目 录

第一章 引言 1
1.1 研究目的、意义及研究背景 1
1.1.1 信息检索的基本概念与发展历史 2
1.1.2 信息检索系统的运行原理 4
1.1.3 信息检索系统的性能评价 6
1.2 网络信息检索的基本概念 7
1.2.1 网络信息检索系统的组成 8
1.2.2 超链接结构分析的形式化基础 12
1.2.3 HITS 算法基本思路及实现 13
1.2.4 PageRank 算法基本思路及实现 16
1.2.5 链接分析算法总述 17
1.2.6 网络信息检索系统的性能评价 18
1.3 面向网络信息检索的用户行为分析概述 21
1.3.1 用户群体行为分析的研究历史 21
1.3.2 用户查询行为的宏观统计分析 22
1.3.3 用户查询信息需求分析 23
1.3.4 针对特定检索用户群体的行为分析 24
1.4 课题研究重点及主要工作 24
1.4.1 用户交互方式与网络信息检索系统的发展 24
1.4.2 网络信息检索系统的信息处理困境 26
1.4.3 本文研究重点：基于用户群体行为分析的网络信息检索 27
1.5 论文各部分的主要内容 28
第二章 基于用户群体行为分析的网页质量评估算法研究 30
2.1 引言 30
2.2 相关工作 31
2.2.1 宏观粒度网络数据质量评估技术 31
2.2.2 微观粒度网络数据质量评估技术 33

2.2.3 冗余页面识别技术 35
2.2.4 网络数据质量评估方式总述 36
2.3 用户信息需求与网页质量的定义 37
2.4 查询目标页面与普通页面的特征差异 39
2.4.1 特征分析的实验设定 39
2.4.2 文档长度特征的差异 41
2.4.3 链接文本长度的差异 42
2.4.4 PageRank 特征的差异 43
2.4.5 入链接个数特征的差异 44
2.4.6 出链接个数特征的差异 45
2.4.7 站点内出链接个数的差异 46
2.4.8 URL 长度特征的差异 47
2.4.9 查询目标页面的查询无关特征总结 48
2.5 基于查询目标页面定位的网页质量评估 49
2.5.1 查询目标页面定位问题的困境 50
2.5.2 基于分类方法的网页质量评估及其局限性 51
2.5.3 基于贝叶斯学习方法的网页质量评估 54
2.6 网页质量评估性能实验 56
2.6.1 性能评价指标 56
2.6.2 实验设定 59
2.6.3 网页质量评估实验 61
2.6.3 页面质量评估算法对检索性能的影响 64
2.7 本章小结 68
第三章 基于用户群体行为分析的用户查询行为研究 70
3.1 引言 70
3.2 网络信息检索用户行为特性 71
3.2.1 用户用多少词来表示查询需求 72
3.2.2 用户是否修改他们的查询 72
3.2.3 用户行为特性对查询处理的影响 73
3.3 研究背景：查询分类在网络信息检索中的作用 74
3.4 用户查询需求的特征分析 75

3.4.1 查询内容特征 76
3.4.2 查询反馈特征 78
3.4.3 链接文本特征 78
3.4.4 查询历史特征 80
3.5 面向查询分类的用户行为特征提取 82
3.5.1 前 N 次点击满足用户需求率 82
3.5.2 前 N 位结果满足用户需求率 84
3.6 基于用户群体行为分析的查询分类算法 85
3.6.1 选择决策树算法进行查询分类的原因 85
3.6.2 查询信息需求分类算法设计 86
3.7 实验结果与讨论 88
3.8 本章小结 90
第四章 基于用户群体行为分析的 搜索引擎自动性能评价研究 92
4.1 引言 92
4.2 相关工作 93
4.2.1 基于 Cranfield 方法的搜索引擎自动性能评价 93
4.2.2 搜索引擎自动性能评价的其他尝试 94
4.2.3 搜索引擎自动评价已有工作总述 95
4.3 查询信息需求与自动性能评价 96
4.4 导航类自动性能评价算法设计 98
4.4.1 查询样例集合的自动生成 99
4.4.2 标准答案集合的自动标注 100
4.5 实验结果与讨论 102
4.5.1 答案自动标注实验结果 102
4.5.2 性能评价实验结果 103
4.5.3 自动性能评价的实验结果讨论 104
4.6 本章小结 105
第五章 网络信息检索系统架构设计的改进 107
5.1 引言 107
5.2 基于网页质量评估的索引模块设计 108

5.3 基于用户需求分类的查询处理模块设计 110
5.4 改进的网络信息检索系统架构 111
5.5 本章小结 114
第六章 结论 115
参考文献 119
致谢与声明 128
个人简历、在学期间发表的学术论文与研究成果 129

第一章 引言

1.1 研究目的、意义及研究背景

在知识经济与信息化蓬勃发展的当今时代，信息与知识的占有和利用程度，越来越成为衡量一个国家和民族进步和发展水平的重要标准。上个世纪最后二十年以来因特网的发展和普及为人们提供了一个全新的信息存储、加工、传递和使用的载体，网络信息也迅速成为了社会成员获取知识和信息的主要渠道之一。1994 年，面向个人用户的浏览服务才诞生，但根据加州大学伯克利分校的统计（Lyman，2003），2002 年为止万维网上存储的信息量就已经超过了 500,000 Terabyte。如此迅捷的增长速度是人类历史上任何一种媒体的发展过程中不曾有过的，而如此规模的信息资源在将人类占有的几乎所有知识纳入其中的同时，也给资源的使用者带来了如何充分开发与利用的问题。

为了解决网络信息资源利用的问题，将传统的信息检索技术与网络应用实际相结合的网络信息检索应运而生。网络信息检索的目的，是快速、准确和最大程度地从互联网上的海量信息资源中定位并获取高质量的相关信息。网络信息检索技术及其实现产品——网络搜索引擎的发展，已经为人们访问网络信息资源提供了巨大的便利。第 16 次中国互联网络发展状况统计报告指出，中国搜索引擎用户已占互联网用户的 95.2%，绝对用户数近 1 亿人，包括搜索引擎在内的网络信息检索工具已经成为当今获取信息的主要手段之一（CNNIC，2005）。

尽管商业搜索引擎的技术发展和商业运行模式已经取得了相当的成功，但不可否认的是，自上世纪末以 Google 为代表的一系列优秀检索工具出现以来，网络信息检索技术并没有出现用户可以感受到的质的性能提高，网络信息检索技术的发展现状还远远无法满足人们快捷高效获取网络信息的要求。网络搜索引擎返回的查询结果有时依然数目浩繁，让人无从着手；虚假信息、过时信息和垃圾信息也很大程度上影响了人们依靠搜索引擎获取信息的有效性。高效利用搜索引擎获取网络信息，一定程度上还是依靠查询者的技巧及其获取互联网信息的经验。近年来中文搜索引擎市场的激烈竞争和关于“第三代搜索”概念的争论也正反映出了互联网用户对当前搜索引擎检索性能的更高需求。

阻碍网络信息检索研究和发展水平进一步提高的因素有许多，如 Web 数据环境的繁杂特性和一日千里的发展速度、互联网用户庞大与复杂的信息需求等都为网络信息检索工具的设计和实现带来了巨大的困难。但最根本的因素，还在于搜索引擎现行采用的“关键词查询+选择性浏览”的信息交互方式。这种快捷、直接的信息获取方式使得搜索工具以极快的速度普及，成为人们访问互联网活动中的必需。但它所带来的用户信息需求传递的瓶颈问题却也成为了阻碍网络信息检索工具发展的最大困难之一。利用对检索用户行为的深入挖掘与分析，对互联网的数据环境和用户查询需求进行重新认识，从而使网络信息检索工具能够更加有效的处理检索需求，是本论文研究关注的重点。

本论文的研究课题“基于用户群体行为分析的网络信息检索研究”在本世纪逐渐得到国内外研究人员的关注，是一个较新的研究方向，也具有广泛的应用前景和现实意义。深入的探索和研究，有可能促进整个网络信息检索领域的发展，也有可能有效提高互联网络信息环境的使用效率，使得互联网信息更好的为人们的生产和生活需要服务。

1.1.1 信息检索的基本概念与发展历史

1945 年，在二次世界大战即将胜利之际，曼哈顿工程和美国自然科学基金的创立人 Vannevar Bush 提出了这样一个想法：在 2010 年左右，世界上应该有一种工具，它能够使人们最方便快捷的获取所有图书馆中藏有的知识，这个构想对应的就是信息检索工具的雏形。在 Bush 的支持下，自 1950 年起美国政府就开始了对信息检索相关研究的支持，而随着计算机技术的发展和迅速普及，信息检索作为应对信息爆炸问题的主要手段而迅速发展起来，其研究领域也由最初的科学技术领域扩展到人类活动的各个方面。尚未走到二十一世纪第二个十年的我们，已经能够很大程度上享用信息检索技术带给我们的快捷与方便：数字图书馆、互联网搜索引擎、遍布各行各业的全文搜索系统等等已经成为我们生活中不可或缺的一部分，大大提高了我们获取信息的效率和便捷程度。

1951 年，Calvin Mooers 首次提出了“信息检索（Information Retrieval, IR）”这一概念，给出了信息检索的主要任务：协助信息的潜在用户将信息需求转换为一张文献来源信息列表，而这些文献包含有对其有用的信息。信息检索涵盖了对信息的描述以及为了查询而对信息进行的规范化工作，还涵盖了用于进行检索操作的任何系统、技术和机器。

我们可以从不同角度对信息检索的概念进行认识：从软件设计的角度，信息检索工具可以被认为是一个输入信息需求，输出检索答案列表的特殊软件；从用户交互的角度，信息检索的过程可以被认为是一个信息用户与对应信息源的作者之间的交互通讯过程；从人工智能的角度，信息检索工具则是一个处理以物理形式存在的知识记录为主要目的的专家系统。概括地讲，人们一般把信息检索定义为从文档集合中返回满足用户需求的相关信息的过程，它是研究信息的获取(acquisition)、表示(representation)、存储(storage)、组织(organization)和访问(access)的一个学科，当前，信息检索也通常被归类于计算机学科与图书情报学科交叉学科的范畴。

信息检索有着漫长的发展历史，约 4000 年前，人类就开始有目的地组织信息，一个典型的例子就是图书中的目录。1845 年，Cowden Clarke 夫人耗费 16 年时间对莎士比亚的 5 部戏剧作品进行了手工索引，可以被认为是建立检索系统的先驱。然而，信息检索学科真正取得长足发展则是在计算机诞生并得到广泛应用之后，20 世纪六、七十年代，人们开始使用计算机为一些小规模科技和商业文献的摘要建立文本检索系统；1980 年前后则诞生了大量的商用大规模数据检索系统，如 Dialog, MEDLINE 等。但总体而言，这时的信息检索系统的应用还局限在极少数领域，其核心驱动力也是集中在提高科学研究的效率，与普通用户的应用相距甚远。

互联网的出现和计算机硬件水平的提高使得人们存储和处理信息的能力得到巨大的提高，从而加速了信息检索研究的进步，并使其研究对象从图书资料和商用数据扩展到人们生活的方方面面。伴随互联网爆炸性的发展，普通网络用户想找到所需的资料简直如同大海捞针，这时为满足大众信息检索需求的专业搜索网站便应运而生了。现代意义上的搜索引擎的祖先，是 1990 年由蒙特利尔大学学生 Alan Emtage 发明的 Archie。虽然当时 World Wide Web 还未出现，但网络中文件传输还是相当频繁的，而且由于大量的文件散布在各个分散的 FTP 主机中，查询起来非常不便，因此 Archie 应运而生。随着 WWW 的出现，搜索网站的查找对象也从单纯的文件扩展到网页，最早现代意义上的搜索引擎出现于 1994 年 7 月，Michael Mauldin 首次将网络爬虫程序与文本索引程序相结合，创建了现在仍在提供服务的 Lycos 搜索引擎（http://www.lycos.com/）。1995 年，斯坦福大学的两名博士生，David Filo 和杨致远共同创办了基于目录索引结构的

Yahoo!搜索引擎，并成功地使网络搜索概念深入人心，从此搜索引擎进入了高速发展时期。

中文网络信息检索工具的发展历史可以追溯到 1997 年，当时北大网络实验室推出了在教育网内的 Web 信息导航服务即后来的天网搜索引擎。2001 年，百度搜索面世并开始逐渐成为中文搜索引擎市场的领头羊。从 2003 年开始，中文网络信息服务的四大门户网站（新浪、搜狐、网易、腾讯）陆续推出了自己的搜索引擎服务。目前，以搜索引擎为代表的网络信息检索工具仍在迅猛发展的过程中。

1.1.2 信息检索系统的运行原理

信息检索系统（如果没有特别说明，本文中的“信息检索系统”指狭义的信息检索系统，即文本信息检索系统）的基本架构和运行示意如下图所示：

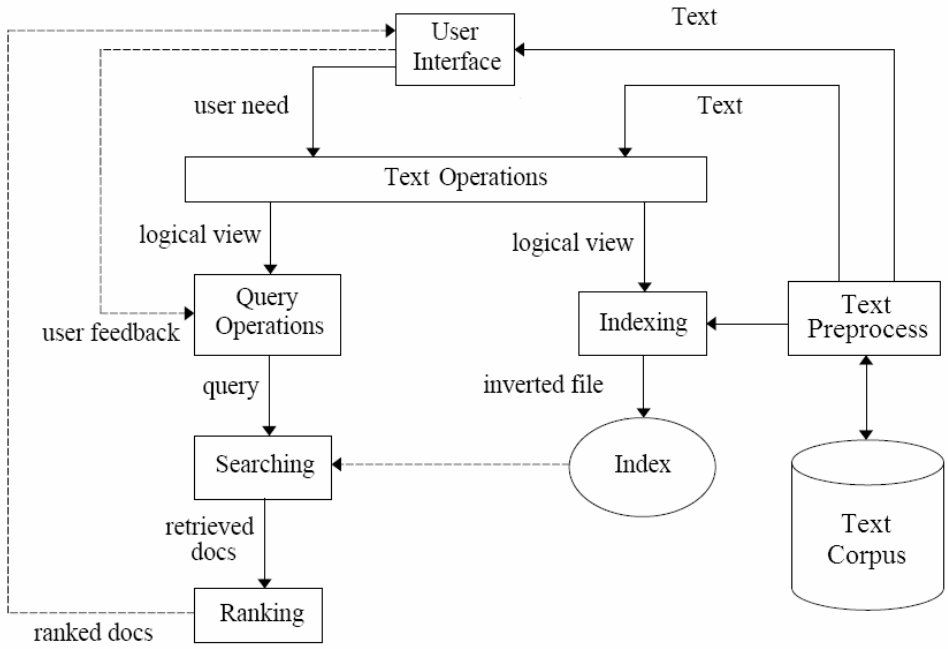

图 1-1 信息检索系统架构及运行原理示意图，修改自 Baeza-Yates et al（1999）

如图 1-1 所示，传统意义的文本信息检索系统是由文本处理、内容索引、查询处理、用户界面等模块组成的。

从数据处理对象的角度分析，信息检索系统面对的处理对象则包括文本语料库与检索用户两方面的内容。具体的运行流程上，系统将语料库中的文本经过取词根（英文）、分词（中文）、去除停用词等操作后输入进索引模块，索

引模块以词项（term）为中心组织倒排索引（inverted index），从而完成系统准备工作。用户进行检索时，首先通过用户界面输入其查询需求（query），系统通过查询模块在倒排索引中定位相关文档集合，再依据相关文档与查询需求的相似度对相关文档进行排序（ranking），并通过用户界面反馈给用户检索结果。

检索系统中，使用何种模型进行查询需求与文档的相似度计算是关键性的研究问题，信息检索研究中的模型研究，大多是针对这类相似度计算模型而进行的。从 20 世纪 60 年代中期以来，人们提出了大量的相似度评价模型。从最初的为一些较小的和较为结构化的文档所设计的特殊模型（如文献记录，包括题目、作者和主题码等），发展到现在具有较强理论基础和能处理多种文档格式的模型。总的来说，当前应用比较广泛的检索模型有如下三种（Spoerri，1995）：

1．布尔模型(Boolean Model)。布尔模型又称精确匹配模型，是许多商业信息检索系统尤其是早期联机信息检索系统的理论基础。在布尔模型中，文档和查询都被表示为索引项的集合。其优势在于：实现简单，计算复杂度低，支持结构化查询等，因此在计算机硬件水平较低的早期信息检索系统应用中得到了青睐。目前，不少图书馆信息检索系统还是基于布尔模型进行搭建的。

2．统计模型(Statistical Model)，统计模型基于文本统计信息（如词项频度统计、文档频度统计等）和统计自然语言处理方法进行文本相似度计算。统计模型从布尔模型的“精确匹配”发展到“最优匹配”，可以说真正把信息检索与传统的数据搜索区别开来。而具体来说，其又包括向量空间模型(Vector Space Model)和概率模型(Probabilistic Model)两类。向量空间模型把文档和查询表示成一个高维空间中的向量，用距离作为相似度的度量。而概率模型则把检索看作是文档表示和查询之间匹配成功的概率估计问题，使用概率理论做指导。统计模型是当前应用最广泛的检索模型，在从小规模纯文本信息检索系统到搜索引擎在内的各种检索系统中广泛应用。

3. 语言与知识模型(Linguistic and Knowledge-based Model)。与统计模型不同，基于语言和知识的检索模型试图考察查询背后的用户真实需求，从而更加准确的理解用户意图，获得更好的检索效果。目前较常用的做法是通过考察用户查询和相关文档之间的语言模型的相似程度，来进行文档相关度的计算，即计算按照文档的语言模型有多大可能性生成用户查询。语言和知识模型在本世纪初以来得到了较为广泛的应用，目前信息检索研究领域应用的较为广泛的 Lemur/Indri 系统，就是基于此模型进行内容相似度计算的。

以上我们对信息检索系统的整体运行流程和其中的相似度计算模型进行了简单介绍，总体而言，信息检索系统的架构设计自上个世纪五十年代信息检索问世以来没有经历过大的改动，而直到今天，这个架构即使在处理海量规模网络文本数据的问题上也是能够满足应用需求的。当前信息检索研究的重点，一方面在于利用自然语言处理方面的研究成果对检索相似度模型框架进行修正；另一方面则在于对检索系统的应用对象加以拓展，针对专门检索系统如网络信息检索系统、多媒体检索系统等开展研究和分析。我们的工作重点集中在第二方面，即利用用户群体行为分析方法对网络信息检索的处理流程和系统架构进行改进，因而对于传统检索系统的技术细节不再赘述。

1.1.3 信息检索系统的性能评价

信息检索系统的评价是信息检索研究中的核心课题之一，客观可靠的检索效果评价方式是任何一项信息检索研究中均需要考虑和设计的内容。目前采用较多的评价是基于 Cleverdon et al（1966）提出的 Cranfield 评价框架（A Cranfield-like approach）。Cranfield 框架涉及到的组成部分包括标准评测语料库、查询样例集合、标准答案集合和相应的评价指标等。对于传统文本信息检索系统的评价而言，由于语料库、查询和答案集合都来自于手工标注人员的整理，因此检索系统的评价指标是其关注的核心。

传统文本信息检索系统的检索性能评价一般采用精确率—召回率的评价指标，这一评价体系自 Kent et al（1955）提出精确率和召回率的概念之后，一直应用在信息检索系统的评价研究中。精确率（也称查准率, Precision）和召回率（也称查全率, Recall）是其最基本的效果评价指标。其具体的定义为：

$$\begin{aligned} \text{精确率} &= \frac{\#(\text{检索系统返回的结果集合} \cap \text{标准答案集合})}{\#(\text{检索系统返回的结果集合})} \\ \text{召回率} &= \frac{\#(\text{检索系统返回的结果集合} \cap \text{标准答案集合})}{\#(\text{标准答案集合})} \end{aligned} \tag{1-1}$$

由此可见，精确率是对“系统返回的结果中有多少比例是正确的”进行评价；而召回率则是从“系统把正确结果中的多少返回给了用户”的角度进行评价。精确率和召回率从不同的方面共同对系统的检索性能给予评价，因此在评价过程中单独评价某个方面的数值结果意义不大。

大量的检索评价实验结果表明，精确率与召回率之间存在着特定的关系：在一个信息检索系统中，当精确率和召回率达到一定程度以后，两者就会呈现出非线性的反变关系。即在精确率提高的同时，召回率会持续下降；反之，在召回率提高的同时，精确率也会持续下降。值得引起注意的是，只有当精确率和召回率达到一定程度以后，两者之间才会呈现出这样反变关系。如果精确率和召回率都很低，那么两者完全可以同时得到提高。精确率与召回率之间的这种反变关系，从评价的角度反映了信息检索系统检索策略选择的困难：把更多的结果作为相关结果返回给用户，自然会提升结果召回率，但不正确的结果也会增多，从而降低精确率；反之，只把系统选择的最精确的结果返回给用户，则系统的精确率会得到提升，但不少实际相关的结果就无法返回，从而降低召回率。这也说明，精确率与召回率在评价过程中必须统一考虑，才能对系统的性能给出全面的评价。

将精确率与召回率共同加以评价的方式有很多，最常用的方式包括 11 点平均精度与 F-评价（F-measure）等。

把召回率分为 0%、10%、20%、…、100%这 11 个等级，分别计算它们对应的精确率，无法直接计算的点则可以用插值法等方法来确定。这样绘制出来的曲线可以直观的反映对一个查询的检索效果。当我们需要衡量检索算法在检索多个不同的查询时总的检索性能时，则可以对所有查询，在同一个召回率等级上对各个精确率数值取平均，而 11 点平均精度就是对 11 个召回率等级上对应的精确率取平均。这是目前较常用的综合评价召回率与精确率的方法之一。

除了 11 点平均精度以外，F-评价（F-measure）也是一个对检索精度和召回率综合考察的指标。它的定义是：

$$F-measure=\frac{2}{1/precision+1/recall} \tag{1-2}$$

其中 Precision 和 Recall 分别为检索系统返回的结果文档集合对应的的精确率和召回率。

1.2 网络信息检索的基本概念

上个世纪最后二十年以来，Internet 的发展和普及为人们提供了一个全新的信息存储、加工、传递和使用的载体，Web 信息迅速成为了社会成员获取知识

和信息的主要渠道之一，于是将传统的信息检索技术与网络应用实际相结合的Web 信息检索应运而生。Web 信息检索技术的发展，已经为人们访问网络信息资源提供了巨大的便利如拥有最大规模网页索引量的 Google 系统（ http://www.google.com/ ），IBM 公司的 Clever 计划（http://www.almaden.ibm.com/cs/k53/clever.html）等都是其中检索效果和性能比较突出的代表。

随着网络普及程度的提高，网络搜索引擎在我国互联网用户的网络使用中也起到了举足轻重的作用，2005 年 7 月公布的中国互联网络发展状况统计报告指出，当前中国搜索引擎用户已占互联网用户的 95.2%，绝对用户数超过 9500 万人（CNNIC，2005）。包括搜索引擎在内的网络信息检索工具已经成为网络用户获取信息的主要手段，84.5%的用户指出搜索引擎是他们得知新网站的主要途径，而 58.2%的用户认为搜索引擎是他们最主要的获取网络信息的途径。

中文互联网搜索市场日趋激烈的竞争也从另一个侧面反映了搜索引擎的用户吸引力和市场价值。2004 年之前百度（http://www.baidu.com/）基本涵盖了中文互联网全部的搜索流量，但从 2005 年开始，传统意义上的四大中文门户网站（新浪、搜狐、网易、腾讯）均陆续推出了自己的搜索引擎服务，而 Google，Yahoo!，MSN Search 等跨国品牌也日渐重视中文搜索，分别推出自己全新的中文搜索服务。网络搜索引擎技术和市场发展的巨大需求使我们的相关研究能够较迅速的转化为实际应用产品，但同时也对研究成果的可靠性和可行性提出了更高的要求。

1.2.1 网络信息检索系统的组成

网络信息检索系统的组成与第 1.1 节中介绍的一般信息检索系统有相似之处，其核心的文档处理与查询处理过程与传统信息检索系统的运行原理基本类似，但其所处理的数据对象即网络数据的繁杂特性决定了网络信息检索系统必须进行系统结构的调整，以适应处理数据和用户查询的需要。

图 1-2 所示是一个典型的网络信息检索系统的系统架构图示，总体而言，网络信息检索系统由信息收集、信息处理和查询服务三个模块组成。

从具体运行方式上说，系统根据站点/网页的 URL 信息和网页之间的链接关系，利用网络蜘蛛在互联网上收集数据；收集的数据分别通过链接信息分析器和文本信息分析器处理，保存在链接数据库和文本索引数据库中，同时，网页

质量评估器依据网页的链接关系和页面结构特征对页面质量进行评估，并将评估的结果保存在索引数据库中；查询服务器负责与用户的交互，它根据用户的检索需求，从索引数据库中读取对应的索引，并综合考虑查询相关性与页面质量评估结果之间的关系，给出查询结果列表反馈给用户。

