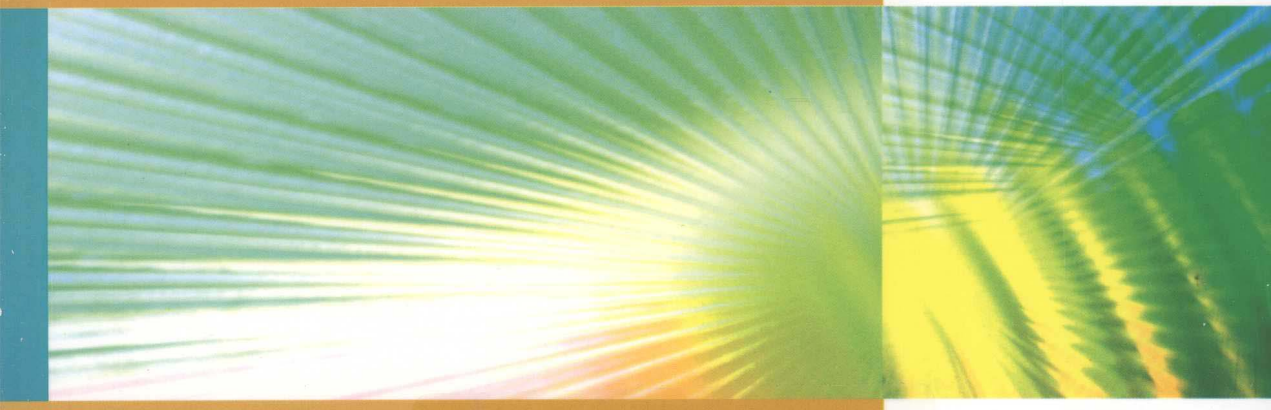
新一代信息科学与技术丛书

搜索引擎效果评测
——基于用户点击日志分析的方法与技术

何 靖 李晓明 著

Search Engine Evaluation:
Methods and Techniques Based on Clickthrough Analysis



高等教育出版社
HIGHER EDUCATION PRESS

新一代信息科学与技术丛书

搜索引擎效果评测

——基于用户点击日志分析的方法与技术

Search Engine Evaluation:

Methods and Techniques Based on Clickthrough Analysis

何 靖 李晓明 著

Sousuo Yinqing Xiaoguo Pingce

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目(CIP)数据

搜索引擎效果评测——基于用户点击日志分析的方法与技术/何靖,李晓明著. -- 北京:高等教育出版社,2012.5

ISBN 978-7-04-034470-7

Ⅰ. ①搜… Ⅱ. ①何…②李… Ⅲ. ①互联网络-情报检索 Ⅳ. ①G354.4

中国版本图书馆 CIP 数据核字(2012)第 071961 号

策划编辑 刘 英 责任编辑 冯 英 封面设计 张 楠 版式设计 杜微言
责任校对 刘春萍 责任印制 毛斯璐

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 北京市卫顺印刷厂
开 本 787mm×1092mm 1/16
印 张 9.5
字 数 180 千字
购书热线 010-58581118
咨询电话 400-810-0598
网 址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.landraco.com
http://www.landraco.com.cn
版 次 2012 年 5 月第 1 版
印 次 2012 年 5 月第 1 次印刷
定 价 49.00 元

物 料 号 34470-00

前　言

当前，互联网上的信息不仅越来越多，而且形式与内容也呈现出越来越广泛的多样性。同时，随着社会信息化的深入普及，互联网信息已经越来越成为人类社会运行不可或缺的一种基本要素。在这种背景下，搜索引擎在人们的日常工作和生活中发挥着越来越重要的作用，同时也对搜索引擎本身不断提出挑战，因而搜索引擎技术成为近十年来一直方兴未艾的一个热点研究领域。作为一个相对比较窄的领域，研究热度持续时间之长，影响之深远，应用范围之宽广，在计算机技术发展史上是少有的。

