. 关于查全和查准率的新认识[J]. 西南民族大学学报(人文社科版),2009(02):283-285.
- [3] 吴广印. 知识获取“五要素”的研究与实践[OL]. <http://www.docin.com/p-14735336.html>.
- [4] 吴广印,胡亚莉. 基于Internet的汉语后控全文检索系统的研究与开发[J]. 图书情报工作,2002(03):91-95.

- [5] 张琪玉. 情报检索语言实用教程[M]. 武汉: 武汉大学出版社, 2004.
[6] 邓汉成, 王瑛, 王敬芳. 从检索实例看查全率与查准率之间的关系[J]. 情报学报, 2000(6): 237-241.
[7] 沈艳红. 信息检索中检索词的选择对查全率的影响[J]. 情报探索, 2006(11): 73-74.
[8] 李育婧. 文献检索中提高查全率与查准率的方法探讨[J]. 图书馆学研究, 2002(11): 92-95.
[9] 科技创新辅助决策支持系统[OL]. <http://ssc.wanfangdata.com.cn/stads/download.htm>.

作者简介

吴广印, 男, 中国科学技术信息研究所研究员, 北京万方数据股份有限公司总工程师, 研究方向: 数据库内容管理、存储与检索。通讯地址: 中国科学技术信息研究所 100038。E-mail: gywu@wanfangdata.com.cn

On "Recall / Precision Ratio"

Wu Guangyin / Institute of Science and Technical Information of China, Beijing, 100038

Abstract: "Recall / precision ratio" is the first essential factor to evaluate information retrieval system. We now assess this indicator mainly from the software, but the real impact of this indicator is the system management of data quality, processing depth, knowledge base and other factors. In this paper, the relationship between "recall / precision ratio" and the processing depth of "five elements" in knowledge acquisition, i.e. subjects, people, institutions, themes, funds and other key data items is discussed in detail.

Keywords: Information retrieval system, Recall ratio, Precision ratio, Five elements, Post-controlled retrieval, Front-end navigation

(收稿日期: 2009-02-24)

业界动态

苹果宣布80天售出300万台iPad

【搜狐IT消息】6月22日消息, 苹果公司今天表示, 在iPad美国上市80天内, iPad销量已经达到300万台。苹果在声明中称: iPad是一款革命性、魔术性产品, 它允许用户与应用程序、内容、互联网相连, 连接的方式更为紧密、更有趣。

苹果CEO乔布斯称: “人们之所以喜欢iPad, 主要是因为它成为日常生活的一部分。我们正努力将这款魔术性产品送到全球更多人的手中, 下个月它将会在9个国家上市。”

得益于多触点触摸屏技术界面、大屏幕、高性能显示装置的采用, 开发者已经为iPad开发了1.1万个应用程序。iPad几乎可以运行App Store商店中的22.5万个应用, 包括用户已经通过iPhone和iPod购买的应用。

用户可以上网, 阅读和发送邮件, 分享图片, 观看HD视频, 听音乐, 玩游戏, 看书, 做更多其它的事。iPad只有0.5英寸厚, 重1.5磅, 比任何上网本都轻薄, 电池可以续航10小时。

按照数据计算, 苹果平均每天售出37.5万台iPad, 平均每小时1562台, 每分钟26台。

苹果没有透露每个月的具体销量, 不过, 在第一个月苹果宣布销量过100万, 到5月31止再度卖掉200万台。iPad最初在美国上市, 目前已经登陆澳大利亚、加拿大、法国、德国、意大利和日本, 7月还有9个国家将开售iPad。(路尘)

来源: http://it.southcn.com/9/2010-06/25/content_13176479.htm 查询日期: 2010-06-24