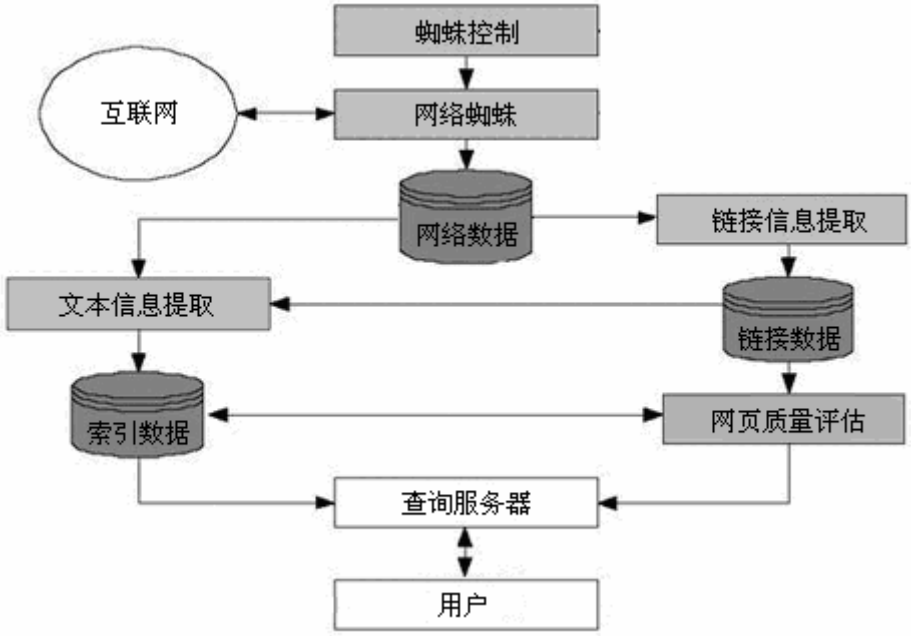

图 1-2 网络信息检索系统架构示意图

下面将逐一介绍网络信息检索系统各组成模块的主要运行方式：

1．信息收集模块

信息收集模块包括图 1-2 中的“蜘蛛控制”和“网络蜘蛛”两部分，“蜘蛛”这个称呼形象的描述出了信息收集模块在网络数据形成的“Web”上进行信息获取的功能。总体而言，网络蜘蛛从种子网页出发，通过反复下载网页并从文档中寻找未曾见过的 URL，达到访问其他网页得以遍历 Web 的目的。而其工作策略一般则可以分为累积式抓取（cumulative crawling）和增量式抓取（incremental crawling）两种。

累积式抓取是指从某一个时间点开始，通过遍历的方式抓取系统所能允许存储和处理的所有网页。在理想的软硬件环境下，经过足够的运行时间，累积式抓取的策略可以保证抓取到相当规模的网页集合。但由于 Web 数据的动态特

性，集合中网页的被抓取时间点是不同的，页面被更新的情况也不同，因此累积式抓取到的网页集合事实上并无法与真实环境中的网络数据保持一致。

与累积式抓取不同，增量式抓取是指在具有一定量规模的网络页面集合的基础上，采用更新数据的方式选取已有集合中的过时网页进行抓取，以保证所抓取到的数据与真实网络数据足够接近。进行增量式抓取的前提是，系统已经抓取了足够数量的网络页面，并具有这些页面被抓取的时间信息。

面向实际应用环境的网络蜘蛛设计如 Brin et al（1998）和孟涛 等（2005），通常既包括累积式抓取，也包括增量式抓取的策略。累积式抓取一般用于数据集合的整体建立或大规模更新阶段；而增量式抓取则主要针对数据集合的日常维护与即时更新。

在确定了抓取策略之后，如何从充分利用网络带宽，合理确定网页数据更新的时间点就成了网络蜘蛛运行策略中的核心问题。针对这个问题，Brewington et al（2000）考察了网页的变化规律并推算出了保持一定的时新性需要的最低网络带宽。Cho et al（2000a）将网页的生命周期视为一个泊松过程，并根据网页历史更新数据计算泊松过程参数，借以估计可能的下次更新时间；Cho et al（2000a）和 Cho et al（2003）则建立了网页的“时鲜性”（即保持搜索引擎网络数据与实际网络数据的一致性）的概念，并对如何评估和保持网页的时鲜性提出了解决方案。

总体而言，在合理利用软硬件资源进行针对网络数据的即时抓取方面，已经形成了相对比较成熟的技术和实用性解决方案，这方面目前所需解决的主要问题，是如何更好的处理动态网络数据问题（如数量越来越庞大的 Web2.0 数据等），以及更好的根据网页质量修正抓取策略的问题。

2．查询服务模块

查询服务模块是指网络信息检索系统中，直接与用户查询需求交互的组成部分，具体到图 1-2 中，主要是包括“查询服务器”部分。从功能上讲，这个模块主要完成两方面的任务：首先是用户交互，即收集用户查询并用符合用户行为习惯的方式对查询结果进行反馈；其次是信息查询和检索，使用收集到的用户查询再由信息收集模块抓取，并由信息处理模块建立索引的网络数据上进行检索。这其中涉及到的核心技术包括：检索系统网络服务的建立和维护、面向网络信息的检索模型建立等。

网络信息检索系统严格意义上说也在提供一种网络服务（Web Service），因此它的建立和运行、维护机制符合网络服务系统的一般规律。但由于网络信息检索用户行为习惯存在宏观上的统计特性，系统的网络服务也应进行优化以更好的适应这种行为习惯。Silverstein et al（1998）和余慧佳 等（2007）的工作都说明，尽管网络信息检索用户所提出的查询数目异常繁杂（根据 Sullivan（2006），Google 每天需要处理超过 1 亿次的用户需求），但搜索需求中的大多数是重复的，重复的比例超过用户所提交查询需求的 70%以上乃至更多。因此在检索系统的网络服务部分加入某种类型的缓存机制就是非常必要的，对检索结果序列或检索结果页面本身的缓存都有助于大大减轻网络数据检索的负担。

面向网络信息的检索模型研究一直是网络信息相关研究的重点，研究的主要方法一般集中在以下两点：改进传统文本信息检索技术以适应网络信息需求；不对传统检索技术进行改动，而尝试对传统检索结果与网络信息质量评估结果进行融合。前者主要的研究工作包括对网络半结构化文本检索方式的研究和对链接文本检索方式的研究（张敏 等，2005）（Craswell et al，2001）；而后一方面主要的研究工作则集中在将网络链接关系分析算法（如 PageRank）的结果融合到传统文本检索结果模型的尝试（Craswell et al，2005）。而这方面最重要的研究平台之一是 TREC 的网络信息检索（Web Track）和超大规模文本检索（Terabyte Track）论坛。上述不少研究所基于的实验数据都来自于这两个论坛从 2001 年以来的积累，我们研究中涉及到的一部分实验结果也是基于这个研究领域比较公认的平台得出的。

总体而言，面向用户的查询服务技术是网络信息检索技术研究的热点之一，而主要的研究方向，则是基于传统文本信息检索技术，融入网络信息检索的应用需求元素，借以建立新的检索模型或对原有模型进行改进。当前，这方面研究集中在对用户信息需求的挖掘与发现、改进网络信息检索系统的用户交互方式等方面上。

3．信息处理模块

图 1-2 中，除去上两个模块之外的部分都可以归入信息处理模块的范畴，将纷繁复杂的网页数据整理成可以被检索系统高效、可靠、方便使用的格式是这一模块的工作。具体来讲，信息处理包括对传统内容信息的处理和对非内容信息（主要是链接结构信息和文本结构信息等）进行处理两个方面的内容。

对文本内容信息的处理与传统文本信息检索系统的信息处理模块功能类似，最主要的目的是建立以词项（term）为中心的文本倒排索引，以便检索系统中检索功能模块进行内容相似度计算使用，提高系统检索效率。这方面的主要研究工作集中在对倒排索引结构合理设计上，如是否记录文字位置信息以及记录的形式等。对中文信息检索系统而言，在这一部分还研究其特殊的预处理过程，即分词并建立词项的过程。由于这方面的研究较多涉及到实际工程层面的实现原理与技术细节，因此相关研究工作的介绍并不多，仅有部分工程技术报告涉及（张俊林 等，2006）（赵乐 等，2006）。

与纯文本数据相比，网络数据不仅包括非结构化的内容信息，也包括一定程度的结构化信息，这些结构化信息在网络信息检索工具评价数据质量、挖掘数据相关性等方面发挥着十分重要的作用。因此，网络信息检索中对非内容信息的处理是其与传统文本信息检索系统最大的区别所在，而其中应用最广泛的是利用超链接结构分析方法的网络数据质量评价技术，因此在本节的剩余部分将对这方面技术的主要发展情况进行讨论。

1.2.2 超链接结构分析的形式化基础

超链接结构信息是网络信息环境与传统信息媒介的最大区别之一，与用户查询需求乃至页面内容均相对独立的超链接结构，成为研究者借以评价网页质量的重要依据。

超链接是指两个网页或网页的两个不同部分之间的一种指向关系，源网页是指包含超链接的网页，超链接一般在源网页 HTML 源码中表现成如下的文字形式：

<A HREF="http://www.tsinghua.edu.cn/">清华大学主页</A>

目的网页是被超链接所引用的网页。在上述例子中，用户在源网页中可以看见的描述链接的内容被称为“链接文本”（在上述例子中链接文本即为“清华大学主页”），链接文本的特殊颜色和下划线格式表示它是可以被点击的超链接。各种基于超链接结构分析的页面质量评估算法，都是围绕对于链接关系图及链接文本的应用而展开的。

在 2001 年 SIGIR 会议上，Craswell 等人对链接结构分析算法的应用方式进行了分析，提出网页超链接结构所具有的以下两个特性（用假设的方式表述）事实上是各种超链接算法得以成立的基础（Craswell et al，2001）：

如果存在超链接 L 从页面 P_{source} 指向页面 $P_{destiny}$，则 P_{source} 与 $P_{destiny}$ 满足：

假设 1：（内容推荐假设）页面 P_{source} 的作者推荐页面 $P_{destiny}$ 的内容，且利用 L 的链接文本内容对 $P_{destiny}$ 进行描述。

假设 2：（主题相关假设）被超链接连接的两个页面 P_{source} 与 $P_{destiny}$ 比随机抽取的两个页面有更大的概率有内容相关性。

由假设 1 可以推知，拥有较高入链接个数的网页得到更高程度的推荐，并应当在页面质量评估中得到较高的评分。在较小规模特定内容网页语料库中的实验（Bharat et al，1998）证明利用这个假设设计的算法能够有效的挑选出高质量网页。因此，在实际网络环境中，为网页增加更多的入链接也成为通过作弊提高网页在搜索引擎中排名的主要方式之一。假设 1 还指出链接文本的重要特性，即它是对目标页面内容的相对客观的描述，利用这一特性设计的算法被许多研究证明是有效的提高网络信息检索质量的手段（Craswell et al，2001）（Upstill et al，2003）。

假设 2 的正确性被多项研究所证明（Davison et al，2000），它也是某些将内容分析与链接分析相结合的算法如行为扩散算法（spreading activity, SA 算法）的理论基础。利用这个假设，检索算法有理由把与相关网页在链接关系上相近的网页同样认定为比较相关，这也为不少重复/冗余网页判定算法提供了有力的在网页内容分析之外的分析途径。

由于链接分析算法是现有的大部分网络搜索引擎评价页面质量中不可或缺的部分，因此针对链接算法的作弊方式也层出不穷，如垃圾网页之间大量相互引用等。这些作弊方式导致上述两个链接结构分析算法成立的假设在现实网络环境中受到巨大的挑战，也决定了链接分析算法无法单独被应用于网页质量评估的操作。但一方面经过改进的链接分析算法还是可以为页面质量评估提供参考；另一方面，在经过数据清理之后的近似理想的网络环境中，它们还是可以发挥其挑选内容相对高质量网页的作用，因此链接分析算法仍旧是当前研究的热点之一，也是我们工作中需要重点参考的对象。

1.2.3 HITS 算法基本思路及实现

HITS 算法是由 Kleinberg 在 90 年代末提出的一种链接分析算法（Kleiberg，1999），与随后我们将介绍的 PageRank 等实用性算法不同，HITS 算法更大程度上是一种实验性质的尝试。它必须在网络信息检索系统进行面向内容的检索

操作之后，基于内容检索的结果页面及其直接相连的页面之间的链接关系进行计算。这使得在实际应用环境中使用 HITS 算法变得十分困难，尽管 Bharat et al（1998）的工作说明，通过算法改进和专门设立链接结构计算服务器（Connectivity Server）等操作，可以实现一定程度的在线实时计算，但这对于每天要处理超过 20 亿次用户需求的商用搜索引擎而言，这样的计算代价仍然是不可接受的（Sullivan，2006）。

尽管如此，但 HITS 算法仍在学术界和产业界都获得了非常多的关注，IBM 公司甚至基于改进后的 HITS 算法开发了专门的检索应用系统 Clever 系统（尽管此系统并没有投入真实的网络信息检索服务）。这是与 HITS 算法设计本身所具有的高度的数学严谨性相关的，但更重要的，是因为 HITS 算法的设计符合网络用户评价网络资源质量的普遍标准，因此能够为用户更好的利用网络信息检索工具访问互联网资源带来便利。

HITS 算法对网页进行质量评估的结果反映在它对每个网页给出的两个评价数值——内容权威度（Authority）和链接权威度（Hub）上。内容权威度与网页自身直接提供内容信息的质量相关，被越多网页所引用的网页，其内容权威度越高；与之相对应的，链接权威度与网页提供的超链接的质量相关，引用越多内容质量高网页的网页，其链接权威度越高。如果我们把一个内容权威度高的网页比作一个味道不错的饭馆的话，那么链接权威度高的网页就是旅游杂志中美食家撰写的一篇推荐饮食地点的文章。

由于网络信息检索所面临的数据对象即万维网数据具有极为繁杂的数据规模，因此，用户所涉及到的绝大多数查询主题都会返回数量繁多的相关查询结果。面对数目动辄上千上万的相关结果集合，绝大多数用户会倾向于查找出结果集合中对自己获取信息最有价值的那一部分网页。HITS 算法所解决的正是这一问题：它所施行的数据集合，就是网络信息检索工具返回的与查询主题相关的结果集合，而其输出的结果，就是对此结果集合中网页的内容权威度和链接权威度的评价。HITS 算法因而被认为能够极大地改善用户的检索体验，也得到了众多研究人员的关注。

从具体施行步骤而言，HITS 算法的施行是一个“迭代—收敛”的过程，传统 HITS 算法的操作步骤可以参照算法发明人 Kleinberg 给出的算法描述，其施行步骤如下所示（Kleiberg，1999）：

HITS（Hyperlink-Induced Topic Search）算法

1．选取网络信息检索系统的结果集合 R

将 R，R 所指向的网页和指向 R 的网页构成的链接结构图称为 N。

对于 N 中的每一个节点 n，设 H（n）是其链接权威度数值；A（n）是其内容权威度数值。

2．对 N 中每一个节点 n，设定其 H（n）和 A（n）的初始值均为 1。

3． 对 N 中的每一个节点 n， $A[n]=\sum_{(n',n)\in N} H[n']$

对 N 中的每一个节点 n， $H[n]=\sum_{(n,n')\in N} A[n']$

4．将 $\vec{H}$ 和 $\vec{A}$ 作规范化处理；

当向量 $\vec{H}$ 和 $\vec{A}$ 未收敛时，返回 3；当 $\vec{H}$ 和 $\vec{A}$ 收敛时则输出算法所计算出的针对 N 中每一个节点 n 的 H（n）和 A（n）的结果。

Kleinberg 同时证明了算法的收敛性，算法施行的过程可以参照图 1-3。

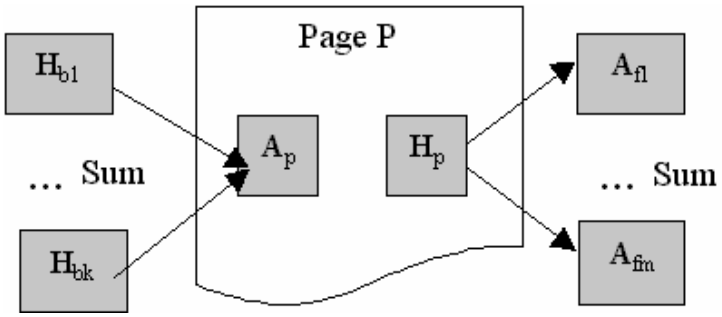

图 1-3 HITS 算法示意图（引自（Chakrabarti et al，1998），有改动）

HITS 算法在特定的应用环境中取得了一定的成功，如在 Chakrabarti 等基于 HITS 算法的小规模试验尝试中，研究者获得了超过 Yahoo!和 AltaVista 手工分类结果的检索性能，但其实验数据的规模较小，实验结果测试集合的标注也缺乏足够的客观性（Chakrabarti et al，1998）。

HITS 算法的施行对象及其迭代算法的本质决定了其不可能在网络信息检索系统中取得大规模的应用。而更多的基于实际网络数据的实验结果证明，这个算法本身在挑选内容或链接质量较高的页面时也并非格外有效，究其原因而言，大致包括以下几点：

A. 站点内部网页在权威度数值上的的相互加强；

B. 网页辅助制作工具自动生成的链接条目的干扰；

C. 与主题无关的网页或者主题漂移。

针对上述缺点， Bharat et al（1998）等人对 HITS 算法进行了相关的修改，具体内容包括忽略站点内部的链接、或者利用网页的内容相似度对 Hub/Authority 值进行初始化等。这些改进获得了有限的成功，但从算法的核心思想而言与 HITS 并没有实质的改变，因此在此不再赘述。

1.2.4 PageRank 算法基本思路及实现

PageRank 算法是 Google 公司的 Brin 等人根据因特网用户浏览模型建立的链接分析算法（Brin et al，1998）。*PageRank* 算法的基本架构和实现思路在实际商用搜索引擎的应用中取得了巨大的成功，并由此得到了研究界的普遍关注，尝试对算法进行性能和效率改进的努力一直到最近也是链接关系分析方面研究的重点之一（Haveliwala et al，2003）（Broder et al，2006）。

PageRank 算法将网络浏览模型作了合乎情理的简化：假设存在这样一名网络浏览者，他从随机挑选的页面开始，按照页面上的链接前进，在每一个页面，浏览者都有可能不再对本页面内部的链接感兴趣，从而随机选择一个新的页面开始新的浏览。

在这种浏览模型下，一个页面被访问到的概率即反映在此页面的 *Rank* 值的大小上。如图 1-4 所示，页面 q_1 包含指向页面 *p* 和 *m* 的链接，则它对 *p* 和 *m* 在 *Rank* 值上的贡献各是它自身 *Rank* 值的一半。

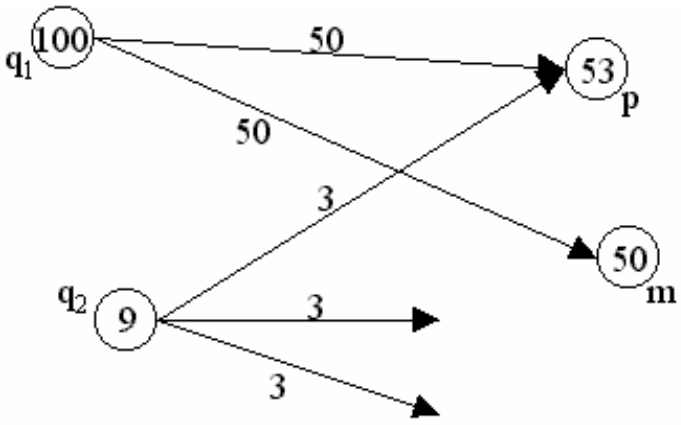

图 1-4 PageRank 算法施行过程的示意图

形式化的说，在 *PageRank* 算法中页面 *P* 被访问到的概率依下式给出：

$$Rank(P) = (1-d)/\#\,pages + d \times \sum_{P_i \in \Omega} \frac{Rank(P_i)}{\#\,outlinks(P_i)} \quad (1\text{-}3)$$

其中，Ω是有链接指向页面 *P* 的网页的集合，而 *d* 是页面 *P* 的重要性因子，由先验知识得出，反映用户认为这个页面有用的程度。简而言之，就是用户会不会从抛弃这个页面而开始一个新的随机访问过程。算法中，上述计算过程被重复进行直到运算结果收敛为止。而作为计算结果的 *Rank(P)*则被用作页面质量的评价参数。

PageRank 算法被作为 Google 的主要成功经验之一广为推介，但他在学术研究的层次上并没有获得想象之中的比较大的成功。Nick Craswell 与 David Hawking 发现，即使在链接分析方法使用较多的主页查找任务中，*PageRank* 算法及其变形也仅仅获得了比纯内容检索略好的结果（Upstill et al，2003）。根据刘悦等的实验，应用 *PageRank* 算法的结果在 TREC 大规模检索数据上“与基准测试数据基本持平”（刘悦 等，2002）。Amento et al（2000）也利用实验验证了至少在小规模数据上，包括 *PageRank*/HITS 在内的各种链接结构分析算法都无法有效的提高纯文本检索的效果。

1.2.5 链接分析算法总述

我们认为，造成 *PageRank* 和 HITS 在内的大量链接分析算法在网络信息检索研究中失效的原因，来自于一般研究所采用数据集合的不完整性。即使现在较为广泛采用的规模较大的.GOV 数据库，其规模也不过覆盖不到 20G 数据，而其链接关系表只涉及了不足 150 万个链接。对于这样链接结构大量缺失的数据集，仅仅依靠链接结构分析评价页面质量是不可靠的。这也是尽管研究界不断汇报链接分析算法对于提高信息检索系统性能没有帮助，而链接分析模块却一直成为商用搜索引擎不可或缺部分的原因。

然而面临真实的网络环境时，链接分析算法又要面临新的问题，那就是数据的繁杂和垃圾、作弊链接的存在，从这个角度讲，无论在实验或是应用的层面，链接分析算法都是解决诸如页面质量评估这样的网络信息检索发展面临问题的可行途径之一而并非全部。

1.2.6 网络信息检索系统的性能评价

在 1.1.3 节中，我们讨论了传统文本信息检索系统的评价问题，对于网络信息检索系统而言，传统的评价方式遇到了巨大的困难：一个不漏的在文档集合中找出所有相关文档，对于文档索引规模动辄以亿计数的网络信息检索系统而言几乎是不可完成的任务；而用户行为的差异也造成了评价指标体系需要进行修正。这造成了传统的 Cranfield 方法和精确率—召回率的评价指标体系对网络信息检索任务并不完全适用，必须在其基础上进行改进。

1. 标准答案集合标注方法的改进

网络信息检索的数据对象即 Web 数据为标注标准答案集合带来了几乎不可逾越的障碍，根据美国加州伯克利大学的统计报告（Lyman et al，2003），2002 年时 Web 数据就包括了超过 100 亿静态页面和超过 1500 亿个动态生成的页面。这使得对于任何查询主题而言，定位完整的相关文档集合都变得不现实，因此这个问题的解决方案注定不可能完美的解决（Hawking et al，1999）。

文本检索会议（Text retrieval conference, TREC）从 1992 年创立之初就将促进大规模文本信息检索的研究作为其首要目的，通过每年组织各种形式的检索评测，TREC 积累了丰富的在大规模文档集合中定位相关文档的经验，其核心技术被称为结果池过滤技术（pooling）。

结果池过滤技术的施行步骤如下：

A. 根据数据规模的大小，选择适当的查询结果集合规模 N。

B. 对于某个查询主题，利用若干检索技术（$T1, T2, \ldots, Ti$）对大规模文档集合进行检索，并分别得出各自的查询结果集合（$RC_1, RC_2, \ldots, RC_n$），其中$|RC_i|=N$，（$i = 1,2,\ldots,n$）。

C. 构建结果池，pool = $RC_1 \cup RC_2 \cup \cdots RC_n$。

D. 对上述结果池进行过滤，利用手工评价结果池中的文档是否与查询主题相关。

经过过滤的结果池中的文档，就可以视作此主题的相关文档集合。结果池过滤技术是基于对检索用户的行为分析而得出的，他的主要思想可以归结成：用户在面临大规模文档集合时只能借助具有一定检索精度的检索工具进行信息获取，因此候选答案集合只可能通过检索工具进行定位。这一定程度上是出于无法手工筛选整个文档集合的权宜之举，但在处理类似网络数据集合这样规模的文档集时却又是必要与合理的。

TREC 在应用结果池过滤技术构建相关文档集合方面积累了相当丰富的经验（Hawking et al，2005）。国内从 2003 年开始，也逐渐开始进行针对大规模网络文档集合检索的相关评测（如国家 863 计划基础资源与评测，SEWM 系列中文 Web 检索测试等），而所采用的构建相关文档集合的方法，也基本沿用了结果池过滤方法。

除结果池过滤技术之外，对于某些特定种类的用户需求（主要是导航类查询需求）而言，还可以利用已有的网络信息资源自动寻找相关文档集合。美国在线公司的 Chowdhury et al（2002）就研究了利用开放目录计划（ODP, Open Directory Project，一个利用志愿者标注网络资源的项目）自动查找导航类查询目标页面的可能性。由于导航类查询目标页面的唯一性，这种尝试得到了成功，但由于缺乏相应的网络资源，对于目标页面集合较大的查询需求而言，这种自动定位不可能得到广泛的应用。

2. 评价指标体系的改进

网络信息检索评价方式与传统检索系统评价的最大区别，在于其描述的网络信息检索系统的检索结果，更大程度上是一个有序的结果集合，因而基于对整个结果集合考察查准率和查全率的评价方式不再适用于网络信息检索的评价。Hawking et al（2005）指出，网络信息检索的主要评价指标包括：

A. 前 n 选精度（Precision at n, P@n)：前 n 篇中满足用户需求文档的比例，例如 P@10=0.4 则表示在检索系统返回的前 10 篇结果文档中有 4 篇文档满足用户需求。这个指标通常用来评测信息事务类或者事务类查询的性能。

B. 前 n 选成功率（Success at n , S@n）：在前 n 篇结果文档中能否有满足用户需求的文档，这是一个全或无的判断。例如 S@10=1 则表示检索系统返回的前 10 个结果中有某个结果满足用户需求。这个指标通常被用于评测网络信息检索系统对各种类型查询的综合性能。

与上述两个评价指标密切相关的典型用户行为是：根据 Silverstein et al（1998）的统计分析，超过 85%的用户只翻看搜索引擎返回结果的第 1 页（一般来说，也就是前十个结果）。这个用户行为决定了尽管搜索引擎返回的结果数目十分庞大，但真正可能被绝大部分用户所浏览的，只有排在最前面的很小一部分而已。用户日志分析结果的不同，决定了评价指标中 n 的取值。