对搜索引擎的研究，主要针对三个基本的问题，即如何改进搜索引擎的服务效果，如何提高搜索引擎的服务效率，以及如何评价搜索引擎的优劣。所谓服务效果，对应的是用户对搜索引擎返回结果的满意程度。所谓服务效率，对应的是搜索引擎对用户查询的响应时间和吞吐率，特别要考虑大量用户同时访问的情况。而搜索引擎评价则是指用科学的方法评测搜索引擎在上述两个方面的表现，尤其是在服务效果方面的表现，通常会涉及指标体系的建立、实验数据的收集以及实验数据综合模型等方面的问题。用户通常不会告诉搜索引擎他们是否满意搜索引擎给出的结果，如果不满意，他们就转向其他搜索引擎。对于搜索引擎来说，争取用户是最根本的目标，流失用户是最大的损失。因此，如何评价搜索引擎质量的好坏是人们非常关注的一个问题。灵敏准确地评价搜索引擎的质量，可以帮助用户对特定的搜索问题选用合适的搜索引擎，更重要的是可以帮助搜索引擎的研究和开发人员甄别与开发高质量的搜索技术及方法。

评估搜索引擎效果的技术通常可分为三类：基于 Cranfield 范式的评测方法、基于用户研究的评价方法和基于用户隐反馈的评价方法。其中，基于隐反馈的评价方法因其自动性和准确性，越来越多地得到学术界和工业界研究者的关注。在用户隐反馈中，最容易获取、使用最广泛的是用户的点击行为。搜索引擎的用户点击日志直接记录了它们，便于被人们深入分析。本书将着重介绍基于用户点击日志分析的搜索引擎评价方法。

在本书的第 1 章和第 2 章，首先给出搜索引擎评价技术和用户隐反馈行为分析方法的综合介绍，有关材料可以作为系统了解搜索评测的一个基础，对于初次涉足该领域的读者会有很大帮助。在第 3 章和第 4 章，介绍两种基于用户点击日志分析的搜索引擎评价方法：归并 - 比较方法和用户点击模型的方法。前者直接比

较多个搜索引擎的优劣，后者对一个搜索引擎给出效果评分。第5章介绍了一种综合考虑结果展示信息质量的评价指标。本书所阐述的技术，不仅有坚实的理论基础，而且便于在实际搜索引擎系统中实现。

本书反映了作者多年的研究心得与成果，主要内容曾在若干重要学术会议上发表，但在本书中描述得更加详尽，同时也有一些首次正式成文的工作，特结集于此，希望对读者有所帮助。本书的内容，不仅包含对搜索引擎评测的概念和一些新技术的介绍，还特别强调了获得这些技术的研究过程，从基本思路、理论分析，到实验设计验证，比较充分地体现了开展搜索评测研究的一种方法论，有志投入该领域学术研究的读者会发现它们的启迪作用。

本书的研究工作先后得到多项国家科技项目的支持，包括国家自然科学基金重点项目（60933004）、核高基项目（2011ZX01042 - 001 - 001）以及863项目（2006AA01Z196），我们在此对有关项目评审专家和相关部门的支持表示感谢。与本书内容相关的工作也不是仅靠作者自己可以完成的，北大网络实验室的闫宏飞、赵鑫、树柏涵为本书的出版都作出了贡献，在此对他们表示感谢。特别是袁文清帮助实现了一个基于第3章研究成果的实验系统，给我们带来了信心和兴奋。还特别应该提到的是，作者曾在美国伊利诺伊大学香槟分校计算机系进行过为期一年的学术访问，其间得到翟成祥教授的悉心指导，在此表示特别的感谢。

作 者

2012年2月

目　录

第 1 章　搜索引擎评价技术

这是一本关于评价搜索引擎的书。近些年来，随着互联网搜索引擎的普及应用，人们对它的研究日益深入成熟，关于搜索引擎出版了不少书籍，国外的有《Search Engines: Information Retrieval in Practice》[1]、《Information Retrieval: Implementing and Evaluating Search Engines》[2]、《Introduction to Information Retrieval》[3]、《Mordern Information Retrieval》[4]等，国内的有《搜索引擎——原理、技术与系统》[5]、《搜索引擎技术基础》[6]、《Web 搜索》[7]、《网络信息检索》[8]，等等。这些书籍一般都是对搜索引擎技术的综合介绍，通常大部分篇幅用在构建搜索引擎的技术上，同时一般也都有一个章节用来讨论评价搜索引擎的技术。随着互联网信息变得越来越复杂，搜索引擎也变得越来越复杂，有效地评价搜索引擎也就变得越来越重要，我们认为有必要单独出版一本书，除了介绍搜索引擎评价的基础知识外，也介绍这方面的一些最新研究成果。