如果说上述两个评价指标是与检索结果中的前 n 篇文档有着十分密切的关系的话，那下列评价指标则更多是由检索结果最前的少数几篇文档是否满足用

户需求决定的：

C. 首现正确结果排序倒数（Reciprocal rank of first correct answer, RR）：

$$RR = \frac{1}{Rank(\ first\ correct\ answer)} \tag{1-4}$$

RR 是指检索系统返回的结果序列中第一个满足用户需求文档出现序号的倒数。RR=1 表示检索系统返回的结果中，第一个结果就可以满足用户需求。这个指标通常用来评价导航类检索的性能，因为这类检索只有一个标准答案可依满足用户需求。

D. 平均精度（Average precision, AP)：

$$AP = \frac{1}{N}\sum_{i=1}^{N} Precision(i) \tag{1-5}$$

AP 是指检索系统返回每一个标准答案文档时精确度的平均值，例如对于某个查询需求，共有 3 篇文档满足用户需求，而检索系统共查找出其中的两篇，分别排在第 1 位和第 4 位，则这次检索的 AP 值是(1+0.5)/3=0.5。这个指标可以用来衡量检索系统对各种类型查询的综合性能。

对于若干不同的查询需求，将每个查询的 RR 或者 AP 进行平均，就得到了评价系统的平均性能评价，相对于 RR 和 AP，它们分别被称为 MRR（mean reciprocal rank of first correct answer）和 MAP（mean average precision）。RR 和 AP 都是着重强调结果序列中最靠前文档相关程度的评价指标，对于只有一个标准答案的查询需求而言，RR=AP。如对某个查询而言，如果两个查询系统都找到了一个标准答案，但分别将其返回在第 1 位和第 2 位，尽管排序非常接近，但前一个系统的 RR 值则是后一个的 2 倍。这种偏向性是与网络信息检索用户期望尽快找到满足需求的结果页面的想法一致的，因此也是一个合理的设置。

由此可见，以上的主要评价指标都与网络信息检索的用户行为密不可分，也正因为它们都试图从某个侧面反映用户需求，才能够在网络信息检索评价中得到广泛的应用。TREC 在近年组织的网络信息检索评测中，无一例外的采用了某个或某几个上述评价指标作为衡量检索系统性能的标准（Hawking et al，2002）（Hawking et al，2003）（Craswell et al，2004）。而针对中文网络信息检索的评测也都采用了类似的评价方式。

除此之外，部分研究者还设计了与用户行为直接相关的评价指标对检索系统的性能给出评价，如 Tang et al（2003）提出的查找长度（用户在找到相邻两篇相关文档时需要阅读的其他文档个数）以及排序相关度（用户认为的相关程度排序与系统实际排序的相关程度）等。这类评价指标更精确的描述了用户检索行为，但由于数据的采集需要给用户的使用带来过多的不便，因此并不适用于大规模的检索系统尝试。

1.3 面向网络信息检索的用户行为分析概述

用户行为分析是网络信息检索技术得以前进的重要基石，也是能够在商用搜索引擎中发挥重要作用的各种算法的基本出发点之一。如 *PageRank* 算法取得的成功（Brin et al，1998），很大程度就被归功于该算法设计的符合用户行为的网页浏览模型：“随机浏览模型”（random walk model）。

检索用户的行为对于网络信息检索系统研究而言是最可宝贵的反馈信息。对于网络通用搜索引擎这种反馈显得格外重要，庞大的用户群体为搜索引擎带来的不仅是巨大的挑战，其群体行为也为搜索引擎带来了调整检索系统算法的必要依据，进而成为推动网络信息检索技术发展的重要动力。面对几乎是不可穷尽的数据对象，用户在使用搜索引擎时的行为不仅直接的对查询结果进行了评价，也间接的对 Web 页面的质量，进而甚至对与搜索引擎相关的一系列商业产品（如竞价广告等）的优劣程度给出一个相对客观的评价。

1.3.1 用户群体行为分析的研究历史

国内外研究者一直对网络信息检索工具（主要是搜索引擎）的用户群体行为分析给予了足够的重视，早在网络信息检索研究开始之前，用户相关反馈在文本检索中的研究就被认为是提高检索精度的主要途径之一。面对复杂的用户需求与更加复杂浩繁的文档集合，通常只有几个字词组成的用户查询成为了影响用户与检索系统之间信息传递的瓶颈，而相关反馈则成为了更明确表达用户意图的有效方案。尽管由于增加了使用成本，用户相关反馈并没有能够在文本检索中得到大规模的应用，但这方面的工作后来却成为了伪相关反馈（pseudo relevance feedback）、查询扩展（Query Expansion, QE）等信息检索核心技术的基础。与此同时，对网络用户群体的行为分析自从 WWW 得到大规模应用以来，

也得到了足够的关注。Cockburn et al（1996），Catledge et al（1995）和 Tauscher et al（1997）等工作就分别在 90 年代中期左右对 Web 用户的浏览行为进行了调研和分析， Byrne et al（1999）还提出了 Web 用户目的分类的概念，以此对用户浏览 WWW 的整个行为进行建模，但上述工作都没有将注意力集中到网络信息检索的范畴中来。

网络信息检索工具得到普及之后，面向网络信息检索的用户群体行为分析得到了更多的关注，主要的搜索引擎技术供应商如 AltaVista，Yahoo!，Excite 等都花费了巨大的人力物力进行基于用户行为日志的相关研究。这是因为：用户群体行为分析的结果对于检索系统结构与算法改进有着重要的指导作用，这是由于只有借助于这种分析，才可能了解纷繁复杂的用户需求背后的统计性规律，并找出隐藏于用户查询背后的真实需求，从而提高检索的效率和效果。

总体而言，这方面的研究成果可以按照其分析对象的不同划分为三类，即针对搜索引擎用户查询特征、查询信息需求特征、和用户群体特征的行为分析三类。

1.3.2 用户查询行为的宏观统计分析

对用户查询行为的宏观分析是针对检索系统用户群体行为分析中较早开展的一部分工作。在 1998 年左右，部分研究者如 Silverstein et al（1998）和 Jansen et al（1998）等就开始对商业搜索引擎的用户日志进行大规模的分析，他们的研究着眼于描述网络检索用户的一般性行为，其主要分析对象包括查询需求的统计特征分布、用户使用检索工具的习惯、检索词同现情况等。Silverstein et al（1998）指出了一系列对检索系统设计有深远影响的用户群体行为分析结果，这些结果包括：

1．用户需求中有大量的重复项，在被分析的近 5.8 亿个用户查询中，有 4.2 亿个用户查询是重复冗余的。这部分内容在第 1.2.1 节进行了讨论，其充分证明了在搜索引擎设计中引入缓存（cache）机制的必要性。

2．77.6%的用户在进行检索的过程中不对其查询进行修改。这突出了如查询扩展、伪相关反馈等协助系统理解用户查询的工作的重要性。

3．超过 85%的用户只翻看搜索引擎返回结果页面的第一页。针对这个现象而设计的各种评测指标应运而生，关于它们的信息将在第 1.4.2 节中详述。

4．尽管搜索引擎提供了若干复杂的查询功能，如组合逻辑查询等，但 80%的查询并没有使用这些功能，这说明使用检索系统的过程中，简便仍是为用户看重的，也说明了各种复杂功能的指定应从用户使用便利的角度出发。

1.3.3 用户查询信息需求分析

对用户查询信息需求的分析是近年来用户群体行为分析研究的热点之一，IBM 研究院首先提出了检索是由“任务驱动”（task oriented）的概念，在他构想的用户检索流程模型中，查询任务决定了用户的查询需求，进而反映在用户的查询词上。Broder（2002）指出，用户的查询任务包括以下三类：

导航类（Navigational）：目标是查找某个特定的站点或者网页。如“上海市政府网站”、“清华大学招生简章”等（摘自百度网站“搜索风向标”栏目，下同）。

信息事务类（Informational）：目标是获取可能位于一个或某几个网页上的信息。如“现代企业制度的形式”、“农村党员队伍状况”等。

事务类（Transactional）：目标是查找能够处理某些以 Web 为媒介的事务的网页。如“连连看下载”、“歌词查询”等。

其中事务类检索是 Broder 重点推崇的概念，他认为当前事务类检索的需求已经相当巨大，但还没有适当的网络检索服务可以有效的完成此类检索。在他提出的搜索引擎分代概念中，能够处理事务类检索的搜索引擎被称为第三代搜索引擎。Broder 同时利用对 AltaVista 用户日志的分析和对用户调研的结果给出了三类检索分别所占的比例，即 20% : 50% : 30%。根据这个分析，信息事务类检索仍然是用户使用搜索引擎的主要任务，但事务类检索也占有了相当大的比重，而导航类检索所占的比重就较低。随后，雅虎公司的 Rose 等人也提出了类似的任务分类概念（Rose et al，2004），他们对 Broder 提出的三类体系进行了更详细的划分，并制定了利用日志对用户查询进行手工分类的详细步骤。他们分析的结果说明，三类检索所占的用户查询比例大致为 15% : 60% : 25%，这一结果与 Broder 的分析基本一致。

对查询任务进行划分的出发点在于，针对三类检索可以使用不同的检索模型、参数，甚至评价方法也随着检索类别的变化而有区别。因此实现检索类别的自动划分对于提高检索性能和增加检索评价的可信度都有非常重要的意义。

TREC 在 2004 年引入了一个检索任务自动分类的子任务，最后的实验结果说明自动分类的正确率约为 60%，能够与手工分类的效果比较接近（Craswell et al，2004）。Lee 等人在加州大学伯克利分校内部使用的一个搜索引擎的用户日志上，得到了 90%左右的自动分类正确率（Lee et al，2005），但他同时也指出，这个正确率是在充分使用了用户点击信息的基础上得到的，如果不使用此类信息（在预测查询分类时，是无法得知用户点击信息的），则正确率与 TREC 的结论类似，即维持在 60%左右的水平。

1.3.4 针对特定检索用户群体的行为分析

针对特定用户（user-specific）的检索系统是近年来备受关注的一个发展方向，通常认为对特定用户行为的学习有助于提高检索系统的检索效果，因此对特定用户的行为分析研究也一直在进行当中。

Navarro-Prieto 等人对不同 Web 使用经验的检索用户的不同行为进行了比较（Navarro-Prieto et al，1999），他们得出的结论是：有 Web 使用经验的检索用户在使用检索工具时比较不容易受检索结果表现形式的干扰。Zukerman 等人则根据用户以往查询需求的信息，利用马尔科夫模型对接下来的用户需求进行预测（Zukerman et al，1999）。这方面的工作被认为对于提高检索工具界面与功能设计有一定的指导作用。在个人信息管理工具日渐强大的今天，这类行为分析也得到了越来越多的重视。

综上所述，用户群体行为分析作为一种极为重要的研究思路与方法，在包括检索系统算法设计在内的网络信息检索研究领域具有极其重要的研究价值和广泛的应用背景。在本文的工作中，我们将试图利用用户群体行为分析的方法解决网络信息检索系统信息传递中的瓶颈问题，从而为互联网信息利用水平的提高做出贡献。

1.4 课题研究重点及主要工作

1.4.1 用户交互方式与网络信息检索系统的发展

针对个人需求的网络信息检索服务是伴随着 1990 年蒙特利尔大学学生创立 Archie 网络文件检索系统而发展起来的，但其技术积累则在上个世纪 80 年代末

就开始了。早在 1988 年，信息检索领域的权威会议 SIGIR（ACM Special Interest Group of Information Retrieval，参见 http://www.sigir.com）就有了专门针对超文本进行检索的相关研究内容。

1994 年，面向个人用户的网络浏览服务诞生；1998 年，个人网络用户超过 5000 万，这标志着网络真正从宏观上成为一种主流的信息传递与获取手段。如果把一种媒体从诞生到拥有 5000 万用户的时间作为其发展速度的衡量标准的话，那么广播的相应发展时间是 38 年，电视是 13 年，互联网则只用了短短的 4 年。而搜索引擎的发展速度更是惊人，1995 年底，作为最早的互联网商用搜索引擎之一的 AltaVista 提供服务的第一天，就有超过 30 万用户感受了搜索引擎的巨大魅力。目前，中国搜索引擎用户已占互联网用户的 95.2%，绝对用户数目超过 1 亿人（CNNIC 2005）。搜索引擎不仅成为了人们获取互联网信息的首选工具，甚至也成为了人们在生活和工作中必不可少的信息渠道之一。

互联网搜索引擎之所以在极短的时间内取得这样巨大的发展，一方面，是由于搜索引擎的处理对象即互联网数据的极大丰富，人们日常生活与工作中所需的大部分信息和知识都可以通过这个数据源获取、或进行交流；另一方面，这也是与搜索引擎工具本身的易用性相关的，这种易用性与互联网搜索引擎所采用的“关键词查询+选择性浏览”的信息交互方式密切相关。而在搜索引擎出现之前的不少信息管理工具，则往往由于没有找到解决用户交互问题的合理方案而陷入困境。

自然语言自动问答系统（如 MIT 开发的 Start 系统 http://www.ai.mit.edu/projects/infolab/；以及现在仍在运行的 AskJeeves 搜索引擎 http://www.askjeeves.com/等）尝试处理用户提出的自然语言问题，但往往被无法正确理解用户需求的问题所困扰。基于数据库研究方法的信息检索系统（如 Konopnicki（1995） 的 W3QL 系统， Mendelzon（1997）的 WebSQL 系统等）通过制定专门查询语言的方式来确保对用户需求的准确理解，但却往往使系统本身的使用变得过于复杂而难以普及。根据 Silverstein（1998）和余慧佳（2006）的用户行为调查工作，近 80%的英文用户和 95%以上的中文用户根本不使用高级查询功能，因此使大多数用户去学习使用专门的查询语言进行信息获取是不现实的要求。

关键词查询找到了一个对用户和计算机都合理的信息传递方式，而用户自行选择性浏览的结果挑选方式对于现有的搜索引擎处理能力而言也比较合适。

这种交互方式与搜索引擎一起成为了人们获取信息的固定模式之一，并对其他信息管理工具如图书馆查询系统、专利查询系统等产生深远的影响。目前针对搜索引擎用户交互方式的改变，例如查询提示、查询纠错等服务，事实上都是基于这一交互方式的补充而非对其进行本质的修改。

1.4.2 网络信息检索系统的信息处理困境

“关键词查询+选择性浏览”是一种符合现有网络信息检索技术发展水平和用户行为习惯的用户交互方式，因此它大大提升了系统的易用性，在网络信息检索系统的迅速普及中起到了不可替代的作用。由于网络用户知识水平的提高和搜索行为习惯的变化是一个较为缓慢的过程，而与网络信息检索相关的基础研究如自然语言处理技术等的发展也很难在短时间内有较大的突破，因此这一交互方式将长期伴随网络信息检索工具的发展。

尽管这种交互方式具有提升系统易用性等不可忽视的优势，但它同时带来的用户信息需求的传递瓶颈问题却也成为了阻碍网络信息检索工具发展的最大困难之一。用户的检索需求往往是复杂的，我们在 1.3.1 节中介绍了用户检索需求的几种主要类型，实际应用环境中细化了的检索需求则通常复杂的多，例如同样是事务类查询，又分为线上资源下载、线下资源获取、交互服务定位等多种细化的需求。由于这些复杂的需求只能通过关键词集合这一十分简单的形式加以描述，因此造成了用户需求和网络信息检索系统之间的信息传递瓶颈，检索系统难以获得用户查询关键词背后的真实信息需求，从而带来了网络信息处理中的极大困难。这种困难主要表现在以下两个方面：

1、在网络数据处理过程方面，尽管搜索引擎声称的网页索引数量不断增加，但用户的使用体验却没有质的增长。2004 年 12 月，Google 报告的索引规模为超过 81 页网页；2005 年 8 月，Yahoo 声称的索引量达到了空前的 192 亿网页。尽管如此，同年，Gulli（2005）的调研结果却说明，没有任何一个搜索引擎可以覆盖互联网上的所有有用资源，覆盖率最高的 Google 也仅仅占有有用资源的 3/4 而已。造成这种问题的核心原因是：搜索引擎难以获知用户的真实信息需求，因而在其数据处理过程中较少考虑、也很难做到选取尽可能满足用户需求的信息资源，造成其数据收集过程无的放矢，缺乏有效的数据质量评估环节。

2、在用户查询需求处理过程方面，由于用户需求仅能通过其查询词表述，搜索引擎很难准确理解用户的查询背后的信息需求。实际应用中，同样的查询

词可能承载不同的信息需求，因此在查询词准确表达用户需求这一假设基础上设计的传统文本信息检索模型自然难以很好的满足用户需求，而基于此类模型的查询反馈和伪相关反馈等工作就更加无从谈起。搜索引擎无法准确获知用户的信息需求，因而不可能按照信息需求的不同挑选结果或进行倾向性的排序，只能将与查询关键词相关的网页资源全部反馈给用户，这造成了用户检索过程中返回的结果列表动辄成千上万，但其中真正符合用户需求的却非常少。

网络信息检索系统所面临的这两方面困难，其核心原因在于现有信息交互方式造成的用户信息需求传递瓶颈。如果不能在深入理解用户信息需求的情况下，从网络数据处理和用户查询处理两个角度改进算法与模型设计，就很难在现有的检索系统信息交互手段下进一步提升检索性能。

1.4.3 本文研究重点：基于用户群体行为分析的网络信息检索

“关键词查询+选择性浏览”的信息交互方式造成了上节中所述的网络信息检索系统技术发展困境。由于这种信息交互方式将长期伴随信息检索系统的发展而存在，因此在这种交互方式的限制下尽可能缓解用户信息需求传递的瓶颈问题，是提高网络信息检索系统性能的必由之路。

如 1.3 节所述，用户群体行为分析的研究方法在信息检索研究的发展过程中发挥了十分重要的作用，很大程度上推动了网络信息检索系统的算法设计和效果评估技术的发展。面对用户信息需求传递的瓶颈问题，使用用户群体行为分析方法深入挖掘检索用户的行为特征，借以改进网络信息检索系统的网络数据和用户查询需求处理方式，将是一个可行的解决途径。

这是由于，用户的行为特征是网络信息检索系统可以获得的最直接的了解用户信息需求的方式。由于运行和维护网络服务的需要，搜索引擎记录了其检索用户的查询和结果点击行为，用户的真实查询信息需求正蕴含在这些“关键词查询+选择性浏览”的行为当中。针对用户群体宏观行为的研究发现，实际应用系统中需要处理的绝大多数（超过 90%）的用户查询需求曾经在以往的查询历史中出现过（Liu，2006）。因此在无法实时获得用户信息需求的情况下，通过对用户宏观行为的把握和用户群体历史行为的分析，是完全有可能对当前的用户信息需求做出估计的。

这说明了以用户群体行为分析指导网络信息检索研究的必要性。一方面，网络数据的处理和质量评估过程如果不考虑用户行为特征，则其数据满足检索

用户需求的程度难以保证，有限的计算和存储资源很有可能浪费在对用户无用的数据上；另一方面，用户查询需求的处理过程如果不考虑用户行为特征，则信息检索系统很难准确理解用户真实的信息需求，后续的检索与结果评估过程的可靠性更无从谈起。尽管面向检索系统技术需要的网络数据整合与用户行为挖掘问题早已得到产业界的广泛重视和研究，但是数据获取和处理能力等方面的限制造成学术领域涉及这方面的研究相对较少，面向中文网络数据环境和中文网络用户行为的工作更是凤毛麟角。由于互联网搜索技术在互联网应用技术中的基础地位，这种情况在不久的将来，将有可能妨碍对于中文互联网资源利用的进一步开展。

可见，利用对检索用户群体行为的深入挖掘与分析，对互联网的数据环境和用户查询需求进行重新认识，是一个使网络信息检索工具能够更加有效的处理检索需求的可行方案。用户交互方式引起的信息传递瓶颈造成了用户的信息需求无法充分表达，而我们的努力，在于通过用户群体行为分析的方法，根据群体行为估计个体用户的信息需求，并在网络数据的组织和利用过程中考虑到用户普遍信息需求的影响，借以缓解信息传递瓶颈的影响。

综上所述，基于用户群体行为分析的研究思路尝试解决困扰网络信息检索工具发展所面临的问题，既是互联网应用水平提高的必然要求，也符合现有的研究发展水平的选择，因此，以这方面的研究内容作为本论文研究关注的重点是必要和可行的。

1.5 论文各部分的主要内容

论文的第一章介绍网络信息检索的研究背景与意义，并对相关研究领域的主要相关研究工作加以介绍，最后提出网络信息检索当前面临的主要问题以及相应的本文研究重点和主要工作。

第二章面向网络数据对象，基于用户群体行为分析的方法解决网络数据质量评估的问题。首先，提出面向信息检索用户需要的网页质量评估标准；然后通过对海量规模中文页面集合的分析，分析用户查询目标页面的查询无关特征；最后提出一种新的基于机器学习方法的网页质量评估模型和算法框架，并对此网页质量评估方法的有效性进行考察。

第三章面向用户查询对象，基于用户群体行为分析的方法解决用户查询需

求的识别与挖掘问题。首先，针对中文搜索引擎用户的宏观查询行为进行分析，考察用户的行为模式与特点；然后基于真实规模的搜索引擎用户日志，提取用户查询、点击行为特征；最后提出一种新的利用基于用户行为挖掘和模式分类方法识别个体用户查询需求类别的算法框架，并对此方法的有效性进行考察。

第四章面向信息检索研究中检索效果的评价问题进行研究，基于用户群体行为分析的方法提出一种新的搜索引擎评价的自动化处理方法。这种方法基于通用的传统信息检索系统评价框架，但把原有框架中涉及人工评价、造成评价反馈缓慢、结果可信度降低的步骤用面向用户日志挖掘的自动方法完成，大大提高了性能评价的公正性和及时性。

第五章介绍基于用户群体行为分析的思路和本文研究成果改进网络信息检索系统架构设计的尝试。主要提出了面向网络数据对象处理的层次索引结构设计方案，以及以宏观用户行为规律为指导的查询处理策略设计方案，并尝试将两方面的设计紧密结合，以在当前的检索系统交互方式限制下尽量缓解用户信息需求传递的瓶颈问题。

在论文的最后一章对本文的所有研究工作给出结论。

第二章 基于用户群体行为分析的网页质量评估算法研究

2.1 引言

如上章内容所述，尽管网络信息检索技术在研究和应用方面都已经取得了很大进展，但以搜索引擎为代表的网络信息检索工具还无法充分满足人们快捷高效获取网络信息的要求，搜索引擎现有的信息交互方式造成了用户信息需求传递的瓶颈并导致了网络信息检索系统发展上的困难。在网络数据处理方面，这种困难主要表现在如何从用户的实际需求出发进行网络数据质量的评估上。

按照加州大学伯克利分校的保守估计（Lyman et al，2003），2003年仅静态Web页面的数目就超过了200亿个。而当时世界上索引量最大的搜索引擎Google只索引到33亿个Web页面，尚不到静态页面数目的1/6 （Sullivan，2005）。因此，提高索引规模成为当时各个商用搜索引擎竞争的焦点。但在2005年，Yahoo!报告的索引量达到创纪录的190多亿页面之后，Gulli et al（2005）的调研结果却说明，尽管搜索引擎的索引量节节攀升，却没有任何一个搜索引擎可以覆盖互联网上的所有有用资源，覆盖率最高的Google也仅占有有用资源的3/4而已。此后不久的2005年9月，Google从其主页上删除了页面索引数量的数据，并解释说“绝对的数字已经不再重要”，经过索引规模竞争之后的搜索引擎更加务实的面对网络数据资源，索引数据的质量而不是数量成为网络信息检索研究更加关注的课题。

造成搜索引擎关注重点从数量到质量转变的核心原因是其面临的巨大的存储和网络服务压力。2003年，当Google只能索引到网络页面资源的不足1/6时（其索引总量仅是静态页面数的1/6，而静态页面资源相比网络上的动态页面资源或称深层网络资源只是冰山一角），就需要耗用超过一万台高性能计算机组成的庞大集群来进行数据存储操作（Singhal，2005）。而目前每天超过2.7亿次的用户查询量更是使搜索引擎系统提供网络服务的难度进一步加大（Sullivan，2006）。面对这样繁杂的数据环境和海量的用户需求，搜索引擎不得不从简单的提高索引数量转而更加关注质量问题，以使其有限的存储和计算资源更加合理的使用。

网络数据质量评估问题的核心，是清除索引中的冗余、低质量、不可信和过时数据，而保证真正满足用户需求的数据能够得到检索算法的关注。在2003年，Google公司的Henzinger等人就在国际人工智能联合会议的大会主题演讲中

提出如何能够独立用户查询评判数据质量的问题（Henzinger et al，2003），并将其列为搜索引擎面临的最重大挑战之一。

2.2 相关工作

面对网络数据中的冗余、低质量、不可信和过时现象，产业界和研究领域的主要解决思路集中在两个方面，即使用网页数据清理的方式清除无用数据，或者使用网络数据质量评估的方式挑选有用数据。这两个方面的问题解决都需要借助于机器学习方法和特定网页的分类技术，只不过前者分类的目的集中在无用数据挑选，而后者集中在有用数据定位。但是，由于网络数据中低质量数据的形式层出不穷，而针对网络页面的均一采样问题本身就是难以解决的问题（Henzinger et al，2003），因此挑选高质量页面比起挑选低质量页面的方法对于网络信息检索系统更加可行。

围绕高质量数据定位的数据质量评估工作按照其评价对象的粒度不同可以分为两类，即Web页面质量评估（宏观粒度）和页面内部元素质量评估（微观粒度），前者以Google公司提出的PageRank算法和IBM公司Clever系统的基石HITS算法为代表；而后面一个思路则集中体现在作为MSN搜索引擎核心技术之一的VIPS（基于视觉效果的网页分块算法）算法上。这些算法被各自企业投入到大规模网络搜索引擎的应用中，并取得了一定的成功。因此，在下文的相关工作介绍中，我们将重点回顾Web页面质量评估，Web页面内部元素质量评估方面的工作，除此之外，由于冗余页面识别技术与传统的页面质量评估技术有较大的区别，也将专门介绍相关工作的开展情况。

2.2.1 宏观粒度网络数据质量评估技术

根据Yi et al（2003）的分类方法，Web环境中的噪音信息可以分为不同粒度，全局噪音是指页面级别的噪音信息，如镜像站点页面，过时页面，虚假页面，纯广告页面等；而局部噪音则是页面内部的噪音信息，如广告条、导航条等。处理全局噪音是页面级别数据质量评估的任务，而清除局部噪音则是页面内部元素级别数据质量评估的工作，两者是互相补充的关系。

比较微观粒度的网络数据质量评估，宏观粒度的评估或称网页质量评估在搜索引擎的结构体系中发挥着更重要的作用。高质量网页中有可能存在着无用

的内容块，但低质量网页中却很难找到高质量的网页内部信息：这说明对低质量网页进行微观粒度的数据质量评估实际上是没有效果的。因此在搜索引擎处理网络数据的过程中及时对网页质量进行准确的评估，有利于提升搜索引擎整个运行过程的效率和性能。

当前网页质量评估的主要工作集中在超链接关系分析方面，超链接结构关系是网络数据区别于普通文本数据的最大差别之一，因此也成为网络信息检索相关研究的重点。网页是超链接结构分析所关注的主要数据对象，而超链接结构分析算法的主要目的，往往也是为了从网络结构图（Web Graph）上找出重要的节点（网页）。因此网页质量评估工作借助超链接结构分析是自然的选择，而媒体领域对Google及其*PageRank*算法的关注，更是使链接结构分析算法成为研究界和产业界关注的重点。网页质量评估方面的相关工作大都集中在利用PageRank框架进行某些特定应用需求的改进（Haveliwala et al，2003）（Benczur et al，2005），或对Brin et al（1998）提出的PageRank传统算法进行效率上的提升（Mcsherry et al，2005）（Broder et al，2006）。

在第1.2节，我们回顾了超链接结构分析的形式化基础，并介绍了两个应用最广泛的超链接分析算法：*PageRank*和HITS算法的基本思路与实现。包括这两个算法在内的所有链接结构分析算法的出发点都可以认为是由Craswell et al（2001）等总结的两个超链接结构特性，即1.2.2节中提到的内容推荐假设与主题相关假设。链接结构分析算法必须在这两个假设成立的情况下方为合理，而假设成立的前提则是网络上绝大多数的网页作者在添加网络链接时的行为符合假设的规定。搜索引擎所依靠的网页质量评估算法依靠这两个假设，而搜索引擎当前的运营体制却无法保证假设的成立。

这是因为，随着搜索引擎逐渐成为用户访问未知网站的最主要方式，一个网站想要被用户点击浏览，最好的办法就是出现在搜索引擎的比较有利排序位置的检索结果中。然而，搜索引擎竞价排名机制的引入，造就了搜索引擎作为网络信息获取渠道和广告媒体的双重身份，即：为获取更高的搜索引擎排名以吸引网络用户关注进而得到某方面利益的途径有两条：利用搜索引擎作为网络信息获取渠道的身份提供更高质量的信息，或利用搜索引擎作为广告媒体的身份提供较高的广告推广费用。搜索引擎及其运营机制，在高质量的网络信息与真实的物质财富之间形成了直接的联系，而这种联系的公平性则是由搜索引擎的运行算法和运营机制决定的。

对于网站而言，改变其网页之间链接结构花费的成本微乎其微，而只要其能掌握相当数目的网页资源，就足以进行对链接分析算法结果产生影响的链接结构改变。这种几乎不耗费成本的改变有可能带来的却是搜索引擎排名的提高和访问量乃至物质财富的提升。因此针对链接结构进行这种“改变”以获取更高搜索引擎排名的“作弊”方式层出不穷，这也成为困扰链接分析算法发挥效用的最大问题。

面对这种针对链接结构的“作弊”方式，通过链接结构分析研究发现“作弊”网站的方法是一条可行的途径（Gyongyi et al，2004），但更可能产生效果的方法则是增加“作弊”的成本。如果在网页质量评估的过程中不仅采用链接结构分析的方法，而是充分采集高质量网页各方面的内容和非内容特征，则会更加准确地评估页面质量，并增加通过“作弊”方式提高网页质量的成本。正如Henzinger et al（2003）所指出的，采用除链接结构特征之外的多种特征共同评价网页质量，是更合理和更全面的评价网页质量、清理网络垃圾的必由之路。

2.2.2 微观粒度网络数据质量评估技术

由于数据挖掘研究的需要，基于页面内部结构分析的数据质量评估研究水平要高于Web页面级别的相应研究。这方面的工作在数据挖掘相关研究中被称为数据清理环节，甚至从万维网开始出现之前，数据挖掘方面的研究人员就开始就当时存在的一些半结构化文档格式（如BibTeX等）进行数据清理方面的研究。相比信息检索更加关注网页文本的内容以及重要性特征而言，数据挖掘注重研究的是如何将半结构化或非结构化的数据形式转化为结构化的数据形式，并引入数据库技术进行处理，这导致了两种目的下的数据清理研究的差异。这说明面对数据挖掘和面向信息检索需求的网页数据清理在去除网页噪声信息的层面上是相同的，差别在于两者对于何为噪声的定义不同。

我们研究的重点是面向信息检索需要的网络数据质量评估，因此重点关注的也只是清除页面局部噪音的相关研究工作，而并非如何查找数据中的错误、不一致现象等传统数据挖掘研究关心的内容。然而，网页局部噪音的清除问题在包括网络信息检索（Web IR）、信息抽取（Information Extraction）、甚至网络数据挖掘预处理的研究中都发挥着十分重要的作用。Büchner et al（2000）指出，包括去除局部噪音在内的数据预处理过程是网络数据挖掘中最费时费力的一个步骤，甚至可以花费全部工作精力的80%以上。因此局部噪音处理一直是相

关领域研究的重心之一，总体来看，可以分为两大类研究：

第一类研究集中在清除某种特定类型的垃圾信息，如Davison等人提出了一个基于学习方法的去除Web页面中广告条的算法（Davison et al，2000b），而Kushmerick等人则利用类似的方法实现了页面中无关链接与垃圾链接的去除（Kushmerick et al，1999）。类似算法都依赖于一定规模的训练数据，并且仅关注于某一方面局部噪音的清理，难以应用于大规模的网络数据清理实践。

与第一类研究不同，第二类研究更注重从页面整体的角度考虑问题，这一阶段研究的特点是将页面进行“分块”处理。这种处理的出发点是由于HTML结构的特点，网页的组成一般有明显的“块”（block）组成，每个块的内容都有特定的主题，如图2-1中，块VB1_1是统一的Logo信息，块VB4则是版权信息，块VB2-1是导航链接信息，而块VB2_2才是页面的主题内容。如果能够实现合理分块并挑选出网页主题内容所在的块，则其它块的内容自然可以作为局部噪音信息丢弃，从而实现页面内部的数据清理。

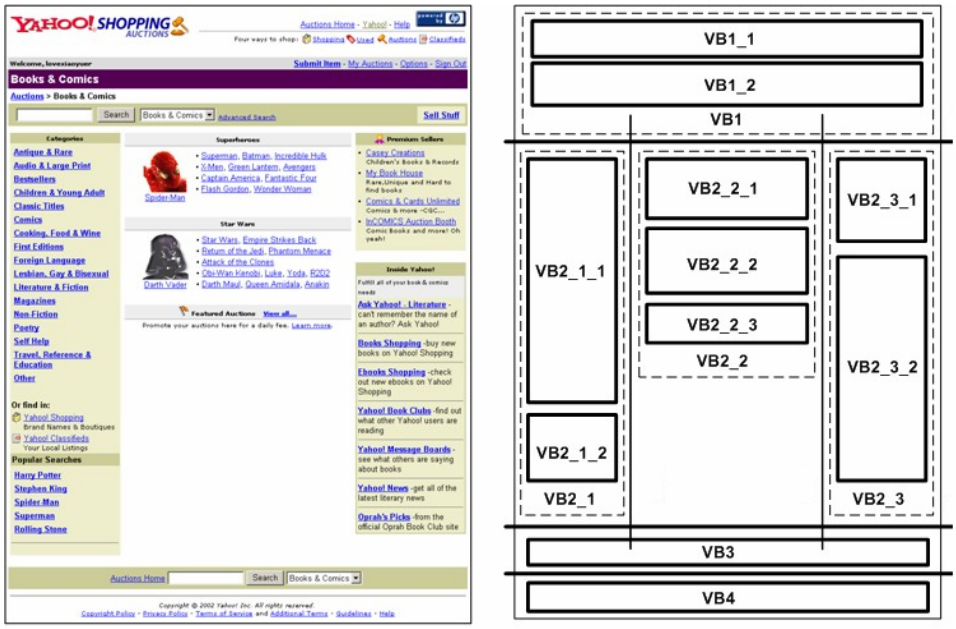

图 2-1 Web 页面的分块情况（引自 Cai（2004），有改动）

Lin等人提出了一个从Web页面中提取信息量最大的页面块的算法（Lin et al，2002），他们的分块算法基于HTML页面中的<TABLE>标签，而信息量计算的依据是根据块中所出现单词的词频所计算出的信息量数值。Bar-Yossef et al

（2002）提出的算法把Web页面清理定义成一个模板频度检测问题，如果某个页面模板被发现频繁的使用，则这个模板就会被定义成噪音信息并加以去除。Yi et al（2003）中的算法综合了上述两种算法的优势，它认为Web页面清理必须依赖于站点信息，利用对同一站点内页面的比对，可以很容易的发现页面的模板信息；它同时基于页面的HTML结构构造站点模板所对应的树结构（称为站点风格树，Site Style Tree, SST），利用信息量计算的方法可以得到SST中哪一部分是重点内容的判断，进而实现页面的数据清理。

微软亚洲研究院相关研究人员提出的基于视觉效果的页面分类算法则在页面内部数据清理方面进行了更为细致的研究，他们首先提出了一个基于用户视觉特征的网页分块算法（Cai et al，2004），随后利用机器学习算法对用户浏览网页的习惯进行建模，得出不同网页块之间的重要性量度计算方法（Song et al，2004），从而实现了对低重要度块的清理；他们还进一步对基于网页块为基本单位的链接结构分析算法也进行了分析（Cai et al，2004b），证明以网页块的粒度进行网络信息检索是一条可行的途径。

综上所述，Web页面局部噪音去除方面的研究成果，已经较好的利用网页内部的结构特征完成页面内部数据清理和分块质量评估的任务，相比较页面级别（针对全局噪音）的网页质量评估而言，这方面的研究要成熟和完善的多。

2.2.3 冗余页面识别技术

除垃圾和低质量页面之外，网页数据清理要处理的另一类无用页面是网络环境中存在的大量重复文本即冗余页面。2004和2005年公布的两次中国互联网络发展状况统计报告指出，用户在互联网上查询信息时遇到的最大问题就是重复信息太多（有超过40%的用户认为这是他们面临的最大问题）（CNNIC，2004）（CNNIC，2005）。根据同样于这两年发布的两次中国互联网络信息资源数量调查报告统计，中国互联网中网页的重复率在25%以上（CNNIC，2004b）（CNNIC，2005b）。大量的冗余页面直接影响到网络信息检索系统的运行效率和效果，因此判定并去除冗余页面，是网络信息检索系统网页处理中必不可少的步骤。

相比去除低质量页面，进行页面质量评估的努力，冗余页面判定算法研究的起源要早得多，当前的研究水平也要高得多。最早在20世纪70年代初，就有学者出于知识产权保护的需要，开始了软件源码复制/剽窃检测的研究，Ottenstein

et al（1977）还提出了基于属性计数（attribute counting）检测软件剽窃的方法。但真正的面对自然语言复制检测的研究还是在20世纪90年代以后出现的，Mander等人提出了基于字符串匹配的方法来衡量文件之间相似度的算法，并应用于其开发的SIF工具，借以在大规模文件系统中寻找内容相似的文件（Manber，et al 1994）。1995年，Google公司创始人之一的Brin首次提出了文本复制检测机制（Copy Protection Method）及相应的算法（Brin et al，1995），这一算法与后来的改进算法SCAM（Stanford Copy Analysis Method）（Shivakumar et al，1995）都是基于文本向量化（Vector Space Model）的观点来进行文本重复程度的检测。

重复网页判定工作也随着WWW的发展和网络信息检索工具的诞生而逐渐走上历史舞台，网络数据的庞大规模和检索工具实时性的要求造成任何一种重复网页判定算法必须把处理效率作为首要考虑的因素之一，因此这方面算法设计的重心一般都集中在网页特征降维上，即如何使用适当数目的网页特征取得较高效率与较好去重效果的折衷。这方面，Broder提出的DSC算法（Digital Syntactic Clustering）（Brode et al r，1997）与其改进DSC-SS算法（DSC-super shingle）是比较早期的解决方案，它们通过将网页文本分块的方式减少网页特征，但仍采用了文档间相互比较相似性的方式来判重，因此算法效率依旧较低。2000年，Campbell等人提出了后来被称为I-Match的去重算法的雏形（Campbell et al，2000），这一算法基于网页文本计算对应每个网页的唯一签名码（用Hash键值表示），可以直接基于Hash键值是否冲突来判断页面重复，这大大的提高了网页去重算法的效率。

尽管大量用户仍然指出冗余网页的存在影响了他们查询网络信息的努力，但从网页去重算法自身而言，去重的精确率与召回率已经普遍在80%以上，算法效率一般也在网络信息检索工具可以容许的范围内。从搜索引擎的应用实践上讲，比较成熟的搜索引擎系统也已经很少针对同一个查询主题返回大量的冗余页面，用户面临的冗余网页问题，大多是从网络浏览的体验中遇到，而非是检索工具造成的，因此尽管冗余网页判定是信息检索工具进行网络数据处理中必不可少的一部分，也有一定的提高余地，但并不准备将其作为本文主要研究的方向。

2.2.4 网络数据质量评估方式总述

综上，我们针对网络信息检索工具针对网络数据质量评估问题的各种努力

进行了一些评述，尽管我们对几种方法分别进行介绍，但实际操作中，往往几种方法会结合进行，以便达到更好的数据处理效果。任何一个不涉及数据质量评估模块的搜索引擎，将在当今的网络数据环境中寸步难行，更谈不上为用户提供高质量的检索服务。

在评述的过程中，我们也可以发现，宏观粒度的网络数据质量评估即Web页面质量评估技术是整个数据质量评估过程的中心环节，在现有的网络信息检索工具架构下，它决定了数据索引的组成，因而对检索工具的结果反馈和用户体验产生直接的影响。而相关方面的研究工作也存在较多的不足之处，因此，基于用户群体行为分析方法使最可能满足用户需求的网页进入索引，也就成为我们这方面工作的核心内容。

2.3 用户信息需求与网页质量的定义

在深入讨论网页质量评估方法之前，有必要从用户信息需求的角度对网页质量给出更明确的定义，以便使我们的质量评估方法能够更好的为缓解用户信息需求的传递瓶颈问题服务。

按照1.3.1节的论述，网络信息检索用户的查询信息需求可以分为导航类查询、信息事务类查询和事务类查询三个类别，从查询目标页面个数的角度进行分析，则导航类查询的查询目标页面个数通常唯一；而其余两类查询的查询目标页面个数不唯一。由于用户的查询信息需求通过其查询目标页面得以表达，因此我们有必要对查询需求对应的目标页面类别进行区分。根据TREC网络信息检索子论坛多年来的工作积累，三类查询信息需求及其对应的目标页面之间的关系可以表述如表2-1所示（Hawking et al，2002）（Hawking et al，2003）。

表 2-1 查询信息需求类型与查询目标页面之间的对应关系

查询信息需求类型	检索任务	目标页面类型	用户需求比例
导航类	特殊需求页面定位	特殊需求页面	约 20%
导航类	主页定位		
信息事务类	主题过滤	关键资源页面	约 80%
事务类	主题过滤		

参考 Broder（2002）中的搜索引擎用户行为调研

上表中，导航类检索按照查询目标页面的类别不同细分为特殊需求页面定位任务和主页定位任务两类，主页定位任务的目标页面是站点/子站点的主页，而特殊需求页面定位任务的目标页面则是主页以外的页面。而由于信息事务类检索和事务类检索都具有查找目标不唯一且不确定的特点，因此对应的检索任务都是主题过滤任务，而其查询目标页面也都是关键资源页面。

关键资源页面可以理解为关于查询主题权威可信，对用户获取信息最有用的页面。如对于“mp3 下载”这个查询，http://mp3.baidu.com 或者 http://d.sogou.com/就是比较典型的关键资源。根据 TREC 网络信息检索子论坛的定义（Craswell et al，2004），关键资源页面应当是某个关键站点的入口页面，此站点需要提供关于某个主题的可靠信息。

需要特别指出的是，这里所提到的入口页面不一定是通常意义上的“主页”，它可能是大规模站点的接入页面，也有可能是某个子站点或者某一类页面集合的接入页面。如 http://www.nida.nih.gov/drugpages/marijuana.html 是美国药物滥用治疗研究所（NIDA）关于大麻方面信息的关键资源页面，但它并不是一个通常意义上的“主页”，这个页面只是为用户提供一个 NIDA 站点内部相关信息页面的访问索引。

从定义中可以看出，关键资源页面之所以“关键”，是因为它提供给用户一个关于某个主题的可靠信息的入口。用户通过关键资源页面，可以比较快捷的查找到所需要的信息。同时，某个主题的关键资源页面数要比其相关页面数少得多（相关页面动辄成百上千，而关键资源页面往往只有几个到十几个），这也方便用户将注意力集中到少数一些与自己的查询主题最贴切的页面上。

按照表 2-1 中查询信息需求类型与查询目标页面的对应关系，网络信息检索的目标页面一般分为特殊需求页面和关键资源页面两类。值得注意的是，尽管用户所提出查询对应的目标页面千差万别，但是就绝大部分的用户需求而言，其目标页面（特殊需求页面或关键资源页面）不可能是低质量、过时与不可信的页面，因为这些页面不符合用户通常的信息需求。

这说明，网页质量评估工作的核心，是定位出高质量网页即可能被作为用户某个查询的目标的页面，而页面质量的评价，应该与其成为查询目标页面的可能性直接相关。由于查询目标页面的类别只有两类，因此这就将网页质量评估工作转化成了针对特殊需求页面和关键资源页面的网页分类任务。

由于网页质量评估是网络信息检索工具处理网络数据的必要环节，网络数据的索引、用户需求查询等环节都需要依赖于质量评估的结果。从提高系统运行效率方面考虑，质量评估需要离线完成，即质量评估不能依靠即时的用户查询需求进行。然而，特殊需求页面和关键资源页面是针对特定用户查询需求的称谓，这又决定了页面质量评估不能脱离用户的查询信息需求存在。因此，页面质量评估的过程只能依靠对搜索引擎记录的用户查询行为进行分析和挖掘才能实现。

根据对用户行为的分析，定位查询目标页面的样例集合；进而对这部分查询目标页面的网页特征进行分析，借以找出此类页面与普通页面的特征差异；最后施行基于分类方法的页面质量评估，这是我们进行网页质量评估工作的总体思路，而寻找查询目标页面与普通页面的特征差异则是这个方法得以顺利施行的基础。

2.4 查询目标页面与普通页面的特征差异

按照第 2.3 节中的查询目标页面分类体系，我们将重点考察特殊需求页面及关键资源页面这两类页面与普通页面存在差异的网页特征。Web 数据的繁杂特性造成了这种特征不能仅仅从理论分析中得出，而应该从考察大规模网络数据统计特性的角度进行挖掘和分析。由于质量评估过程所依靠的网页特征是独立于即时的用户查询需求而分析出的，因此，这些特征也被称为“查询无关特征”。

2.4.1 特征分析的实验设定

对于查询目标页面和普通页面差异的分析是必要的：一方面，按照上文的论述，查询目标页面与普通页面的差异是本文数据清理工作的出发点；另一方面，这种分析对于我们了解查询目标页面的实质也是大有裨益的。

根据 Kraaij et al（2002）和 Upstill et al（2003），查询无关特征可以用来提高某些特殊类型检索的性能（如主题过滤、主页查找等）；而我们之前的工作也说明一些查询无关特征可以用于关键资源页面的挑选（Liu et al，2004）（Liu et al，2005）。在下文的分析中，我们借鉴了这些工作所采用的特征，并引进了一些新特征以发现查询目标页面和普通页面之间的差异，这些特征包括：

1. 文档长度：以词数（英文单词数/中文字数等）计算的文档长度；

2. 链接文本长度：链接到该页面的链接文字的长度总和，反映了此页面被多少个页面所链接，以及这些页面对该页面的重视程度；
3. PageRank：根据 Brin（1998）的标准算法计算出的反映用户访问该页面概率的数值；
4. 入链接个数：有多少个页面链接到该页面；
5. 出链接个数：该页面链接到多少个页面；
6. 站点内出链接个数：该页面链接到多少个本站点内部的页面；
7. URL 长度：按（Kraaij 2002）的描述，根据长度不同将 URL 分为四个类型，分别为 ROOT 型（域名），SUBROOT 型（一级目录），PATH 型（多级目录），FILE 型（其它）。

上述都是在以往工作中被证明可以一定程度反映页面质量的查询无关特征，为了比较查询目标页面与普通页面在这些特征上的差异，必须考察大规模语料库上两者之间的分布情况。.GOV 是 2002 年从.gov 域抓取的规模为 125 万页面的基于真实网络环境的语料库，共包括近 20G 的数据，并在 2002-2004 年的 TREC 评测中使用。我们之所以选择.GOV 而不是其它更大规模的语料库，原因在于基于.GOV 已经进行了多年的信息检索评测研究，有充分的被 TREC 评测人员手工标注的高质量页面充当后续工作的训练与测试集合。进一步的，.GOV 语料库来源于真实网络环境，在其上所进行的特征差异结论也应当普遍适用。在考察查询目标页面的查询无关特征方面，我们采用了如表 2-2 所示的训练集合。

表 2-2 查询目标页面训练集合的组成情况

	关键资源训练集合	特殊需求页面训练集合
页面集合来源	TREC2002-2003 主题过滤任务	TREC2001-2003 导航类检索任务
页面集合大小	327	860

由于 TREC2002 主题提炼任务的目标是查找关键资源站点所包含的页面而不一定是关键资源页面（Hawking et al，2002）（Hawking et al，2003），因此专门对这部分的答案集合进行了手工筛选，以找出标准答案页面对应的站点/子站点入口页面。

例如：TREC2002主题提炼任务的599号主题“scientific research misconduct”，原有的答案包括子站点http://ori.dhhs.gov/html/misconduct/内的24个页面，手工筛

选后，将这些页面用这个子站点的入口页面http://ori.dhhs.gov/html/misconduct/casesummaries.asp代替。

根据对表 2-2 的训练集合以及.GOV 页面的查询无关特征分析，我们发现查询目标页面与普通页面之间存在着较明显的特征差异，下面将介绍各查询无关特征对应的具体的实验结果：

2.4.2 文档长度特征的差异

文档长度特征在本文中是指 HTML 文档中所有文本（不含结构化信息文本）包含的词数。图 2-2 记录了查询目标页面与普通页面之间在文档长度特征上的分布差异。

图中，横坐标为文档词数，纵坐标为相应文档词数的文档个数在文档全集中所占的比例。

由图中可以看出，关键资源页面和特殊需求页面这两类查询目标页面在文档长度特征上的分布十分相似，而两者的分布曲线与普通页面有一定的差别。尤其是在文档长度较小时，查询目标页面基本没有分布：只有不足 1%的查询目标页面长度小于 1 千词。

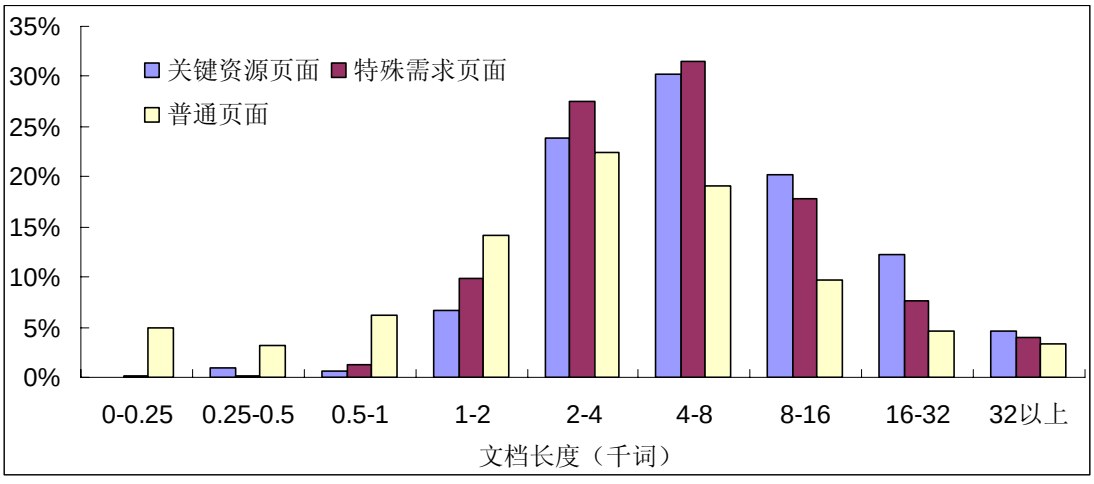

图 2-2 普通页面与查询目标页面在文档长度特征分布上的差异

页面长度较小的网页除掉垃圾页面外，很大一部分是单独提供图片或音频、视频浏览的网页，对于网页检索的需求而言，它们不可能提供太多用户需要的信息，因此查询目标页面的长度不会太短是合理的统计结果。

2.4.3 链接文本长度的差异

链接文本是描述页面间超链接的文字，如清华大学网站主页（http://www.tsinghua.edu.cn/）上链向清华大学教育基金会主页的链接，其描述文字“清华大学教育基金会”就是链接文本。具体到 HTML 代码中的形式为：