搜索引擎是最重要的信息处理系统之一，它可以帮助人们从海量数据中找出能够满足他们需求的信息。搜索引擎评价的目标是准确而又高效地判断搜索引擎的优劣。搜索引擎评价推动了搜索引擎技术的发展。譬如，自 1992 年以来，信息检索评测会议（Text Retrieval Evaluation Conference，TREC）每年召开一次，目的在于推动搜索引擎技术的进步。会议每年都定义一系列搜索任务，如 Web 搜索、问答系统、博客搜索，等等，世界范围内的所有学术界和工业界的研究机构都能够报名参加。他们各自研发搜索算法，完成这些任务，提交算法获得的搜索结果。信息检索评测会议对这些搜索算法的结果进行评价、比较和分析[9-11]，从而能够发现真正有效的搜索引擎技术。自信息检索评测会议召开以来，依赖严格而有序的评测，甄选出很多有效的搜索引擎算法，同时也积累了很多可以重复使用的数据集。系统地利用评测作为一种动力来推动技术的发展，是信息检索（搜索引擎）领域的一个重要特色。评估技术也成为信息检索研究持续关心的三大基本问题之一（另外两个问题是文档相关性和对用户需求的理解[1]）。

1.1　目标、角度和方法

为了评价搜索引擎，首先需要定义什么是一个“好”的搜索引擎。不同的人群目标不同，对于“好”的搜索引擎的界定是不同的。对这些不同的目标，应该有不

同的评价标准。

对于普通的搜索引擎用户来说,当他们有信息需求的时候,希望可以通过搜索引擎,快速地找到有用的信息,来满足他们的需要。因此,搜索引擎最主要的目标,是能够既快又好地满足搜索用户的信息需求,这也是本章讨论的重点,将在1.3节到1.11节中进行介绍。

近年来,搜索引擎成为普通用户获取互联网信息的主要途径。因此,对于互联网信息的发布者来说,搜索引擎也成为他们发布的信息传播的主要渠道。信息发布者希望他们发布的信息能够被用户通过搜索引擎获取。譬如,对于一个正在进行产品推广的公司来说,他们希望通过搜索引擎,其产品能够被有购买倾向的潜在用户发现。对于政府机关来说,他们希望发布的政策、法规能够被群众所了解。这部分的研究,将在1.12节中进行介绍。

大部分的搜索引擎评价研究是以满足普通搜索用户的信息需求作为目标的。但是,搜索用户对搜索引擎的满意程度,很难被定量地表示。一般的,用户的满意程度会取决于三个主要的方面:搜索引擎的效率、结果质量和界面。

因此,可以从这三个角度对搜索引擎进行评价。

- 搜索引擎系统的效率(Efficiency)是指它处理用户搜索请求所耗费的代价,主要指时间代价。这部分研究将在1.10节中介绍。
- 搜索引擎系统的结果质量(Effectiveness)是指它提供的结果文档序列能够满足用户信息需求的程度。这部分内容是本书讨论的重点,将在1.3节到1.9节中予以介绍。
- 搜索引擎系统的界面(Interface)通常指搜索引擎交互界面的易用性、可学性、可理解性等。这部分内容将在1.11节中进行介绍。

评价搜索引擎的好坏,主要有三种方法:基于用户研究的方法、Cranfield范式的方法和基于用户隐反馈的方法。

基于用户研究方法(User Study)[12,13]通过招募一些搜索用户,给他们指定一些搜索任务,根据他们利用搜索引擎完成任务的质量(任务是否完成,完成的好坏)、效率(完成任务的时间)和他们报告的满意程度来评价一个搜索引擎的好坏。这部分研究内容将在1.9节中进行介绍。

最常使用的搜索引擎评价方法是Cranfield范式(又称批量评测方法或面向系统的评测方法),它使用一套可重用的静态评测集来评测不同搜索引擎系统的好坏。这部分研究将在1.3节到1.7节中进行介绍。