清华大学教育基金会

一般而言，按照第 1-2 节中所述的内容推荐假设，链接文本被认为是对所链向网页的较客观的描述，这是由于链接文本的作者与所链向页面的作者通常不同，因而有可能给出除网页本身作者之外的评述。链接文本长度一方面反映了页面被链接的多少，另一方面也反映了链接到它的页面对它重视程度的高低，因为作者一般倾向于用较完整的文字描述其认为重要的链接。这个特征的分布情况如图 2-3 所示。

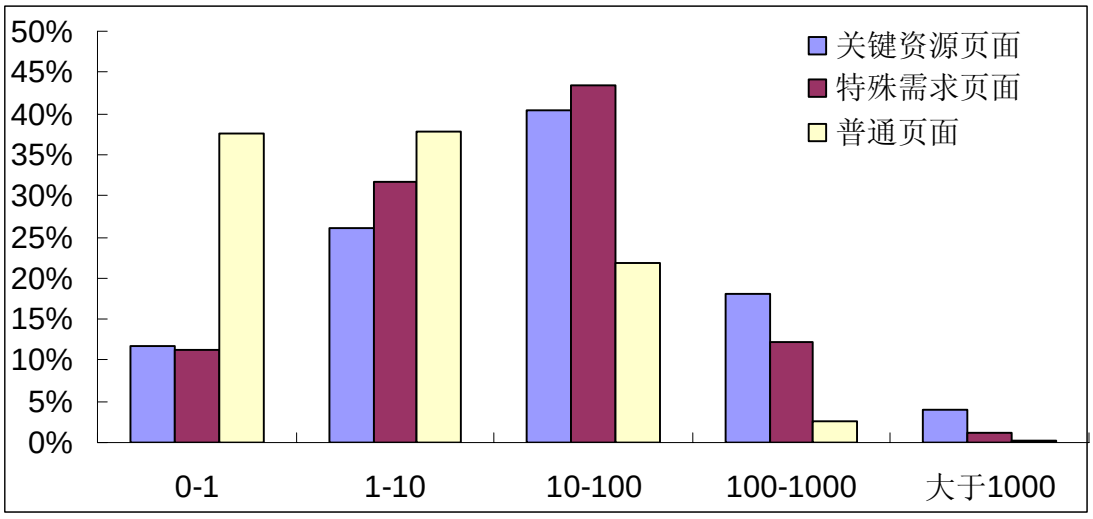

图 2-3 普通页面与查询目标页面在链接文本长度特征分布上的差异

图中，横坐标为网页对应的入链接文本词数，纵坐标为相应链接文本词数的文档个数在文档全集中所占的比例。

链接文本长度特征的分布情况证实了我们对这个特征作用的估计，即：链接文本长度反映了页面受其他页面重视的程度。两类查询目标页面在这个特征上的分布曲线也比较相似，关键资源页面的链接文本长度总体而言略大于特殊需求页面。而查询目标页面整体的链接文本长度则明显不同于普通页面，有超

过 3/4 的普通页面只有不足 10 个链接文本词，而这个比例在查询目标页面中则不足 40%。

2.4.4 PageRank 特征的差异

关键资源页面与特殊需求页面是用于满足不同检索需求的页面，就页面本身的特性而言，其最大的区别在于关键资源页面通常包含较多链向其他高质量页面的链接（为满足用户获取某方面信息的需要，提供延伸阅读的选择是必不可少的），而特殊需求页面则没有这方面的特征。这决定了两者在链接关系特征上会有较大的差别。

PageRank 特征是链接结构特征的一种，Google 搜索引擎的成功使之成为公众眼中搜索引擎核心技术的代名词之一。尽管不少研究结果如 Upstill et al（2003）证明 Google 本身使用的 *PageRank* 算法与其公布的算法有一定的出入，但 Brin et al（1998）的算法本身是一种有充分理论依据的比较可靠、可信、可行的算法。

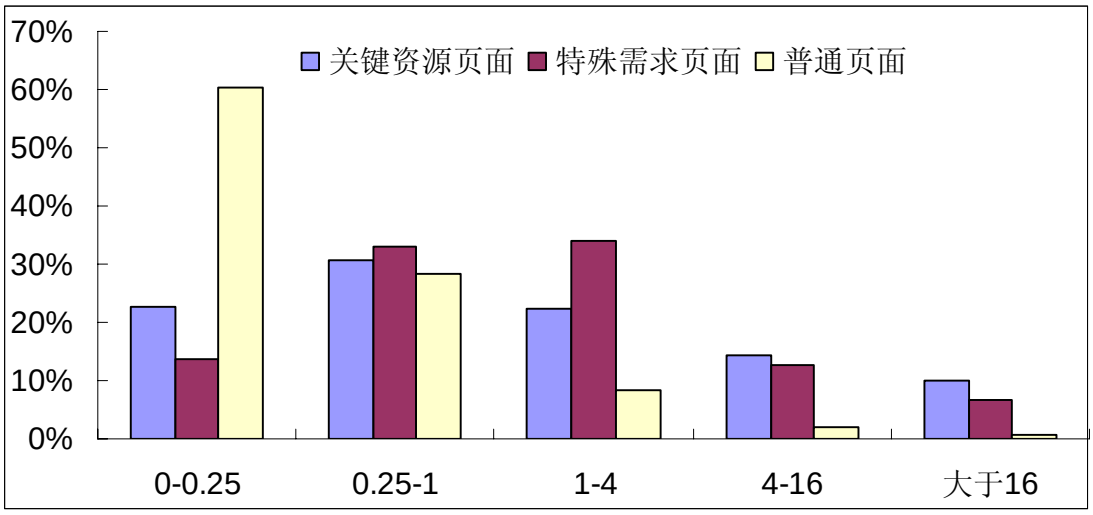

图 2-4 普通页面与查询目标页面在 *PageRank* 特征分布上的差异

我们使用传统 *PageRank* 算法（如第 1.2.4 节所述）对.GOV 语料库进行了链接关系分析，并考察了 *PageRank* 数值的页面分布情况如图 2-4 所示。图中，横坐标为网页对应的 *PageRank* 数值（已经进行归一化），纵坐标为相应 *PageRank* 值的文档个数在文档全集中所占的比例。

由图中可以看出，查询目标页面的 *PageRank* 数值大小比普通页面有明显的

优势，而特殊需求页面的 *PageRank* 总体而言比关键资源页面略大。如分别有 87%的特殊需求页面和 77%的关键资源页面其 *PageRank* 值大于 0.25，但普通页面的 60%相应数值都小于 0.25。

PageRank 数值一定程度上可以被看作页面质量的评价量度，因此查询目标页面作为被用户挑选的质量较高的页面具有较高的 *PageRank* 值是合理的。特殊需求页面之所以成为查询目标页面，是由于其被用户之前浏览过，而 *PageRank* 数值最直接的定义即是指页面被用户随机访问到的可能性，从这个角度讲，成为特殊需求页面的网页，其 *PageRank* 数值也应相对关键资源页面（用户不一定访问过）为高，才能成为导航类检索的目标。

2.4.5 入链接个数特征的差异

按照 1.2 节中提到的内容推荐假设，入链接个数直接反映了该网页被其他网页作者所推荐的程度。对于查询目标页面而言，能够被用户关注的页面往往也被其他网页作者所推荐（以便提高自己网页的链接质量），因此其入链接个数也一般较大。根据 Upstill et al（2003）的分析结果，入链接个数特征与 *PageRank* 特征在百万页面规模的语料库上进行分析时有类似的分布特性，我们的实验结果也证明了这一点。入链接个数特征在普通页面和查询目标页面之间的分布差异如图 2-5 所示。

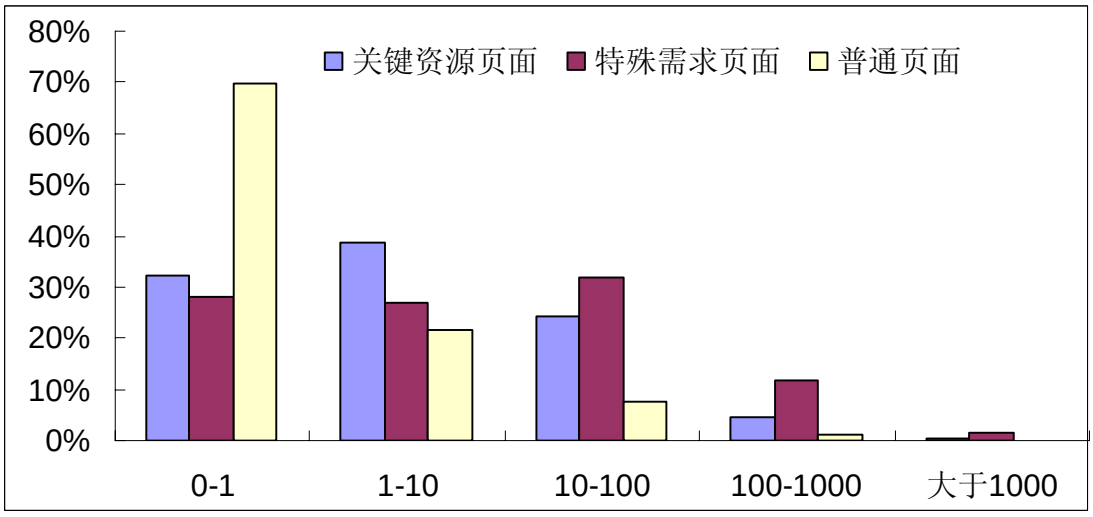

图 2-5 普通页面与查询目标页面在入链接个数特征分布上的差异

图中，横坐标为网页对应的入链接个数，而纵坐标为相应入链接数的文档个数在文档全集中所占的比例。

从图中我们发现有 68%的关键资源页面和 72%的特殊需求页面拥有超过 10 个入链接；而普通页面集合中对应的比率只有不足 30%。这说明查询目标页面的入链接个数从总体而言要多于普通页面，这可以由查询目标页面包含了大量的可靠信息、因此有大量的页面作者向自己页面的读者推荐这些页面的事实来解释。

2.4.6 出链接个数特征的差异

如果说入链接个数特征一定程度上反映的是页面的受欢迎程度的话，那么出链接个数特征则是反映了页面引导浏览者纵深阅读的能力。按照关键资源页面的定义，关键资源页面应当不仅本身提供信息，而且能够提供足够的相关网页链接，以保证用户能通过访问这一个页面总括性的了解相关主题信息，并能够进一步选择阅读。从这个角度来说，关键资源页面的出链接个数应当偏多。图 2-6 反映的是普通页面与查询目标页面在出链接个数上的差异，试验结果证明这个推断是正确的。

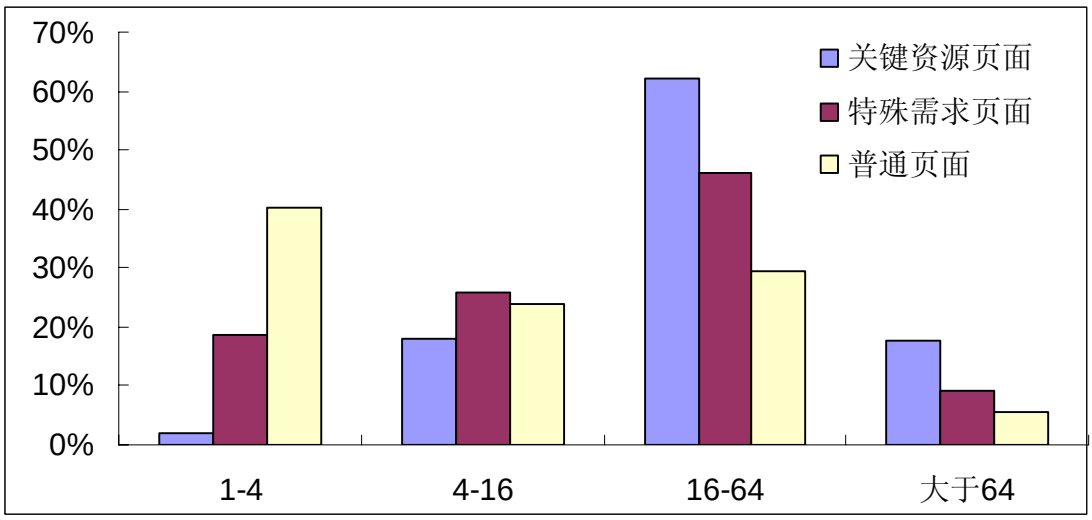

图 2-6 普通页面与查询目标页面在出链接个数特征分布上的差异

图中，横坐标为网页对应的出链接个数，而纵坐标为相应出链接数的文档个数在文档全集中所占的比例。

根据对数据样例集合的统计，关键资源页面的平均出链接数为 55，而特殊需求页面和普通页面的对应数值分别为 37 与 28。这说明关键资源页面的出链接个数明显多于特殊需求页面，而特殊需求页面的出链接个数略多于普通页面。图 2-6 也显示了类似的实验结果，几乎所有关键资源页面的出链接数都大于 4，而超过 40%的普通页面有不到 4 个出链接。

2.4.7 站点内出链接个数的差异

由于关键资源页面是关键资源站点的代表和用户访问的接口，因此它应当具有能够直接链接到本站点/子站点内的大多数页面，也即必须具有 Craswell et al（2001）提出的“导航功能”（Navigational function）。

为了更清楚的表述关键资源页面的这个链接特征，我们把页面出链接划分为指向站点内部的出链接和指向站点外部的出链接两类。由于关键资源页面首先是站点内部高质量页面的代表，而关键资源页面在整个出链接个数上已经有比较明显的分布优势，因此我们有理由认为关键资源页面在站点内部出链接个数上有更明显的特征。

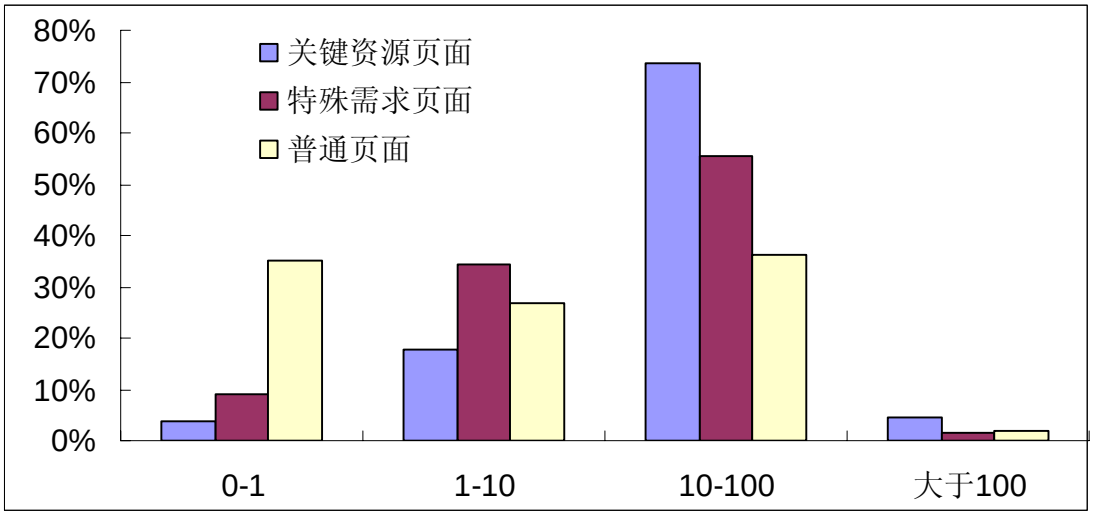

图 2-7 普通页面与查询目标页面在站内出链接个数特征分布上的差异

图中，横坐标为网页对应的指向站内页面的出链接个数，而纵坐标为相应出链接数的文档个数在文档全集中所占的比例。

图 2-7 中的试验结果证明了我们的推断，仅有不足 20%的关键资源页面站点内部出链接个数少于 10，而特殊需求页面中的相应比例是 44.7%，普通页面中的比例则是 64.4%。

2.4.8 URL 长度特征的差异

图 2-8 记录了查询目标页面与普通页面在 URL 长度特征上的分布。URL 长度特征是表示页面 URL 种类的一个非内容特征，由 Kraaij 等人提出，页面的 URL 划分为 ROOT，SUBROOT，PATH，FILE 四类，其中 ROOT 类的 URL 只包括域名，而从 ROOT 类到 FILE 类其 URL 中的“/”逐渐增多（Kraaij et al，2002）。

按照信息、事务类查询的检索需求，关键资源页面有较大的可能是网站或其子站点的首页，这就决定了其 URL 有较大的可能是 ROOT，SUBROOT 或 PATH 类型。而对于导航类检索中的特殊需求页面定位任务而言，特殊需求页面只是用户经常访问到的页面，它并不一定在 URL 上有明显的特征，因此其在 URL 长度分布上可能会与普通页面比较类似。

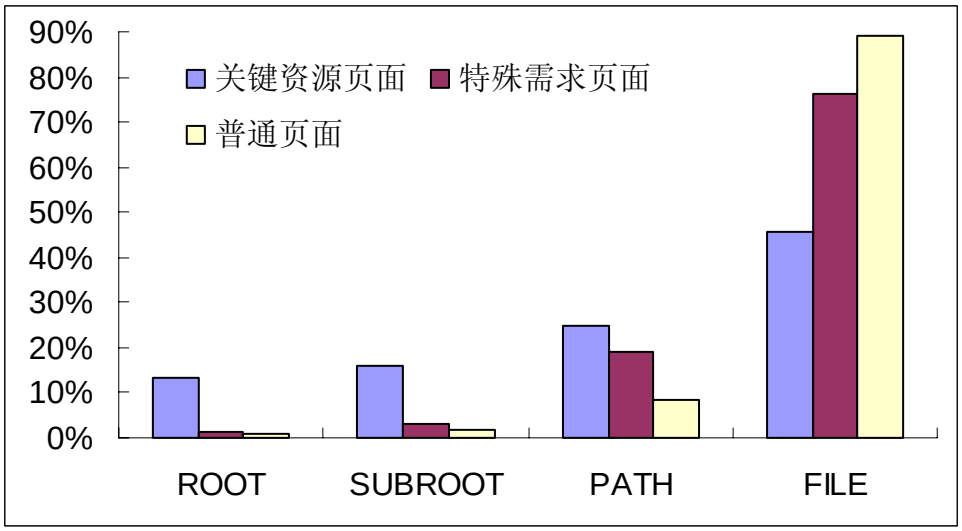

图 2-8 普通页面与查询目标页面在 URL 长度特征分布上的差异

图中，横坐标为网页对应的 URL 长度类型，而纵坐标为相应类型的文档个数在文档全集中所占的比例。

由图2-8可以看出，FILE类型的网页在.GOV语料库中占了大多数（近90%）；但在关键资源训练集中，FILE与非FILE类型的页面比例基本相等，非FILE类型

的页面反而更多一些。这说明关键资源页面的URL更倾向属于非FILE类型。相比较而言，特殊需求页面的URL长度特征分布则基本与普通页面相同，这与我们的推测是一致的。

2.4.9 查询目标页面的查询无关特征总结

本节以上所述的各个实验结果图示中，两类查询目标页面（关键资源页面与特殊需求页面）具有不同于普通页面的查询无关特征，而两类查询目标页面之间也存在着一定的特征差异。各类页面之间的统计意义上的分布差异情况可以参见表 2-3 所示：

表 2-3 查询目标页面与普通页面的相关系数比较

查询无关特征	关键资源页面 vs 普通页面	特殊需求页面 vs 普通页面	关键资源页面 vs 特殊需求页面
文档长度	0.3491	0.4878	0.9775
链接文本长度	0.7185	0.8198	0.9793
PageRank	0.5684	0.0695	0.8198
入链接个数	0.6517	0.5289	0.8858
出链接个数	0.2640	0.6728	0.8725
站点内出链接个数	0.5888	0.7226	0.9221
URL 长度	0.9639	0.9888	0.9923

表 2-3 中，相关系数是统计学研究中常用的数值标准，这里用来表示待比较的两组页面数据之间的相似程度，相关系数越小，说明两组页面之间的相似程度越小；相关系数越接近 1，则说明两组页面在这个查询无关特征上的表现越相似，其具体计算公式为：

$$\rho_{X,Y}=\frac{\sum_{i=1}^{n}(x_i-\sigma_X)(y_i-\sigma_Y)}{n\times\sigma_X\times\sigma_Y} \tag{2-1}$$

其中，σ_X 和 σ_Y 分别表示 X 和 Y 两组数据的平均值，而 n 为数据规模。

由表 2-3 的内容可以得出以下几个结论：

1．关键资源页面与普通页面的最大差异在于与出链接相关的几个特征上，关键资源中出链接/站内出链接个数较多的页面比例明显多于普通页面，相关系数的数值比较也说明关键资源页面在出链接特征方面与普通页面有非常大的不同。这是由关键资源页面在用户获取信息方面所发挥的作用决定的，根据我们之前的工作，关键资源页面本身不一定提供大量可用信息，但一般都能够提供较多高质量页面的链接（Liu et al，2004）（Liu et al，2005）。

2．特殊需求页面与普通页面的最大差异在于 PageRank 特征上。不同特征之间的相关系数比较中可以看出，PageRank 是区别特殊需求页面与普通页面的最重要特征。根据 Brin et al（1998），PageRank 反映的是互联网浏览用户访问到某个页面的概率大小。由于特殊需求页面一般都是用户曾经访问到的页面，因此它在 PageRank 数值上与普通页面有较明显的区别也就是合理的了。

3．特殊需求页面与关键资源页面的相似性远大于它们各自与普通页面之间的差异，如在文档长度、链接文字长度、URL 长度等特征上两类页面都表现得非常相似。在 PageRank 和出链接特征上两类页面有一定差异，这也验证了这两个特征分别是特殊需求页面与关键资源页面的主要特征。

综上所述，特殊需求页面与关键资源页面具有类似的查询无关特征，而这两类查询目标页面与普通页面之间的查询无关特征差异则较大。以上基于大规模语料库的实验结果说明，查询目标页面这一类与用户查询直接相关联的网页具有可以离线获得的查询无关特征。

查询目标页面具有查询无关特征是有其客观合理性存在的，查询目标页面能够满足用户某方面的信息需求，则其必定具有某些方面的内容（包括文本内容和链接内容）特征，因而在页面内部结构上具有某些特性；进一步的，其在 Web 环境中必然会引起其他页面作者的注意，使之具有特殊的链接关系特征。接下来，我们可以基于这些查询无关特征，采用机器学习方法完成在 Web 数据集合中定位这类可能满足用户信息需求的页面的任务。

2.5 基于查询目标页面定位的网页质量评估

根据第 2.4 节的内容，查询目标页面与普通页面之间存在着查询无关特征的差异，因此利用这些特征定位可能满足用户信息需求的查询目标页面，进而提

高检索系统处理网络数据的性能是可能的目标。然而使用何种机器学习和模式分类方法进行特征综合，并实现查询目标页面定位还是需要进一步考虑的问题。

2.5.1 查询目标页面定位问题的困境

查询目标页面分类问题与一切大规模数据上的网页分类问题一样，面临着训练样例获取的问题。任何机器学习算法必须依赖于一定数量的训练样例才能实现，但是对于面向网络数据的学习算法而言，大量的研究者如 Yu et al（2004）则证明主要的困难集中在如何获取有充足代表性的训练样例反例上。

由于查询目标页面有较为明确的定义，人们对这类页面也有较明确的感性认识，因此正例相对较易获得，特别是 TREC 的网络信息检索任务提供了大量的标准答案页面作为正例页面的可靠来源。相比较而言，反例的获得就困难得多，Web 页面多种多样，作为 Web 页面中绝大部分的非查询目标页面更是数目繁多，情况繁杂。内容不可信，内容太少，没有高质量的入链接或者出链接，有太多广告信息等等，都可能是页面无法成为关键资源页面的理由。Google 公司的 Henzinger 等人近年来多次指出：Web 页面的均一采样在当前是一个难以克服的困难（Henzinger et al，2003），因此获得有代表性的反例集合绝非易事。

既然获取反例是一项非常困难的任务，那么一个可行的实现查询目标页面分类就是选取那些可以不依赖反例信息进行学习的算法。尽管没有反例信息，但那并不意味着我们仅有正例信息可用，结合实际问题，往往有许多辅助信息可以让分类器变得更加有效。

1998 年，Denis 等人提出了基于正例和未标注样例的统计信息进行学习的模型，并证明了这个方法是 PAC（可能近似正确）可学习的（Denis et al，1998）。随后的工作中， DeComite et al（1999）和 Letouzey et al（2000）则验证了 k 阶合取范式及 C4.5 决策树算法都可以进行适当的变形以进行类似的学习。2004 年，我们的工作也说明了使用 ID3 决策树算法进行基于正例及未标注样例学习的有效性，当时在.GOV 数据集合上使用这个算法取得了关键资源分类的较好效果（Liu et al，2004）。然而，这些算法都有较强的依赖性，它们必须依赖于全部样例中正例比例的先验知识，而这种知识往往是不容易在实际应用中获得的。

单类 SVM 和 Yu et al（2004）提出的正例学习框架（Positive Example Based Learning framework，PEBL）都是基于 SVM 算法提出的解决上述问题的方案（Tax et al，2002）（Manevitz et al，2001），其中 PEBL 在相同条件下被理论和实践

都证明可以有更好的分类效果。但对于利用查询无关信息分类查询目标页面这个任务来讲，这两种算法都不是最好的选择，这是因为原有的算法框架都是为处理与文本内容相关的分类问题而设计的，其特征规模动辄成千上万，而我们要处理的任务特征数目在几个到十几个，特征数目在数量级上的差异使 PEBL 等算法不一定有效，进一步考虑到 SVM 算法本身在低维空间中的分类效果和训练效率不占优势，这促使我们寻找比现有算法更有效的解决查询目标页面分类问题的方法。

2.5.2 基于分类方法的网页质量评估及其局限性

针对查询目标页面分类的问题，我们提出过两种可以用于查询目标页面分类的算法，分别是基于改进决策树方法的网页质量评估算法和基于改进 K-均值方法的网页质量评估算法。两种算法的性能和适用范围各有差别，其共同点则是可以基于有限的正例和海量的未标注样本进行学习，实现查询目标页面的分类。

Liu et al（2004）提出了一种基于改进决策树算法的查询目标页面分类算法，算法的核心思想是利用正例样本和未标注样本共同估计反例样本的统计分布特征，以解决难以获得充分代表性的反例样例集合的问题。算法获得了比使用传统决策树算法在随机挑选的反例集合上进行学习更优的效果。

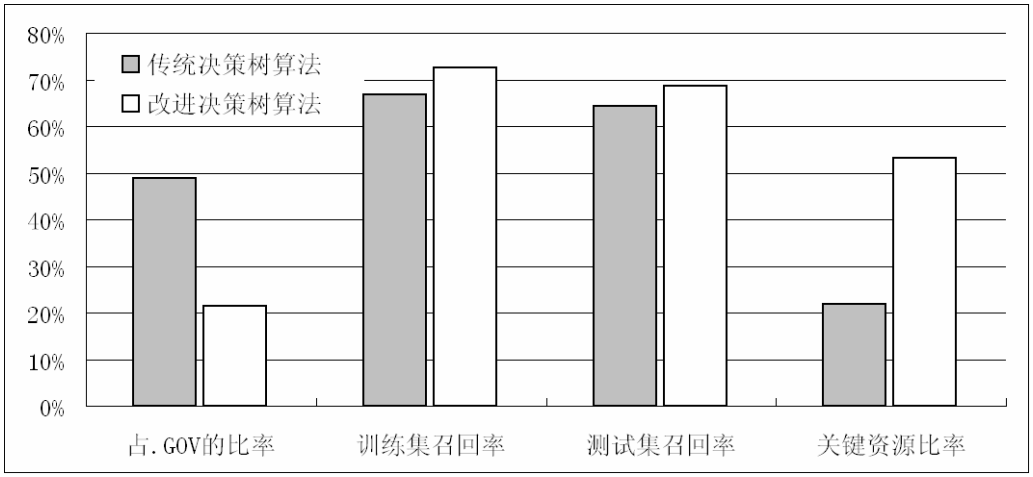

图 2-9 改进决策树算法结果集合的查询结果页面覆盖率情况

如图 2-9 所示，在基于 TREC2004 的实验集合上，改进算法分类得到的查询目标页面集合质量明显高于传统决策树学习算法。传统方法得到的结果集合包

含的页面数量是改进方法得到的结果集合的两倍多，但其关键资源页面集合的覆盖率无论是在测试集合还是在训练集合上却都低于改进方法得到的结果集合，改进决策树学习方法的关键资源覆盖率则是传统学习方法的 2.4 倍。这说明改进的决策树学习算法利用反例的统计性能能够比较理想的估计真实情况。而由于处理能力的限制，基于随机抽取少量反例的方法很难保证抽取到的反例样本的特性能够与真实的网络环境一致，这也验证了本节所述的查询目标页面反例样例获取困难确实存在。

尽管基于改进决策树的网页质量评估算法能够对可能成为用户查询目标的页面进行分类，但其需要对正例页面（即用户查询目标页面）在页面全集中的比例进行预估。预估的结果直接会对决策树的建立过程产生影响进而影响分类效果。这个数值对于大多数应用环境而言是难以给出准确的估计的，即使其可以作为系统参数进行调整，但参数调整的反馈过程会是一个繁杂的过程。

为了解决参数预估的问题，我们需要采用一个迭代的过程来修正可能被预估错误的参数，具体的步骤是：如果我们能对正例页面的比例进行一个初步估计 K(1)的话？利用这个初步估计得到的页面分类结果可以为我们提供一个进一步的正例比例估计 K(2)，以此类推，如果页面集合在整个特征空间中的分布比较适当，或说我们选取的特征空间比较合适的话就可以找到一个迭代方法逐步逼近正例页面真实比例的方案，这个方案类似于传统 K-均值聚类的方法（有些文献中也称为 C-均值聚类）。与传统 K-均值聚类算法不同的是，这个基于迭代方法估计整理页面比例的网页质量评估算法具有一定数量的正例样例作为训练集，因此不是一个完全的非监督算法，而可以被称为一个半监督学习方法。

我们基于以上思路提出了查询目标页面迭代分类的方法（Liu et al，2005）（Liu et al，2005b）。方法类似于 PEBL 的基于正例学习框架（Yu et al，2004），但将其中的 SVM 分类环节用较高效率的最小距离分类方法替代。算法可以将整个页面集合划分成查询目标页面集合和非目标页面集合两类。图 2-10 记录了算法在.GOV 页面集合上施行的情况。

图 2-10 中，横坐标为迭代次数，纵坐标为分类出的正例页面集合覆盖目标页面测试集合的比例（召回率数值），两条曲线分别表示起始时正例样例比例的估计值为 1/3 和 1/6 的情况。从图中我们可以发现，尽管两条曲线对应的初始情况不同，但都能在合理的迭代次数之内收敛到相同的收敛点，这也可以理解

为至少对于.GOV 数据集合而言，现有的算法和特征空间是可以完成检索目标页面分类的任务的。

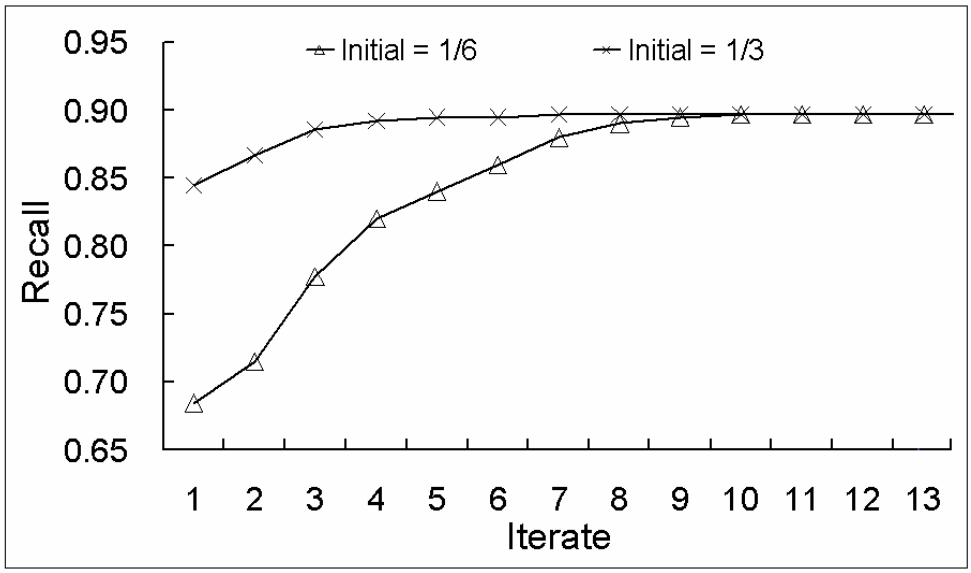

图 2-10 K 均值迭代算法在.GOV 页面集合上施行的情况

同时需要指出的是，基于 K-均值聚类的网页质量评估算法虽然不需对正例页面比例进行准确的预估（在一定范围内的初始预估值都可以收敛到真实值）；但其迭代算法的效率与真实网络信息检索工具环境的应用要求还有较大的差距。因此此算法更适合于进行针对一般规模封闭页面集合的页面质量评估，而不适用于大规模搜索引擎的使用。

综上所述，我们分别提出了基于决策树算法和 K-均值迭代方法的目标页面分类方法，其中决策树算法需要对查询目标页面在页面全集中的比例进行先验估计，而迭代方法的效率则无法满足网络信息检索系统实时性的需要，这两种算法因而难以在真实网络信息检索工具中得到广泛的应用。

从算法具有以上局限性的根本原因来看：以上算法基于查询目标页面分类的思路进行高质量页面的定位，因此需要获得充分有代表性的正反例信息。尽管算法对正例和未标注样本的统计特性进行了充分的挖掘，但仍然需要对正例样例比例这一参数进行估计。决策树算法直接进行参数估计，带来了其应用范围的局限；K-均值算法采用迭代收敛进行参数估计，则引起了算法效率的降低，同样难以获得广泛的应用。尽管使用查询目标页面分类的方法能够利用单类模式

分类方面充分的研究成果（如第 2.5.1 节所述的种种算法），但这方面的尝试却只获得了有限的成功，这促使我们寻找分类方法之外的页面质量评估算法方案。

2.5.3 基于贝叶斯学习方法的网页质量评估

在上节中，我们分别提出了两种基于决策树算法和 K-均值迭代方法的目标页面分类方法，决策树算法需要对正例页面比例进行先验估计，而迭代方法的效率则无法满足网络信息检索系统实时性的需要。考虑到贝叶斯学习方法的有效性，以及查询无关特征信息来源之间相互独立的有利条件，因此在本节中探讨使用先验概率学习的方法完成工作的可能性。

与前两种基于分类思路的网页质量评估算法不同，贝叶斯学习的方法通过计算网页成为查询目标页面的概率来实现页面质量的评估。跟据我们在 2.3 节中对网页质量的定义，这个概率反映了网页可能满足用户信息需求的程度，因而可以作为网页质量的评价指标。它形式定义为：

$$P(p \in Target\ \ page \mid p\ has\ \ feature\ A_1, A_2, ..., A_n) \tag{2-13}$$

在贝叶斯学习方法的推导过程中，我们首先推导仅考虑单个特征条件下的网页质量的表示方式，再考察多特征条件下的情况。

1．单特征条件下的网页质量表示方式推导

如果我们假定仅有一个特征 A 与网页质量相关，则网页质量可以用下式量化：

$$P(p \in Target\ \ page \mid p\ has\ \ feature\ A) \tag{2-14}$$

利用贝叶斯定理将上式展开，得到：

$$\begin{aligned} &P(p \in Target\ \ page \mid p\ has\ \ feature\ A) \\ &= \frac{P(p\ has\ \ feature\ A \mid p \in Target\ \ page)}{P(p\ has\ \ feature\ A)} \times P(p \in Target\ \ page) \end{aligned} \tag{2-15}$$

(2-15)式中，$P(p \in Target\ \ page)$ 即为查询目标页面在页面全集中的比例，如上文所述，估计此比例带来了改进决策树算法的困境，但在贝叶斯学习的过程中，由于仅需比较某个页面集合中 $P(p \in Target\ \ page \mid p\ has\ \ feature\ A)$ 的相对大小关系，而对于确定的页面集合而言，$P(p \in Target\ \ page)$ 是恒定的数值因此其并不影响相对大小关系的比较。因此在确定页面集合中，(2-15)式可以改写为：

$$P(p \in Target\ page \mid p\ has\ feature\ A) \propto \frac{P(p\ has\ feature\ A \mid p \in Target\ page)}{P(p\ has\ feature\ A)} \tag{2-16}$$

(2-16)式中，$P(p\ has\ feature\ A \mid p \in Target\ page)$ 可以用查询目标页面集合中的拥有特征 A 的网页比例进行计算，而 $P(p\ has\ feature\ A)$ 则等于网页全集中拥有特征 A 的网页比例。因此有：

$$\frac{P(p\ has\ feature\ A \mid p \in Target\ page)}{P(p\ has\ feature\ A)} = \frac{\#(p\ has\ feature\ A \cap p \in Target\ page)}{\#(Target\ page)} \Big/ \frac{\#(p\ has\ feature\ A)}{\#(CORPUS)} \tag{2-17}$$

其中，由于查询目标页面的种类仅有两种（特殊需求页面和关键资源页面），因此可以对其进行近似均一的页面采样。查询目标页面集合中的网页比例从而可以用查询目标页面样例集合中的相应比例来进行估计：

$$\frac{\#(p\ has\ feature\ A \cap p \in Target\ page)}{\#(Target\ page)} = \frac{\#(p\ has\ feature\ A \cap p \in Target\ page\ sample\ set)}{\#(Target\ page\ sample\ set)} \tag{2-18}$$

将(2-17)和(2-18)两式代入(2-16)式，可以得到：

$$P(p \in Target\ page \mid p\ has\ feature\ A) \propto \frac{\#(p\ has\ feature\ A \cap p \in Target\ page\ sample\ set)}{\#(Target\ page\ sample\ set)} \Big/ \frac{\#(p\ has\ feature\ A)}{\#(CORPUS)} \tag{2-19}$$

由于(2-19)式中右侧的各项都可以基于对语料库和查询目标页面的数值统计得到，因此根据此式计算单个特征情况下的页面质量是可行的。

2．多特征条件下的网页质量表示方式推导

现在考虑多个特征 $A_1, A_2, ..., A_n$ 共同影响下的网页质量评估方式。朴素贝叶斯定理保证下式成立：

$$P(p\ has\ feature\ A_1, A_2, ..., A_n \mid p \in Target\ page) = \prod_{i=1}^{n} P(p\ has\ feature\ A_i \mid p \in Target\ page) \tag{2-20}$$

如果能够通过选取信息来源不同的特征，则可以保证各特征之间近似独立（这一点在随后一节的大规模实验中将进行验证），即下式成立：

$$P(p\ has\ \ feature\ A_1, A_2, ..., A_n) = \prod_{i=1}^{n} P(p\ has\ \ feature\ A_i) \tag{2-21}$$

则多特征条件下的网页质量可以用下式表示：

$$\begin{aligned} & P(p \in Target\ \ page \mid p\ has\ \ feature\ A_1, A_2, ..., A_n) \\ & = \frac{P(p\ has\ \ feature\ A_1, A_2, ..., A_n \mid p \in Target\ \ page) P(p \in Target\ \ page)}{P(p\ has\ \ feature\ A_1, A_2, ..., A_n)} \\ & \approx \prod_{i=1}^{n} \frac{P(p\ has\ \ feature\ A_i \mid p \in Target\ \ page) P(p \in Target\ \ page)}{P(p\ has\ \ feature\ A_i)} \\ & = \prod_{i=1}^{n} P(p \in Target\ \ page \mid p\ has\ \ feature\ A_i) \end{aligned} \tag{2-22}$$

即多特征条件下的网页质量可以用多个单特征条件下的查询目标页面概率的成绩进行近似。如果将(2-19)式代入(2-22)式，则可以得到：

$$\begin{aligned} & P(p \in Target\ \ page \mid p\ has\ \ feature\ A_1, A_2, ..., A_n) \\ & \propto \prod_{i=1}^{n} \left(\frac{\#(p\ has\ \ feature\ A_i \cap p \in Target\ \ page\ sample\ set)}{\#(Target\ \ page\ sample\ set)} \Big/ \frac{\#(p\ has\ \ feature\ A_i)}{\#(CORPUS)} \right) \end{aligned} \tag{2-23}$$

网页质量与其成为查询目标页面的概率直接相关，而这个概率依照(2-23)式，可以基于正例页面集合和未标注页面集合中的查询无关特征分布情况进行计算。

基于语料库统计方法计算查询目标页面概率，并依照其大小对网页质量进行评估，这是我们基于贝叶斯学习方法进行网页质量评估的核心所在。

从算法的实用性上，由于算法不需进行高质量页面比率的估计，因此比较前文提到的决策树学习方法可以在更广泛的环境中得到应用；同时，算法判定页面质量的复杂度为线性复杂度（只需进行简单的逻辑判断和普通算术运算），比前文提到的基于迭代方法的页面质量评估具有高的多地算法效率，也适用于大规模网络信息检索系统的实际应用。

2.6 网页质量评估性能实验

2.6.1 性能评价指标

据我们所知，针对网页质量评估进行性能评价的工作非常有限，在我们描述基于分类的网页质量评估工作的论文中，由于把网页质量评估作为一个页面分类问题来处理，因此我们使用了传统的精度—召回率模型来衡量评估算法的

有效性（Liu et al，2004）（Liu et al，2005）（Liu et al，2005b）。精度（查准率）与召回率（查全率）的一般化定义可以参照第 1.1 节，在进行评价时，由于无法判断实验结果集合中所有的查询目标页面，因此在精度和召回率的计算中必须与分类算法一样，进行查询目标页面比例的估计。

引入如下两个假设：

1. 查询目标页面训练集合是页面全集中所有高质量页面的一个均一采样；

2. 查询目标页面占页面全集的比例为 R。

在上述假设下，召回率可以通过查询目标页面测试集在实验所得到的结果集合中的覆盖度来估计，即：

$$recall = \frac{\#(实验结果集合 \cap 关键资源测试集)}{\#关键资源测试集} \tag{2-24}$$

在精度的计算中，由于查询目标页面的总数可以用页面总数和 R 来计算，因此精度表达式为：

$$precision = \frac{\#(.GOV\ 页面集合) \times R \times recall}{\#(实验结果集合)} \tag{2-25}$$

由(2-25)式可见，利用传统的精度—召回率模型来进行页面质量评估算法的性能评价同样无法避免查询目标页面比例估计的问题。考虑到我们已经摒弃了基于分类思路的页面质量评估方法，因此有必要为页面质量评估引入新的性能评价方式。

考虑到基于学习的页面质量评估算法可以为页面集合中的每个页面赋以重要性权重（用页面成为查询目标页面的概率表示）。因此我们需要评价的实质上是一个按照权重进行排序的序列其排序的合理性，一个理想的质量评估算法应当把重要页面排在序列中靠前位置，而把低质量页面排列在序列中靠后的位置。具体到实验操作，较好的质量评估算法把查询目标页面测试集合中的页面排列在序列中尽量靠前的位置。

根据这个需求，我们定义如下评价指标“高质量页面平均召回率”(High quality page Average Recall, *AR*)：

$$AR = \sum_{i=1}^{\#(Original\ Set)} Recall(i) \Big/ \#(Original\ Set) \tag{2-26}$$

其中 *Recall(i)*表示序列中前 *i* 个页面组成的页面集合中的高质量页面召回率（用查询目标页面测试集的召回率近似），即：

$$Recall(i)=\frac{\#(HighQuality(i))}{\#(HighQuality\ Set)} \tag{2-27}$$

与 1.3.2 节中提到的平均精度（*AP*）的概念类似，高质量页面平均召回率是针对被评估页面序列的合理性的评价指标。

我们引入 $s=\frac{i}{\#(Original\ Set)}$，进一步对(2-26)式进行如下变换：

$$AR=\sum_{i=1}^{\#(Original\ Set)}Recall(i)\Big/\#(Original\ Set)=\sum_{s=\frac{1}{\#(Original\ Set)}}^{1}Recall(s) \tag{2-28}$$

其中 $Recall(s)$ 的定义与 $Recall(i)$ 相同，这样，当页面集合规模较大时，*AR* 就可以用下式表示：

$$AR=\int_0^1 Recall(s)ds \tag{2-29}$$

即在 *s-Recall* 图中，*AR* 的大小可以通过 *Recall(s)*图线下覆盖的面积进行计算。假设我们从页面质量评价排序最靠前的网页开始，遍历页面序列，则遍历过程中的某个时刻，*s* 的物理含义可以理解为我们已遍历的页面占页面总数的比例；而 *Recall(s)*则是我们已遍历的高质量页面占高质量页面总数的比例。

图 2-11 是页面比例—高质量页面召回率曲线的示意图。其中，横坐标是页面比例 s，而纵坐标则表示对应 s 的高质量页面召回率 Recall(s)。从图中，我们容易看出三条不同的曲线的面积之间满足大小关系 $Area(C1)>Area(C2)>Area(C3)$。因此，它们的所对应的页面质量评估算法的性能也满足 $AR(C1)>AR(C2)>AR(C3)$。

如果我们对三个算法的特性进行分析，则也可以得到类似的结论：C1 曲线对应的页面质量评估算法性能较高，因为其可以在保留很少原始集合页面的情况下覆盖相对较高比例的高质量页面（如 10%的页面集合对应的高质量页面召回率达到了约 35%）。而 C2 曲线对应的页面质量评估算法则类似于随机为页面赋以重要性权值，因为其对应的页面集合比例与高质量页面召回率比例呈线性

增长关系。类似的，C3 曲线对应的则是一个为高质量页面赋以低重要性权值的页面质量评估算法。

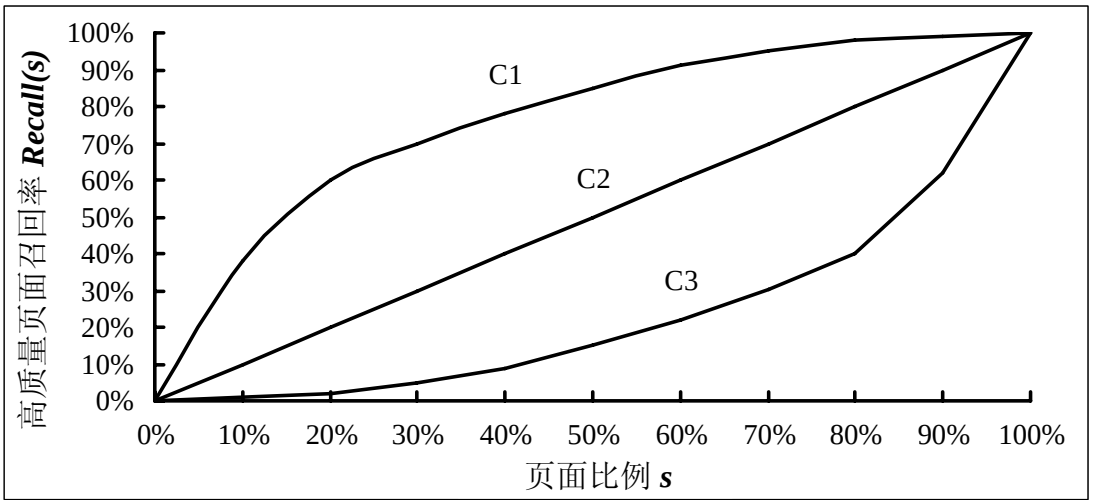

图 2-11 页面比例—高质量页面召回率(*s-Recall*)曲线示意图

依照(2-29)式，我们也可以推知，一个有效的页面质量评估算法，其 *AR* 值大于 1/2，一个挑选低质量页面的质量评估算法，其 *AR* 值小于 1/2，而 *AR* 值等于 1/2 的页面质量评估算法，事实上对页面随机赋以重要性权重。同样，我们也可以推知 *AR* 值的取值范围为(0,1)，*AR* 取值越接近 1 的页面质量评估算法越为有效，反之取值接近 1/2 的算法则效果甚微。

2.6.2 实验设定

我们基于两个不同规模和类型的语料库进行了网页质量评估的实验。它们分别是.GOV 语料库和 SOGOU 语料库，两个语料库的数据统计信息如表 2-4 所示：

表 2-4 .GOV 语料库与 SOGOU 语料库的数据统计信息

	网页数量	语言	所占磁盘空间	平均页面长度	域名限制	抓取时间
.GOV	1,247,753	英文	18.1G	15.2k	.gov	2002 上半年
SOGOU	37,205,218	中文	558.0G	15.0k	无	2005.11

其中，.GOV 是由 TREC（文本信息检索会议）收集的评测用网络语料库，TREC 在 2002 至 2004 年之间应用.GOV 组织了其网络信息检索论坛的评测项目。.GOV 的组织方式十分完备，由于网页局限在.gov 域中进行抓取，因此数据质量也较高，因此被许多信息检索研究人员应用于实验研究中。但其抓取的时间较早，网页抓取范围的限制也造成其无法很好的代表整个互联网数据的情况。尽管有这些不足，.GOV 毕竟是当前最容易获得和处理的网络数据语料库之一，我们基于.GOV 语料库进行网页质量评估算法的实验，具有便利其他研究人员进行比较实验的优点。

由于.GOV 无法很好的代表当前互联网数据的现状，更特别的，它没有涉及我们重点关注的中文互联网数据，我们于 2005 年底与 Sogou 公司配合，利用其搜索引擎抓取工具抓取了约 3700 万规模的中文网页数据用于实验，称为 SOGOU 语料库。根据 CNNIC（2006），中国互联网在 2005 年底时约有网页 8.7 个，我们所抓取到的页面占当时页面总数的约 4.3%。抓取也没有进行 URL 域的限制，可以说，SOGOU 语料库很大程度上反映了中文互联网环境近期的总体情况。尽管此语料库仅针对中文网页进行抓取，但只要我们所使用的查询无关特征不涉及中文本身的特性，则质量评估的方法还是可以考虑被移植到其他语言的网络环境中去。

总体而言，.GOV 语料库限制抓取域为.gov，其数据质量较高，而 SOGOU 则更能代表当前中文互联网的实际数据情况。

为进行网页质量评估方面的相关实验，我们为这两个语料库进行了查询目标页面样例集合的标注和整理工作。对于.GOV 而言，TREC2002-2004 的网络信息检索论坛积累了一部分由 NIST 人员手工标注的查询目标页面（如 2.4 节所示）。而对于 SOGOU 语料库，我们需要进行专门的查询目标页面标注。我们采用了 Hawking et al（2005）描述的 TREC 使用的 pooling 方法进行页面标注，其具体施行步骤如下：

1. 收集一定数量的有代表性的用户查询需求；
2. 利用各商用搜索引擎（如 Baidu，Google 中文，Yisou 等），对上述查询需求进行处理，并收集搜索引擎返回的结果；
3. 将各搜索引擎返回的结果构建结果池（pool），由评测人员评价结果池中的结果是否是查询目标页面；
4. 检查评测人员选择的查询目标页面是否出现在 SOGOU 语料库中。

我们从 Sogou 搜索日志中选取了 650 个覆盖各种类型（信息、导航、事务等）和各种内容（电影、电视、歌曲、软件、小说等）查询需求的用户查询，并利用其进行了查询目标页面的标注。最终构建的 SOGOU 查询目标页面集合共有查询目标页面 18630 个，我们随机选用其中 1630 个页面进行算法训练，并选用其余页面作为检验算法效果的测试集。