另外一个新兴的重要研究方向是通过挖掘实际用户的搜索行为,如查询输入、点击、浏览等,发现其中所包含的隐反馈信息,利用这些信息对搜索引擎的质量进行评价。这是本书介绍的重点内容。将在第2章介绍隐反馈的建模方法,在第3章和第4章介绍两种基于隐反馈的评价方法,在第5章中介绍一种可利用隐反馈获得特征进行评价的新指标。

本书的研究定位,包括搜索引擎评价的目标、角度和方法见表1-1。

表 1-1 本书研究的定位

分类体系	类 别
搜索引擎评价的目标	普通搜索用户、信息发布者(满足他们的需求)
搜索引擎评价的角度	结果质量、效率、交互界面
搜索引擎评价的方法	用户研究、Cranfield 范式、基于用户隐反馈分析

1.2 符号定义

为了使本书的风格保持统一,本节定义了一些符号,见表 1-2。这些符号将在整本书内使用。

表 1-2 本书常用的符号定义

符 号	含 义
$q;Q$	信息需求(查询),信息需求集合
$d;D$	文档,文档集合
$u;U$	用户,用户集合
$\|X\|$	以符号 X 表示的集合的大小,即它包含的元素个数
$L(d)$	文档 d 在文档序列 L 中的位置
$R(q;d)$	文档 d 和信息需求 q 的相关程度
$S(q;L)$	对于信息需求 q,指标 S 获得的文档序列 L 的质量得分
$d_1 >_R d_2; d_1 <_R d_2$	文档 d_1 比 d_2 更加相关/更不相关
$L_1 >_p L_2; L_1 <_p L_2; L_1 =_p L_2$	文档序列 L_1 和 L_2 相比,质量更好/更坏/一样好
$L_1 >_s L_2; L_1 <_s L_2; L_1 =_s L_2$	通过某种评测方法 s 认为,文档序列 L_1 和 L_2 相比,质量更好/更坏/一样好

1.3 Cranfield 范式评价方法

最常使用的搜索引擎评价方法是 Cranfield 范式(也称为批量评测方法或面向系统的评测方法),它使用一套可重用的静态评测集来评价搜索引擎系统的好坏,是一种很自然的实验性评测方案。之所以采用这个名称是由于这种评价方法是英国 Cranfield 学院的 Cleverdon 等人最初提出并实践的。一个 Cranfield 评测集包括一个文档集合 D、一个信息需求集合 Q 以及一个标注了信息需求和文档相关性的关系集合 $\{R(q,d) \mid q \in Q, d \in D\}$。一般来说,每一个信息需求可对应一个或者多个相关的查询词语,而表示文档与信息需求相关性关系的集合通常是由人工完成的。为了评价一个搜索引擎,首先它要对 D 中的文档进行预处理并建立索引。对

于 Q 中的每一个信息需求 q，该搜索引擎在文档集合 D 上进行搜索，获得一个结果文档序列 $L^{(q)}$。通过一个评测指标 S 对搜索获得的文档排序结果 $L^{(q)}$ 和相关性标注集合进行对比，计算获得该系统对于该信息需求的一个得分 $S(q,L^{(q)})$。搜索引擎对于信息需求集合 Q 中的所有信息需求的得分，构成了这个系统的搜索结果质量得分的一个样本。不同搜索系统结果分数样本一般可以通过显著性测试[13-15]进行比较，从而获得可靠的系统质量比较结果。使用这种评测范式的时候，如果评测集合和指标选择恰当的话，这种在实验室环境中获得的分数能够反映系统的真实表现效果的好坏[16-18]。

Cranfield 评测方法包含两个重要的组成部分：一个静态的评测集合（文档集合 D、信息需求集合 Q 和相关性标注集合 $\{R(q,d) \mid q \in Q, d \in D\}$）和一些用于打分的指标。

常见的评测集合包括 TREC 组织构建的评测集[11]、亚洲 NTCIR（National Text Collections for Information Retrieval）组织构建的评测集[19]、欧洲 CLEF（Cross Language Evaluation Forum）组织构建的跨语言检索评测集[20]和中国中文 Web 信息检索论坛构建的评测集[21]等。这种评测集往往是可重用的。一旦被构建，如果有新的搜索引擎算法，就可以使用已有的查询集在文档集上做搜索，再依赖已有的相关性标注，来评测新的搜索系统。

不同的打分指标通常具有不同的表达能力和适用范围[11,22-25]。由于评测集通常是静态的，并且对于每个搜索系统，评分指标都可以获得一个绝对的分数，不同的搜索系统的优劣就可以通过这些分数数值来比较了。因此，可以用这种方法方便地评测任何改进搜索算法的新想法，而不用重新构造新的测试集来进行比较。Armstrong 等人[26]集成以往 TREC 中参加评测的组织以及公开发表论文中各种检索系统在不同评测集上的结果，创建网站 www.evaluatIR.org，以供所有信息检索领域的研究者研究和比较。

Cranfield 评测范式已经有较长的发展历史，源于传统信息检索系统评测的需求，远早于万维网的发展。在 20 世纪 60 年代，英国 Cranfield 学院的图书馆学家 Cyril Cleverdon 首先在他的两个信息检索项目中采用了这种评测方法[27]。在他的第二个项目中，他采用了一个包含文档集、信息需求集和完整的相关性集合的测试集对几种检索方式进行了评测。20 世纪 70 年代，Karen Jones[28]提出了一个理想的评测集所应有的特征，但这样的“理想评测集”一直到 1992 年第一届信息检索评测会议召开的时候才开始真正地建立。该会议是一个年度性的会议，它采用 Cranfield 评测范式建立针对各种搜索任务的评测集，然后对各种搜索系统进行评测。自从信息检索评测会议（TREC）举办以来，信息检索领域进入了一个高速发展阶段。在万维网诞生后，搜索引擎作为一种特殊但很快取得支配地位的信息检索系统出现，对其技术的评测在信息检索评测会议上得到了极大的重视。

在 1.4 节和 1.5 节，将分别介绍 Cranfield 评测范式的两个重要组成部分：评测指标和评测集。

随着互联网的发展,真实世界中文档集的规模越来越大。为了反映真实世界的这种变化趋势,评测集中的文档集合也越来越大,但是日益增大的文档集,导致很难有足够的人力资源来对每一个(查询,文档)对都给予一个相关性标注。因此,测试集的标注的不完整性程度变得越来越高。该问题成为 Cranfield 评测范式这类方法在近年来所遇到的最大的挑战,很多研究提出一些解决这个问题的策略,这些研究工作将在 1.6 节中介绍。

1.4 评测指标

本节介绍 Cranfield 评测范式中常用的一些评测指标,它们虽然大都产生于 Cranfield 范式的框架中,但其中许多也适用于其他的评测方法(例如本书重点介绍的用户隐反馈方法)。应该注意的是,由于搜索引擎与传统信息检索系统的应用环境有显著差别,以前提出的一些指标不一定很适合搜索引擎。这种认识成为本书第 5 章所体现工作的出发点,在那里,我们提出了一个新的搜索引擎评价指标:有效时间比(Effective Time Ratio,ETR)。

读者可以认识到,一个搜索引擎的好坏最终取决于总体用户体验,任何评测指标都是对用户体验的近似。指标的好坏对应这种近似程度的高低。由于搜索引擎应用环境的复杂,用户的需求多样且表达方式有限,很难讲某一指标就是最好的。常常可以说的只是某一指标更适合什么情形。这也是在信息检索领域不断有人提出新指标的原因之一。在理解指标局限性的同时,也应该看到指标的意义。从 20 世纪 90 年代中期万维网兴起开始,近 20 年过去了,最初的一些搜索引擎有些已经不存在,目前最受用户欢迎的搜索引擎也都是后来才出现的,实现了后来居上的目标。同时,现在也有一些新的搜索引擎实现了新的超越。广大互联网用户在搜索引擎之间的迁移是搜索引擎质量对比的最好风向标。然而,这种迁移是需要时间的,一个有雄心的搜索引擎不是要在投入市场后等待迁移,而是应该在投入市场前预测是否会发生所希望的用户迁移。按照一定的指标来对系统进行评估,就是进行上述预测的重要基础。