SOGOU 语料库及其对应的查询目标页面集合是我们所知最大规模的相关测试语料，这保证了在其上所进行的页面质量评估实验的可靠性。

2.6.3 网页质量评估实验

我们分别在.GOV 和 SOGOU 语料库上进行网页质量评估实验，选用的查询无关特征除第 2.4 节指出的特征外，还包括为中文网页质量评估设计的页面编码特征和页面镜像个数特征。

使用页面编码特征的原因是我们发现，SOGOU 语料库中绝大部分高质量网页的编码是 GBK 格式，而页面镜像特征则是针对中文网页中冗余页面数目较多的现象设计的。

为了验证 2.5 节中的查询无关特征独立性假设，我们将算法中使用的查询无关特征在 SOGOU 语料库中进行了特征相关度计算，如表 2-5 所示，每两个特征对应的相关系数值列在表内两特征交叉对应的位置。

表 2-5 查询无关特征之间的特征独立性

	URL 类型	编码	PageRank	镜像数	页面长度	URL 长度	入链接个数
URL 类型	1.00	0.15	0.15	0.01	0.04	0.10	0.00
编码		1.00	0.20	0.00	0.06	0.30	0.00
PageRank			1.00	0.01	0.06	0.03	0.05
镜像数				1.00	0.01	0.10	0.00
页面长度					1.00	0.04	0.00
URL 长度						1.00	0.02
入度							1.00

由上表可知，算法所采用的大部分查询无关特征之间相关程度很小，因此算法的特征独立性假设是可以近似成立的。值得注意的是，在 SOGOU 语料库中

PageRank 特征与入链接个数特征的相关度远远小于这两个特征在.GOV 语料库中的相关度，这验证了 Craswell et al（2005）指出的，PageRank 算法在页面集合规模扩大时数值特性变化较大的结论。

在 SOGOU 语料库上施行页面质量评估算法的 *s-Recall* 曲线如图 2-12 所示，其中虚线的部分表示 *AR*=1/2 即算法无效的位置。

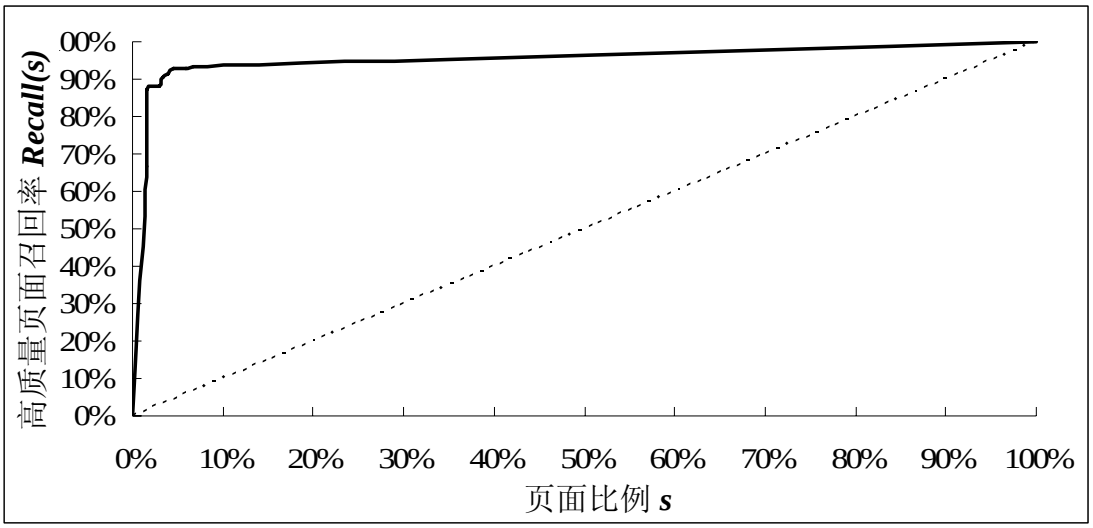

图 2-12 页面质量评估算法对应的页面比例—高质量页面召回率(*s-Recall*)曲线

由图 2-12 可知，我们在 2.5 节中提出的基于贝叶斯学习的页面质量评估算法是可行与有效的，其 *AR* 值为 0.9064，已经相当接近于 *AR* 值的上限 1。我们进一步将我们的质量评估算法与传统的基于链接分析的质量评估算法作了比较。基于链接分析的质量评估算法方面，我们选用了 PageRank 算法作为页面质量评估的标准，以此与综合使用各种特征的贝叶斯学习方法进行比较。

比较实验的结果如图 2-13 所示，其中横坐标分别代表：仅使用 *PageRank* 算法进行质量评估；使用贝叶斯学习方法，但不使用 *PageRank* 特征；使用贝叶斯学习方法，但不使用入链接个数特征；以及使用所有特征的贝叶斯学习方法。纵坐标为高质量页面平均召回率数据。

由图 2-13 可以看到，使用贝叶斯学习方法进行网页质量评估，比传统的利用链接关系分析算法有更优的性能。我们方法的 *AR* 值比仅利用链接关系特征的 *PageRank* 方法对应的 *AR* 值提高 6.62%。这说明：首先，单纯使用链接关系特征的页面质量评估效果不如综合使用各种查询无关特征的效果；其次，我们所采

用的贝叶斯学习方法能够有效的将各种查询无关特征进行综合，并获得比传统基于链接分析的质量评估方法更合理的结果。

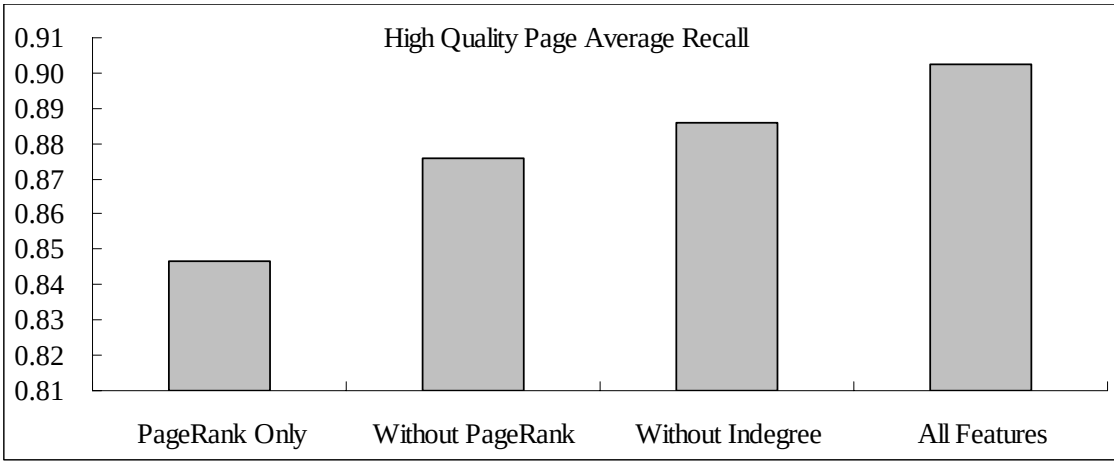

图 2-13 基于贝叶斯学习的页面质量评估算法与基于链接分析的相应算法比较

在图中我们还可以发现，即使算法不利用 *PageRank* 和入链接结构特征，也可以获得比单纯使用 *PageRank* 方法更优的效果。这也说明，基于贝叶斯学习进行页面质量评估依靠的不是某个特征的评估能力而是各方面特征的综合效果。这也符合 Google 公司的 Henzinger et al（2003）作出的“一种（比 PageRank）更优的页面质量评估算法应当不仅使用页面之间的链接信息，也应当利用页面内部的信息”的预测。

我们进一步考察了被应用于页面质量评估的各个查询无关特征的评估性能，具体的评价方法为：考察算法在不使用某个特征时的质量评估结果变化情况。如果算法不使用某个特征时，其性能出现明显下降，则说明此特征对于算法具有较大的影响；反之，如果算法不使用某个特征时的性能没有明显变化，则此特征对于页面质量评估的影响就不大。

具体的实验结果如表 2-6 所示。实验结果说明，链接结构特征（*PageRank* 与入链接个数）是算法中起作用最大的两个特征，但是这两个特征去除之后的实验结果也没有使整体结果有明显的下降（*AR* 值分别下降了 3.12%和 1.95%），这进一步明确了我们的结论，即页面质量评估的效果来自于各特征的综合效果和贝叶斯学习方法的使用。

表 2-6 查询无关特征在页面质量评估算法中发挥的作用

不使用的特征	丢弃特征之后的质量评估结果 *AR*
PageRank	0.8756
入链接个数	0.8860
URL 长度	0.8984
镜像个数	0.9012
页面长度	0.9031
编码	0.9032
URL 类型	0.9037

2.6.3 页面质量评估算法对检索性能的影响

我们最后考察了页面质量评估算法在.GOV 和 SOGOU 语料库上的综合性能，即被算法给予高评价值的页面集合能否取得较好的检索效果。理想情况下，网络环境中能够成为查询目标的页面只占少数，因此使用比较少量的页面取得与原有页面集合类似乃至更优的检索效果是可能的。

为了证实这个想法，我们首先对.GOV 和 SOGOU 两个语料库上页面质量评估结果进行了考察，分别考察其高质量页面召回率在 95%左右时对应的高质量页面集合比例。

实验结果如下表所示：

表 2-7 页面质量评估算法在.GOV 和 SOGOU 语料库上的不同性能

	页面集合占全集比例	训练集高质量页面召回率	测试集高质量页面召回率
.GOV	52.00%	95.53%	93.57%
SOGOU	4.96%	92.73%	92.37%

由表 2-7 可以发现，页面质量评估算法对.GOV 和 SOGOU 语料库都有较好的效果，但性能也有一定的差别。在保留训练集中约 95%高质量页面的情况下，.GOV 需保留的页面集合规模为 52%，而 SOGOU 需保留的页面为 5%左右。这首先说明我们提出的利用较小规模页面集合达到与页面全集类似检索效果的想法是可行的，.GOV 的约一半页面和 SOGOU 的约 1/20 页面就可以覆盖绝大多数的查询目标页面，从而保证检索效果；其次，实验结果也说明两个语料库的页面质量有较大差别，SOGOU 语料库的高质量页面密度远低于.GOV 语料库，

因此其平均的页面质量也要低得多。这与我们在介绍语料库时提到的情况是吻合的，由于.GOV 采集的时间较早，URL 又局限在.gov 域中，因此其页面质量较高是可以想见的。

既然可以利用较小规模页面集合覆盖绝大部分查询目标页面，则质量评估算法能够在减少索引规模的前提下对查询性能产生尽量小的负面影响乃至不产生负面影响也是有可能达到的，为此，我们基于.GOV 语料库和 TREC 的实验环境设计了如下实验。选择这个实验平台的主要原因是：实验平台便于获得，数据的公允性与可靠性强，研究结果易得到广泛承认。

我们分别考察施行页面质量评估算法后，排名约前 1/4，前 1/2 与前 3/4 的页面集合的检索性能。检索性能的评价使用 TREC 网络信息检索论坛 2003 和 2004 年的检索任务。2003 年 TREC 集中在考察主题过滤任务（如表 2-1 所述）的效果，而 2004 年则考察包括主题过滤、主页定位和特殊页面定位在内的三个任务的综合效果。检索结果的评价方式采用了 TREC 通用的 MAP 和 B-pref 评价方式（参见 Buckley et al（2004）），这两个评价指标被包括 TREC 在内的网络信息检索研究所广泛采用。

检索实验采用了 Robertson et al（1994）所提出的 BM2500 方法和默认的系统参数，实验结果如图 2-14 所示。

图中，我们考察了检索性能随页面集合规模变化的趋势，我们发现，无论对于 TREC2003 还是 TREC2004 的检索任务而言，使用较小规模的页面集合取得比原有页面集合更优的检索性能都是可行的。对于 TREC2003 任务，页面质量排名前 1/4 的页面取得了比页面全集优越的多的检索效果（MAP 提高 9.93%，B-pref 提高 28.30%）；对于 TREC2004 任务，去除页面质量排名后 1/4 的页面后，检索结果的 MAP 提高了 1.25%，B-pref 提高了 3.65%。

从检索系统实际应用的角度考虑，只要根据质量评估结果得到的小规模的页面集合检索结果比较页面全集没有明显的下降，即可以认为质量评估算法是值得应用的，因为它毕竟能够在保证检索质量的前提下减少搜索引擎需要索引的页面数量。但我们的实验结果说明，页面质量评估算法甚至能够在.GOV 这样已经拥有较高页面质量的语料库中帮助检索系统提高检索性能。这充分说明了质量评估算法的巨大应用前景。

从图 2-14 中，我们还可以发现：对 TREC2003 的检索任务而言，检索性能在页面集合约占原始数据 1/4 时达到最优值，而对 TREC2004 而言，检索性能最

优值出现在占原始数据规模 3/4 的页面集合上。这种差异来源于两个年度不同的检索任务设置，对于 TREC2003 而言，主题过滤任务的查询目标页面是关键资源页面，且查询目标页面的个数不唯一（一般在 3-5 个）；而 TREC2004 中相当规模的查询对应的是导航类查询任务，其目标页面唯一，一旦在保留较小规模页面集合时，将查询目标页面与普通页面一起丢弃就会造成极大的性能损失（该查询主题 MAP=0），因此需要用更大规模的页面集合以保证其查询目标页面能够被覆盖到。

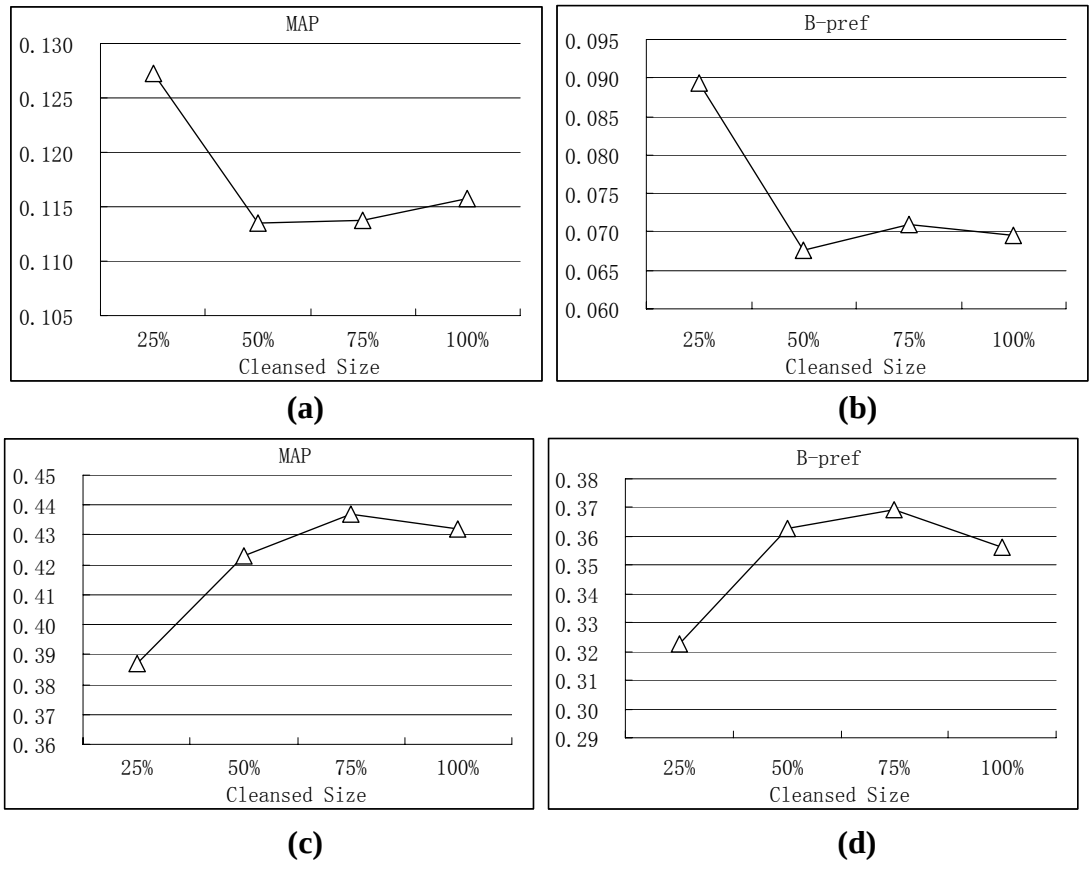

图 2-14 页面质量评估算法对检索性能的影响（a,b：基于 TREC2003 实验平台的结果，c,d：基于 TREC2004 实验平台的结果）

因此我们可以发现，信息检索系统的检索性能决定于以下两个因素：

1. 页面集合规模

利用页面质量评估算法，有可能将页面集合中的低质量页面去除，因此合理减少页面集合的规模有可能提高检索性能。例如，对于 TREC2004 的第 26 号

查询“*The White House President Bush's cabinet*”而言，在进行针对页面全集的检索时，.GOV 语料库中的第 G45-22-1096484 号文件被返回在结果列表的第三位，认为是一个相关结果。但事实上此页面仅仅提到了查询主题中的一些词句而已，其实并未提供与查询主题相关的其它信息，尽管如此，它处在结果列表如此靠前的位置，却明显影响了其它更相关结果被用户关注。而在进行页面质量评估之后，此页面被清除，因此整个主题的检索性能也得到了提高（总体而言 MAP 从 0.149 提高到了 0.500）。

2. 查询目标页面召回率

查询目标页面召回率对检索系统性能会起到至关重要的影响，一个在页面质量评估过程中被摒弃的页面是绝不可能被检索系统返回的，因此一旦查询目标页面被丢弃，就意味着对应查询主题的检索性能会明显下降。对于查询目标页面唯一的导航类检索任务而言，这种影响会更加明显，这一现象也在 TREC2004 的检索实验中得到了证明。因此层次式的索引结构（将质量高的页面置于高速索引中，同时不抛弃质量低的页面，只是将其置于低访问速度的索引服务器中）可能是更适合的信息检索系统处理网络数据的解决方案。

以上两个因素在决定检索系统性能的过程中同时发挥作用，如果查询目标页面容易被质量评估算法给予较高评价，则查询目标页面召回率相对保持较高，页面集合规模起主要作用。如主题过滤任务中，页面集合较小时检索性能更优。反之，当查询目标页面不容易被质量评估算法辨认或查询主题对应的查询目标页面唯一时，则查询目标页面召回率成为主要的决定性因素，如特殊需求页面检索任务中，保留的页面集合就需要较大以保证较高的查询目标页面召回率。

如果我们深入分析 TREC2004 检索实验中不同类型查询对应的检索性能，则可以进一步验证这一质量评估算法影响检索性能的机理。具体的实验结果如图 2-15 所示。

从图中所示的试验结果中我们可以看到，TREC2004 实验结果中信息事务类查询的检索实验结果与 TREC2003 的实验结果曲线十分类似，都是在页面集合规模较小（原集合规模 1/4）时取得较好的检索效果；而导航类查询的检索实验结果则依然是在原集合规模 3/4 时取得最有检索效果。由于 TREC2004 的任务设置中，导航类查询的数目是信息事务类查询的两倍，而且导航类查询对应的 MAP 性能较高，因此总体的实验结果基本还是呈现导航类查询的检索性能变化趋势。这再次说明丢失查询目标页面对于导航类查询的影响要远大于其对信息事务类

查询的影响。

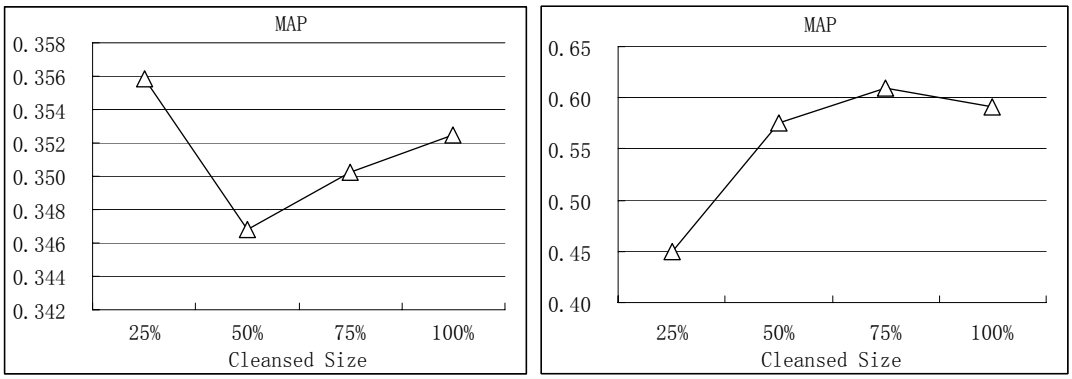

图 2-15 页面质量评估算法对检索性能的影响机理考察（a：信息事务类查询主题对应的检索效果；b：导航类查询对应的检索效果）

2.7 本章小结

网络数据处理是网络信息检索系统中极为重要的环节之一，直接决定了其利用网络数据资源的能力。在网络数据处理环节引入用户查询需求分析的方法，使检索系统有限的存储和处理能力集中在对用户查询需求有效的网页上，是本章的核心内容所在。

为了实现提高检索系统处理网络数据能力的目的，我们设计了基于用户群体行为分析的网页质量评估算法。首先，我们考察了网络数据质量评估的已有工作并将我们工作的重点集中到宏观粒度的网络数据质量评估即网页质量评估上。其次，我们根据用户查询的实际需求将网页的质量定义为其成为查询目标页面的可能性，并利用对大规模网络语料的分析证明查询目标页面具有与普通页面不同的查询无关特征。随后，我们利用贝叶斯学习的方法对查询无关特征进行综合，并以此计算页面成为查询目标页面的可能性。

基于对海量规模网络语料（占中文互联网网页总数的4%以上）的网页质量评估实验证明，基于贝叶斯学习的质量评估算法能够有效的实现页面质量评估，其评估结果比传统的基于网络链接结构分析的评估算法更为合理。在TREC网络信息检索实验平台上的检索实验结果也说明，网页质量评估可以用于提高检索系统的性能，利用较小规模的高质量页面集合达到与页面全集类似乃至更优的

检索结果是完全可能的。

综上所述，我们基于用户群体行为分析的思路实现了一种有效的网页数据质量评估算法，此算法不仅能够准确的评估页面成为查询目标页面的可能性，而且能够在减少索引规模的前提下实现网络信息检索系统的检索性能提高。

第三章 基于用户群体行为分析的用户查询行为研究

3.1 引言

本文所要解决的主要问题，是搜索引擎现有信息交互方式造成的用户信息需求传递瓶颈问题。在用户查询处理方面，这种瓶颈主要体现在用户的信息需求无法被合理的理解，因此搜索引擎只能会与用户查询词直接相关的内容，而无法根据用户需求调整结果内容。

为此，我们希望能够借助对用户行为的分析，挖掘用户行为模式与其信息需求之间的内在联系，借以根据用户行为判断其信息需求，提高网络信息检索工具的性能。在进行这方面工作之前，我们有必要对搜索引擎可能收集到的用户行为信息作一个归纳。

首先让我们考虑搜索引擎用户获取信息的行为流程，我们给出一个简单的搜索引擎用户的行为流程示意图，如下所示：

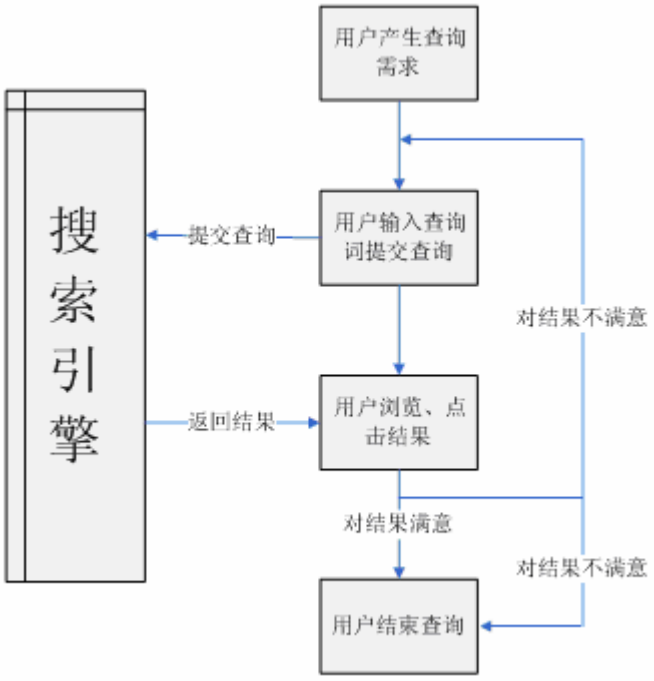

图 3-1 搜索引擎用户行为流程示意图

根据图 3-1 所示，用户产生查询信息需求之后，首先向搜索引擎提交信息需求对应的查询词，而后根据搜索引擎返回的结果进行浏览和点击，如果对结果满意，则结束查询流程；否则可能修改查询词，或直接结束查询选择其它的信息查找方式。

针对这样一个用户行为流程，搜索引擎作为一个网络服务提供者可能收集到用户的行为信息，而这些信息则是我们在不改变用户行为习惯的前提下进行行为分析最可靠的依据。具体而言，这些信息可能包括：

表 3-1 搜索引擎可能采集到的用户行为信息

名称	记录内容
Query	用户提交的查询
URL	用户点击的结果地址
Time	用户点击发生时的日期、时间
Rank	该 URL 在返回结果中的排名
Order	用户点击的顺序号（这是用户点击的第几个页面）
Session ID	用户某次使用搜索引擎过程对应的标识号
Submitter information	浏览器信息，计算机信息

其中，Session 定义为用户打开浏览器进行查询直至其结束查询并关闭浏览器的全过程，而 Session ID 可以理解为用户某次使用搜索引擎时的唯一标识，他可以通过浏览器记录的信息对搜索引擎进行反馈，而搜索引擎可以用其来辨识某个用户一次使用搜索引擎的全过程。