下面,首先介绍用于评测系统在单个信息需求上检索效果表现的指标。根据对相关性函数定义方式的不同,这些指标可以分成三类:二值相关性指标、多值相关性指标以及偏好类指标(这些指标的一个比较完整的列表和分类体系,也可以参考 Demartini 等人的工作[29])。其次,讨论把单个信息需求的分数在多个信息需求上综合的方法,并讨论如何根据两个系统各自的综合性分数比较它们检索效果的优劣。最后,讨论这些指标之间的关系。

1.4.1 二值相关性指标

二值相关性是搜索模型和评测研究常用的一个假设。它假设对于一个信息需求和一个文档要么是相关的,要么是不相关的,即 $R(q,d) \in \{0,1\}$(其中 0 代表不

相关,1 代表相关)。在这个假设下,搜索问题就大大简化了。同样的,相关性的判断也变得非常简单,对于一个信息需求,只需要判断一个文档是相关的或者是不相关的。早期的搜索引擎主要应用于图书馆的资料检索,对于一个查询,输出的结果是一个文档集的一个子集合 $D' \subset D$。搜索系统认为,在这个子集中的文档都是相关的。这类搜索系统称为分类式的搜索系统(即分为"相关"和"不相关"两类)。评价这种分类式的搜索系统,有两个基本指标:精度(Precision)和召回率(Recall)。现在的搜索引擎通常不是分类式的搜索系统,它的输出通常是一个文档序列,而不是一个文档集合。尽管如此,现有的很多评价搜索引擎指标还是从精度和召回率衍生出来的。

精度度量搜索系统排除不相关文档的能力,召回率度量搜索系统发现相关性文档的能力。准确的定义,精度(Prec)是搜索获得的相关文档个数和搜索获得文档总数的比值,而召回率(Rec)是搜索获得的相关文档个数和所有相关文档个数的比值,即

$$\mathrm{Prec}(q,D') = \frac{\sum_{d \in D'} R(q,d)}{\sum_{d \in D'} 1} \tag{1-1}$$

$$\mathrm{Rec}(q,D') = \frac{\sum_{d \in D'} R(q,d)}{\sum_{d \in D} R(q,d)} \tag{1-2}$$

由于相关性的判断是二值的,因此这里 $R(q,d)$ 的取值只能是 0 或者 1。

现代搜索系统返回的是一个结果文档序列(而不是一个结果文档集合),因此精度和召回率不能直接用于评价它的质量。一种变通的方法是,把一个排序式的搜索系统的结果解释为一个分类式搜索系统的结果,然后用精度和召回率进行评价。一种常用的转换方法是在结果文档序列的某一个位置 n 处做一个截断,把排序位置小于等于 n 的文档认为是搜索获得的结果文档集合,把排序位置大于 n 的文档认为是未被搜索获得的文档。因此,对于截断位置 n,就可以计算出相应的精度(Prec@n)和召回率(Rec@n)了。

对于一个结果文档序列的每一个截断位置 $n(n \geq 1)$,都可以得到一对召回率值和精度值(Rec,Prec)。如果把召回率作为横坐标,把精度作为纵坐标,就可以在这个坐标系画出所有截断位置 n 对应的点。对于一个文档序列,一个召回率的数值可能会对应多个序列位置,因此会对应多个精度值。通常,为了曲线的平整,对于一个召回率数值只显示对应的最大精度值,连接这些点构成的曲线被称为 R-P 曲线(Recall-Precision Curve)①。

① 在其他场合,例如在一些教科书上,R-P 曲线可能会描绘出所有相关的点,本书的处理方法旨在突出搜索结果体现出来的主要性态。

例 1-1 对于一个查询 q,一个系统返回的文档的相关性情况是 $L=(1,0,0,0,1)$,表示 $R(q,L_1)=1, R(q,L_2)=0,\cdots$。当召回率为 0.5 的时候,最大精度值是 1(对应第一个排序位置)。当召回率为 1 的时候,最大精度值是 0.4(对应第 5 个排序位置)。另外,当召回率为 0 的时候,默认精度是 1。因此,对应的 R-P 曲线如图 1-1 的蓝线部分所示。

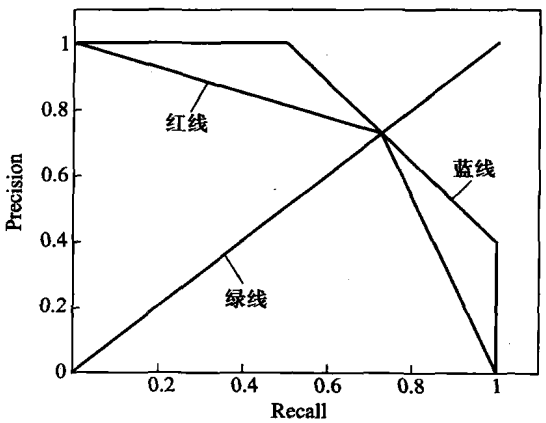


图 1-1 一个 R-P 曲线的例子

对于一个信息需求,一个结果文档序列对应的 R-P 曲线包含了它的搜索结果的完整的精度和召回率的信息。对于两个结果文档序列 L_1 和 L_2 来说,如果 L_1 对应的 R-P 曲线总在 L_2 对应的曲线之上,就表明,从精度和召回率这两方面考虑,L_1 都表现得更好。但是,实际上这种情况很少出现。在大多数情况下,根据两个序列获得的两条 R-P 曲线是相交的。也就是说,在一些截断位置范围内,一个结果序列更好,在另一些截断位置范围内,另一个序列表现得更好。这样的曲线虽然能够帮助我们更好地理解系统的具体表现,但是不便于直接比较两个系统的表现。为了直观地得到两个序列的好坏,需要把 R-P 曲线总结为一个分数,通过这个分数来比较两个系统。

对于 Web 搜索引擎这样的应用,用户只查看一些排列位置很高的文档,因此最前面的文档对用户而言是至关重要的。在这种情况下,在某个截断位置上的精度 Prec@n(n 一般取比较小的值,如 10)可以被直接用来评价系统。但是,由于 Prec@n 只能提供一个位置上的精度,因此它只能提供局部的信息,如例 1-2 所示。

例 1-2 对于查询 q,序列 A 对应的文档相关性是(0,0,0,0,1),而序列 B 对应的文档相关性是(1,0,0,0,0)。很明显,序列 B 的质量高于序列 A 的质量。但是对于指标 Prec@5,有 Prec@5(q,A) = Prec@5(q,B),因此它不能区分这两个序列的质量。

另外一个类似的指标是在某个位置上的召回率 Rec@n 也经常被使用,同样的,它也只能提供一个局部的召回率信息(对于 Web 搜索引擎而言,它很少被使

用。这是因为,用户不用找到全部的相关文档,只需找到能够满足信息需求的一些文档就足够了。而且在 Web 上,相关文档集合往往是未知的)。这两个指标通过两个不同的角度来评价一个搜索系统,并没有提出一个统一的指标。一种常用的做法是采用它们的调和平均值(F 指标)来综合两个分数。

$$F = \frac{Rec \cdot Prec}{Rec + Prec} \tag{1-3}$$

另外,Prec@n 和 Rec@n 这两个指标共同存在的一个问题是,它们都不能很好地被归一化,这一点将在分数综合部分详细讨论。

TREC-2 中采用了 R-prec 作为一个新的指标[10],综合 R-P 曲线上的精度和召回率信息。对于信息需求 q,假设它在整个文档集 D 上的相关文档数是 N_q,那么 R-prec 定义了在截断位置 N_q 上的精度,即

$$R\text{-prec}(q,L) = \text{Prec@}N_q(q,L), \quad \text{其中 } N_q = \sum_{d \in D} R(q,d) \tag{1-4}$$

同样,可以计算在截断位置 N_q 上的召回率

$$\text{Rec@}N_q(q,L) = \frac{\sum_{i=1}^{N_q} R(q,L_i)}{N_q} = \text{Prec@}N_q(q,L) \tag{1-5}$$

从式(1-5)可以看到,在截断位置 N_q 上,精度和召回率的数值是相同的。因此,它实际上是在(Prec,Rec)坐标上,Prec = Rec 这条直线和 R-P 曲线的交点对应的精度和(或)召回率[3],如图 1-1 中蓝线(R-P 曲线)和绿线(精度-召回率等值线)的交点所示。因此,这个数值同时包含了精度和召回率这两者共同的信息[1]。通过简单的推导可以获得,这个数值实际上还等于由(1,0)、(R-prec,R-prec)、(0,1)、(0,0)这四个点构成的一个四边形的面积,而这个四边形又可以看成是整个 R-P 曲线和两个坐标包成的图形的一个近似[22],如图 1-1 中的红线与坐标轴围成的图形所示。所以,该指标包含了 R-P 曲线的综合信息。

另一种总结 R-P 曲线信息的方法是对 R-P 曲线上面的精度做平均。一种平均的方式是对召回率为 0,0.1,…,1 这 11 个点对应的精度做平均。这种方法的一个缺陷是可能并不存在一个截断位置,对应其中的一种召回率数值。

一种更简单的方式是对每一个相关文档的位置,获得一个精度,然后得到这些精度的平均值,这个对应的指标称为平均精度(Average Precision,AP)。形式化的 AP 可以定义为

$$AP(q,L) = \frac{\sum_i R(q,L_i)\,\text{Prec@}\,i(q,L)}{\sum_i R(q,L_i)} \tag{1-6}$$

从几何上来看,平均精度也是一个 R-P 曲线围成的形状的面积的一个近似,

而且它的近似比 R-prec 得到的近似更加准确。除了几何意义以外，平均精度也可以从用户收益的角度来解释[30]。假设用户的收益就是看到文档中相关文档的比例，而用户以一个均匀分布在某一个相关文档结束搜索，那么平均精度就度量了用户的收益的期望。

有一类特殊的信息需求，它们对应的查询被称为有目标的查询（Known Item Query）。提交这类查询的用户，只需要阅读一个相关文档，就能满足这种信息需求。譬如说，用户想要找一个网站的首页，或者找到一个以前访问过的页面，或者想知道一个简单的事实性问题的答案。在这种情况下，系统的搜索效果取决于第一个相关文档的位置。常用的指标是第一个相关文档位置的倒数（Reciprocal Rank，RR），即

$$\mathrm{RR}(q,L)=\frac{1}{\min(\{L(d)\mid R(q,d)=1\})} \tag{1-7}$$

第一个相关文档的排序越高，那么系统在这个信息需求上的得分也就越高。

1.4.2 多值相关性指标

二值相关性指标基于二值相关性假设，这个假设非常强，因此和实际情况会有所不符。在实际情况中，并不是所有的相关文档的有用程度都是完全一样的，有的文档通篇和信息需求都非常相关，而有的文档只有部分内容和信息需求相关。对于用户来说，看到一篇高度相关的文档会比看到一篇部分相关的文档获益更多。为了体现文档不同相关程度的差异，可以使用多值的相关性函数 $R(q,d)\in\{v_1,\cdots,v_K\}$（这里相关性程度可以取 K 个数值中的一个）。基于多值相关性函数的指标就是多值相关性指标。

其中，最为常用的多值相关性指标是折扣累计收益（Discounted Cumulated Gain，DCG）[31]。这个指标基于以下两个假设。

假设 1：高度相关的文档比部分相关的文档对用户更加有用。

假设 2：一个文档被搜索算法排得越高，则它应该对用户更有用。

基于这两个假设，该指标定义了用户从一个文档排序中的收益是他从这个排序中所有文档获益的累加，而一个用户从一个文档的收益是由这个文档的相关性程度和排序位置决定的。它可以形式化地定义为

$$\mathrm{DCG}(q,L)=\sum_i Dis(i)\cdot Gain(R((q,L_i))) \tag{1-8}$$

其中，$Gain$ 是一个收益函数（Gain Function），一般它是随着文档的相关程度单调递增的，即文档越相关，给用户带来的收益越多；$Dis(i)$ 是对位置 i 的一个折扣函数（Discount Function），一般它是随着文档排序单调递减的，即文档排序越靠后，它能给用户带来收益的机会就越小。由于收益函数和折扣函数的选取，它可以有各种变种。Kanoulas 等人[32]通过进行方差成分分析的方法[33]，在多个数据集合上进行