利用上述信息，我们就可以对搜索引擎用户在图 3-1 中的行为进行一定程度的复原，从而分析用户行为与其真实信息需求的关系。

3.2 网络信息检索用户行为特性

针对用户行为的分析和挖掘必须具备相当规模的搜索引擎用户行为信息资源，我们在 Sogou 搜索引擎的协助下，收集了其 2006 年 2 月全月的用户日志，经过对日志的分析，发现其中包括了绝大部分表 3-1 中所涉及的用户信息内容。因此我们可以基于这部分日志信息，进行用户行为特性的分析，并进而考虑利用用户行为特性中的有利资源进行信息需求提取。

我们收集到的日志信息包括了 Sogou 搜索引擎在 2006 年 2 月 1 日至 2006 年 2 月 28 日的 28 天时间内的所有用户查询和点击信息，原始数据规模超过 50G。其中，非空查询共 45,745,985 个，非重复的非空查询共 4,345,557 个，而 Session 个数为 26,255,952 个。

3.2.1 用户用多少词来表示查询需求

我们首先来考察用户查询长度特性，不涉及分词问题，如果我们把用户使用空格间隔开的部分称为“词”，并用词数表征查询长度，则可以得到如下统计信息：

表 3-2 搜索引擎用户查询长度特性

查询长度	占非重复查询总个数的比例
1	44.37%
2	36.22%
3	12.56%
大于 3	6.85%

可见，绝大部分（超过 80%）用户仅使用 1-2 个词进行查询，而使用超过 3 个词以上进行查询的仅占不足 7%。而查询的平均长度为 1.85 个词，这说明用户输入的查询通常都比较短。这个结果与 Silverstein et al（1998）针对英文统计的结果（2.35 个词）有一定差异，这应该主要是由中文和英文两种语言的差异造成的。而在中文搜索中，查询的平均长度更短，这就使中文搜索引擎得到的用户需求信息更少，需要对用户需求有更多的分析和经验，才能更加准确地返回用户需求的信息。

3.2.2 用户是否修改他们的查询

从对查询长度的分析可以看出，用户仅使用很少的词语来表示其信息需求，这显然为搜索引擎反馈合理的查询结果带来了很大困难。CNNIC（2005）报告，用户使用搜索引擎遇到的最大问题是其查到的“重复、陈旧信息太多，有用信息太少”，这相当程度上是由于搜索引擎返回的仅仅是与查询词相关的结果而非最满足用户需求的结果造成的。那么，在进行搜索的过程中，用户又是否会

通过修改查询的方式更合理的表述其信息需求，以获得更好的查询结果呢？我们对用户这方面的行为习惯进行了如下考察：

表 3-3 不同 Session 中用户进行查询次数的分布

查询次数	占 Session 数的比例
1	66.46%
2	17.36%
3	7.33%
大于 3	8.85%

表 3-3 中，我们分析了用户在某次检索（用 Session 进行区分）中所涉及到的查询次数的分布情况，由表可见，大部分的用户（66.46%）在进行检索时不曾改变查询词。而改变查询词的用户是对原有查询进行修改，还是进行另一个内容的查询呢？我们又进行了更深入的分析：

表 3-4 用户修改查询的方式

修改方式	占用户查询修改次数的比例
增加查询词	9.00%
减少查询词	1.43%
彻底修改查询	83.27%
其他修改方式	6.30%

表 3-4 中，我们分析了用户在进行多次查询时，其查询词改变方式的分布情况，由表可见，83.27%的用户直接使用一个新的查询代替原有查询，这基本上是属于进行其它内容查询的范畴，而不是对原有查询的修改。

3.2.3 用户行为特性对查询处理的影响

综上所述，一方面，用户使用的查询词数很少；另一方面，用户又不几乎不进行查询词的修改，这造成搜索引擎获得的用户查询很难真实的反映用户的信息需求。但用户是如何通过搜索引擎返回的并不完善的结果获取有用信息的呢？经过对搜索引擎日志的分析，我们发现，用户在 2006 年 2 月共对 Sogou 搜索引擎的结果进行了 86,538,613 次点击，提交了 45,745,985 个查询，平均每个 Session 内输入了 1.7 个查询词，但却要进行 3.3 次点击。这说明，用户倾向于用

更多次的点击来选择出符合自己信息需求的结果，而不是输入更多的查询词，或根据结果调整自己的查询形式。

用户这种“懒于输入，勤于点击”的行为习惯事实上加重了搜索引擎获取用户信息需求的瓶颈问题，但也为我们带来了宝贵的用户点击行为信息，有助于利用点击行为分析提高搜索引擎的查询处理能力。

3.3 研究背景：查询分类在网络信息检索中的作用

第 3.2 节中，我们对搜索引擎用户行为特性进行了分析，用户“懒于输入、勤于点击”的行为习惯造成其查询的信息需求很大程度上反映在点击行为日志中。但由于用户的查询需求内容繁杂，数量众多，我们不可能针对每个用户及其查询实时分析信息需求。因此将用户查询需求分类，利用用户行为特征判断查询需求的类别，进而指导网络信息检索系统查询处理的过程，是一个可行的解决方案。但这必然对查询分类体系提出更高的要求：一方面要便于基于用户群体行为分析的方法进行分类，以保证分类精度；另一方面也要能够对应不同的类别设计专门的查询模型和检索策略，以保证这种类别的信息需求能得到更好的满足。表 2-1 所列出的查询需求类型，即信息事务类与导航类的两类分类体系基本满足这两个要求，因此为我们所采用。

将用户查询的信息需求分成信息事务与导航两个类别是研究界和产业界比较通行的做法，在搜索引擎算法设计和性能评估中都得到了广泛的应用。分类体系本身就是 IBM 公司 Broder 和 Yahoo!公司 Rose 等基于实际应用需求而提出的。而从 2002 至 2004 年，TREC 网络信息检索论坛也一直基于这个分类体系进行评测任务设置，分别举行过主题过滤任务评测、主页定位任务评测、特殊需求页面定位任务评测和查询类别辨识评测等项目。

在第 2.1 节，我们指出信息事务类与导航类查询的不同信息需求特性，信息事务类一般没有确定的用户查询目标页面，而导航类查询则通常具有确定的目标页面。这种信息需求特性差异也造成了网络信息检索系统在处理两种类型查询时需要采用不同的检索模型和排序策略。

Craswell et al（2003）和 Kraaij et al（2002）指出，链接文本内容和 URL 类型特征在进行主页型的导航类检索时会有很大帮助。而 Zhang et al（2002）的工作也说明，基于链接文本特征进行强调精确匹配的内容检索的方法会提高特殊

需求页面定位任务的性能。与之相对应，Bharat et al（1998）和 Hawking et al（2003）的工作证明，采用链接结构分析方法能够帮助主题过滤任务提高性能。总体而言，精确匹配的检索策略（如基于二元组关系的检索）、链接文本信息和 URL 信息对提高导航类检索的性能有利；而相对模糊的检索策略、查询扩展方法和链接结构信息则会提高信息检索系统处理信息事务类查询的能力。

信息事务与导航类查询需要采用不同检索策略的根本原因在于其对应的用户信息需求不同，如果我们能够利用用户群体行为分析方法及时得到用户查询所对应的信息需求类别，则采用对应的检索策略会大大提高检索系统查询处理的性能，从而提高用户的使用体验。

3.4 用户查询需求的特征分析

2002 年以来，导航类和信息事务类的查询分类框架及其对提高信息检索系统查询处理性能的作用逐渐被研究人员所认识，也出现了一些基于查询自身特征和用户行为特征进行查询需求分类的工作。TREC2004 专门设计了查询需求分类任务，参与者需要利用查询自身特性和检索系统的反馈信息分辨 225 个查询分别是信息事务类别、主页定位类别还是特殊需求页面定位类别；Kang et al（2003）利用链接文本信息和词性标注信息进行查询需求分类，并基于此分类算法指导了检索策略改进；Lee et al（2005）在查询分类中引入用户群体行为分析方法，他们基于 UCLA 的学生群体采集用户行为并进行查询分类，在小规模的实验环境取得了较大的成功（分类成功率约 90%）。

以上工作考察了查询自身特征、链接文本信息、用户行为信息和检索系统反馈信息在查询需求分类中的作用。但总体而言，以往工作缺乏对特征信息的可用性、易用性和分类能力深入的考察，实验数据也与网络信息检索工具的真实应用环境有较大的差别。为了探索一种能够满足真实应用需求的查询需求分类算法，我们基于 Sogou 搜索引擎日志进行了查询分类所依据特征的分析考察。

分析依据的实验数据仍为第 3.2 节介绍的 Sogou 搜索引擎 2006 年 2 月份的用户日志信息。不同的是，我们对其中部分查询的信息需求类别进行了标注，以用于分类算法的训练与测试。查询样例训练集合和测试集合的具体数据及来源如表 3-5 所示。为增加分类效果测试的可靠性，避免标注人员的主观影响，我们使用不同的方式生成查询样例的训练与测试集合。训练集来自于 Sogou 公司

标注人员对查询频度较高的部分查询词的标注结果；而测试集则分别采集于天网（http://e.pku.edu.cn）组织的搜索引擎评测项目和网络上较常用的网页目录Hao123.com。基于这一日志数据和训练样例集合，我们分别考察了可能被用于搜索引擎查询需求分类的各个特征的可用性、易用性和分类能力。

表 3-5 查询样例训练集合与测试集合

	训练集合		测试集合	
类别	导航类	信息事务类	导航类	信息事务类
数量	153 个	45 个	152 个	81 个
来源	手工标注	手工标注	Hao123 网站	天网搜索引擎评测项目

3.4.1 查询内容特征

查询内容特征是最容易获取的查询需求分类依据，但仅仅依靠查询内容实现需求分类是不可靠的作法，TREC2004 的研究结果说明，查询长度对于实现查询需求分类有一定的帮助，而 Kang et al（2003）则报告查询词的词性对提高分类效果也有作用。他们的工作基于如下想法：进行导航类查询时，用户由于曾经访问过目标页面（或听人描述过），因此倾向于使用较多的查询词来表述自己的查询需求，且这些查询词多为名词词性。但我们针对中文搜索引擎用户行为的分析发现，这个现象在中文检索中并不明显。

下图给出了不同信息需求类别的用户查询对应的查询长度的分布情况：

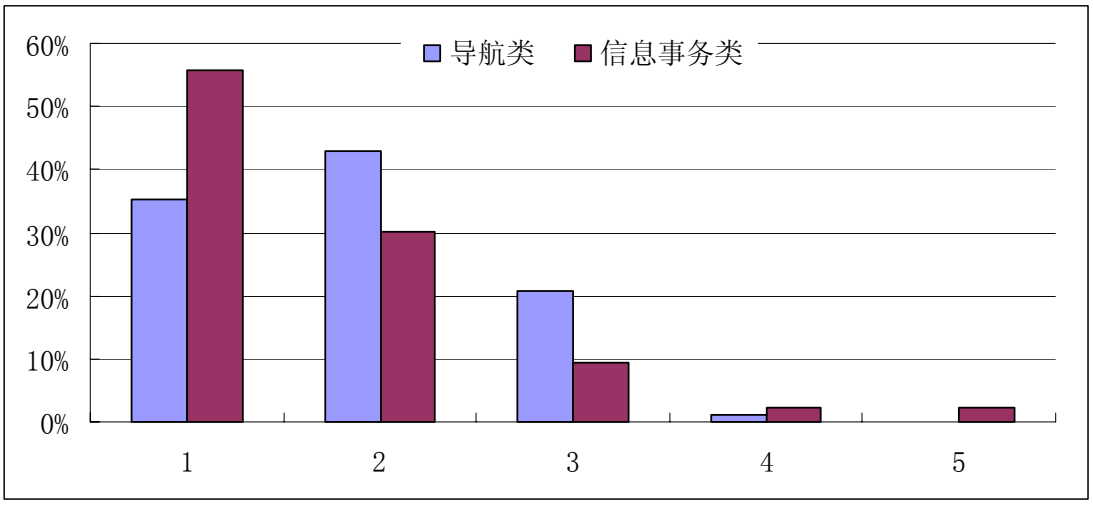

图 3-2 不同信息需求类别的用户查询对应的查询长度

图 3-2 中，横坐标为查询长度的大小（用查询词数表示），而纵坐标是对应查询长度的某种类型的查询所占训练集中总查询数的比例。

我们可以从图中看出，导航类查询的查询长度比信息事务类查询略长，但两者总体而言差别不大，对两类查询对应长度的统计发现，导航类查询平均长度为 1.87 词，而信息事务类查询为 1.65 词。可以说，这个特征很难将两类查询区别开来。

针对查询词词性的分析也没有取得有效的区分效果，图 3-3 中，横坐标为查询词的词性，而纵坐标是对应词性的查询所占训练集中总查询数的比例。

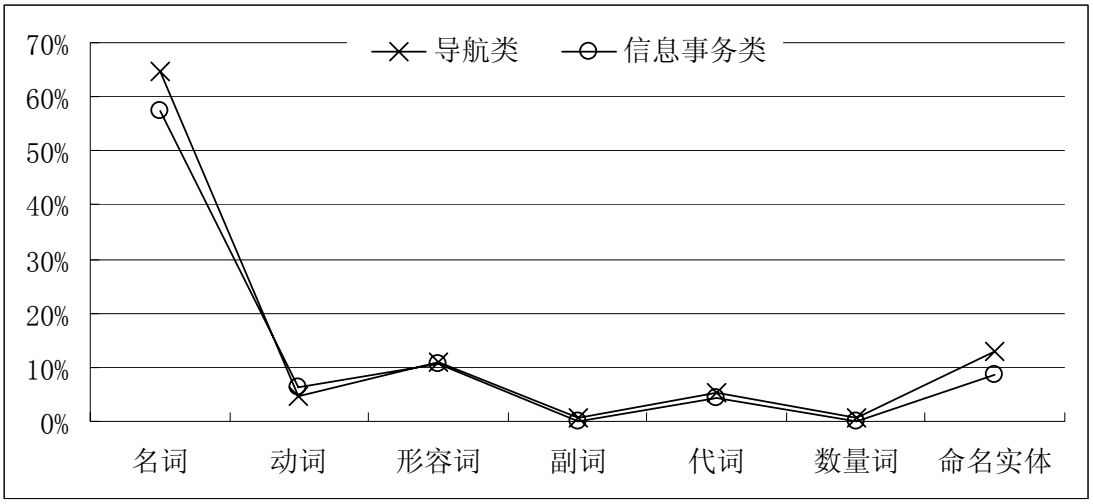

图 3-3 不同信息需求类别的用户查询对应的查询词性

实验所采用的词性标注工具为 Sogou 公司自行开发的词性标注工具 SogouSegmenter，在搜索引擎和输入法相应产品中的应用实践表明，该工具具有较高的分词和词性标注能力，因此对于查询词词性的标注基本认为是可信的，而由图 3-3 可以发现，不同类型的用户查询所对应的查询词性并没有明显的差别，两类查询的查询词主要都集中在名词上。

以上实验证明，对于中文搜索引擎用户而言，其查询词的内容特征对于查询需求分类并没有直接的区分力。这与 TREC2004 和 Kang et al（2003）的实验结论有一定差别，我们认为，这种差别主要是由中英文的语言差异与中文用户特有的“少输入、多点击”的用户行为习惯相关的。如中文用户进行导航类检索时多使用查询目标页面的简称进行查询，如“网易”、“QQ”，很少有人会使用“网易 邮箱 主页”或者“QQ 官方主页”等详细表明其查询目标页面名称

的查询此来进行检索。这与我们在 3.2.1 节中指出的中文搜索引擎用户输入的查询词数目更短的情况是一致的。综上所述，查询内容特征对于中文信息检索用户的查询需求分类，可能发挥的效果十分有限。

3.4.2 查询反馈特征

查询反馈的特征在 TREC2004 的查询分类任务中得到了广泛的应用，其主要思想来自于网络信息检索系统处理导航类查询和信息事务类查询性能上的差异。一般而言，网络信息检索系统处理导航类查询的能力较强，TREC 历年来的实验结果都说明，检索系统处理此类查询的平均性能在 MRR=0.8 左右（Hawking et al 2002）（Hawking et al 2003）（Hawking et al 2005），即平均而言，检索系统将导航类查询对应的查询目标页面返回在其结果列表的第 1.25 位（非常接近首位的位置）。而无论是 TREC 还是基于其他实验平台的检索实验都说明，处理信息事务类查询时，检索系统的最高处理性能在 MAP=0.2 左右，即平均检索精度不足 20%。这一性能上的差异是查询反馈特征得以应用的根本原因。

由于导航类查询的查询性能高，因此绝大部分导航类查询对应的查询结果列表中，第一位都是其查询目标页面，而导航类查询结果的唯一性又使得其他结果不可能是查询目标，这就有可能造成内容相似度数值从第一位结果到其他结果的突变现象。

基于查询反馈特征和查询内容特征，TREC2004 进行了查询需求分类的尝试，最终的实验结果差强人意，取得最优成绩的 Meiji 大学仅取得了 60%左右的查询分类精确率。这说明在无法获得用户行为方面实验数据的情况下，仅利用查询反馈是很难取得较好的分类效果的。

除此之外，由于查询分类特征的获取必须经过检索系统反馈的步骤，因此系统进行查询处理的步骤就由原来的直接检索变为“预检索，查询分类，再次检索”的三步流程，这需要对网络信息检索系统的现有框架进行较大规模的改动，而且也会成倍增加检索系统的运算负担。因此，无论从分类效果还是从特征获取的难度上讲，查询反馈特征都不适宜于进行查询需求分类使用。

3.4.3 链接文本特征

Kang et al（2003）的工作中首先提出了使用链接文本特征进行查询需求分类的思想，其首先假设：如果一个查询与指向某个页面的链接文本相同，则这

个查询可能是以这个页面为查询目标页面的导航类查询。由于查询词可以看作进行导航类查询的检索用户对查询目标页面的描述，而链接文本则通常被认为是网页作者对于其他网页内容的一个概括性描述，因此对于同一个网页而言，其导航类查询词与链接文本有可能一致。当然，这个假设的成立是建立在网页作者与检索用户的描述网页方式类似的基础上的，对于目标网页而言，由于网页作者和检索用户都是相对客观的描述网页内容的群体，因此假设可以认为是有可能成立的。

在假设成立的基础上，Kang et al（2003）提出了一种基于链接文本与查询词匹配度计算的查询分类算法。其首先给出语料库中所有 Web 网页与其链接文本的对照关系；随后，对于每一个查询 *Q*，算法分析 *Q* 与哪些网页的链接文本匹配；对每个拥有与 *Q* 匹配的链接文本的网页，统计其匹配 *Q* 的链接文本条数，并依照这个条数对网页进行排序。这样，对 Q 即可计算下述被称为链接文本比率（Anchor Rate）的统计指标：

$$Anchor\ Rate = \frac{Anchor\ Number\ (\text{1st ranked Web page})}{Total\ Anchor\ Number} \tag{3-1}$$

其中，分子部分是具有以 *Q* 作为链接文本的网页中，*Q* 链接文本条目最多的一个网页对应的条目数；而分母部分则是 *Q* 作为链接文本出现的总条目数。

对于导航类查询而言，由于其查询目标页面唯一，而 *Q* 很有可能就是对查询目标页面的描述，因此 *Q* 很有可能同查询目标页面的链接文本匹配。而对于导航类查询的目标页面而言，其唯一性也保证了链接文本的独特性，因此以 *Q* 为内容的链接文本应当指向相对集中（甚至是唯一）的网页，这就造成查询对应的链接文本比率较高的事实。

链接文本比率的分类特征被应用于查询需求分类中，根据 Kang et al（2003），基于此特征的分类体系在 TREC2001 数据集合上取得了约 80%的分类精确率和约 60%的分类召回率，因此是一种较为有效的分类特征。

但是 TREC 的实验设置毕竟不同于真实网络信息检索系统的应用环境，因此我们进一步对链接文本比率特征的可用性进行了考察。查询词与链接文本相匹配是链接文本比率能够被应用的前提，因此我们首先收集了 Sogou 搜索引擎索引到的网页对应的链接文本（在 2006 年 2 月，其索引规模约为 10 亿网页）。随后，基于上文提到的查询日志集合，我们统计有多大比例的查询词具有与之匹配的网页链接文本。

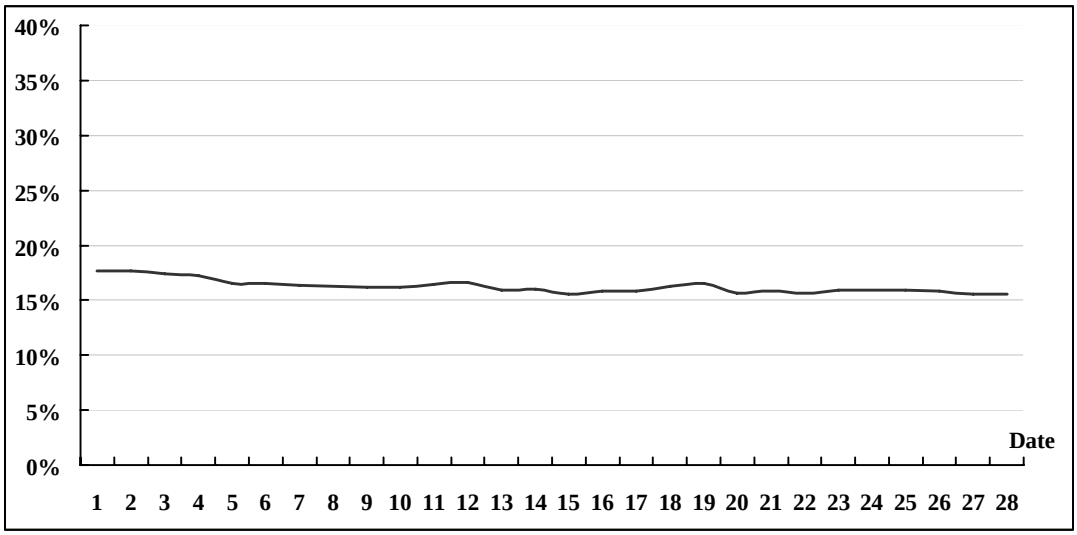

图 3-4 链接文本比率特征的可用性分析

图 3-4 中，横坐标为查询日志中的日期信息，1-28 表示 2006 年 2 月中的不同日期。纵坐标则表示与某个链接文本匹配的用户查询个数占当天总用户查询个数的比例。由图可见，每天中能够被提取链接文本比率特征的用户查询仅有不足 20%。不考虑此特征在真实应用环境中的分类能力，单就其可用性分析，即可发现特征对大部分查询不具备分类能力。Kang et al（2003）的工作中，TREC 实验平台中的绝大部分查询词都是特定设计的，如其所使用的 TREC2001 主页查找任务查询集合是直接采用主页名称作为查询内容的，因此其链接文本匹配比率要远远高于真实应用环境中的情况。

综上所述，链接文本比率特征被以往工作证明能够起到一定的查询需求分类工作，但这个特征对于大多数真实应用环境中的用户查询并不可用。这促使我们寻找其它既具有较高分类能力，又可用于大多数查询的查询需求分类特征。

3.4.4 查询历史特征

由于 TREC2004 的查询分类工作基于查询自身特征和检索结果反馈特征取得的效果十分有限，不少研究人员开始考虑利用检索系统的用户行为特征进行查询需求分类，而这种行为特征则直接反映在用户针对某个查询的历史行为特征上，即利用之前进行查询的用户的行为特性指导对当前用户查询的处理。

Lee et al（2005）的工作提出了“用户点击集中度”（Click Distribution）的概念并以此指导查询需求分类工作。其算法的核心假设是：不同用户具有相同的导航类别查询需求时，他们的点击都会集中在某个网页（或其镜像）上。这是因为：首先，查询内容相同，导航类的检索需求也相同，则目标页面一般会相同（导航类查询的目标页面唯一）；其次，如果这个目标页面在结果页面中，则必然会成为用户关注的焦点因而吸引其点击。

对于某个查询 *Q* 而言，我们可以首先定义“用户点击最集中的查询答案”为：针对 *Q* 的查询中，被不同用户点击的次数最多的查询答案 URL。

则对于 *Q* 的“用户点击集中度”具体计算公式是：

$$\text{用户点击集中度} = \frac{\text{用户点击最集中的查询结果被点击的次数}}{\text{查询Q用户的总点击数}} \tag{3-2}$$

由(3-2)式可知，由于“用户点击最集中的查询结果被点击的次数”必然小于或者等于“查询 Q 用户的总点击数”，因此“用户点击集中度”的取值范围必然在 0 至 1 之间。对于导航类查询而言，由于不同用户的点击都会集中在查询目标页面上，因此其对应的用户点击集中度自然较高。理想状态下，目标页面唯一的导航类查询的用户点击集中度应当为 1。

Lee et al（2005）的小规模实验验证了用户点击集中度的有效分类能力（他们甚至获得了超过 90%的分类精确率），这也说明了用户查询所具有的历史特征可能对当前的查询需求判断起到有益的作用。但是包括“用户点击集中度”在内的查询历史特征同样必须考虑可用性的问题，即有多大比例的用户查询具有查询历史特征。我们基于 Sogou 搜索引擎的用户查询日志进行了这方面的考察，实验结果如图 3-5 所示。

与图 3-4 类似，图 3-5 中的横坐标为查询日志中的日期信息，1-28 表示 2006 年 2 月中的不同日期。纵坐标则表示在此日期之前被查询过（即具有查询历史信息）的用户查询个数占总用户查询个数的比例。由图可见，在进行实验统计的第一天，100%的用户查询是新查询，因为按我们的统计规则，只有在当天之前被查询过的查询才认为是具有查询历史信息。从第二天开始，具有查询历史特征的查询比例激增到 85%左右，到月底的 28 日，有 90%左右的用户查询都具有查询历史特征。这说明，至少对于中文搜索引擎而言，绝大多数的查询都是在以往（至少在最近一个月内）被查询过的。

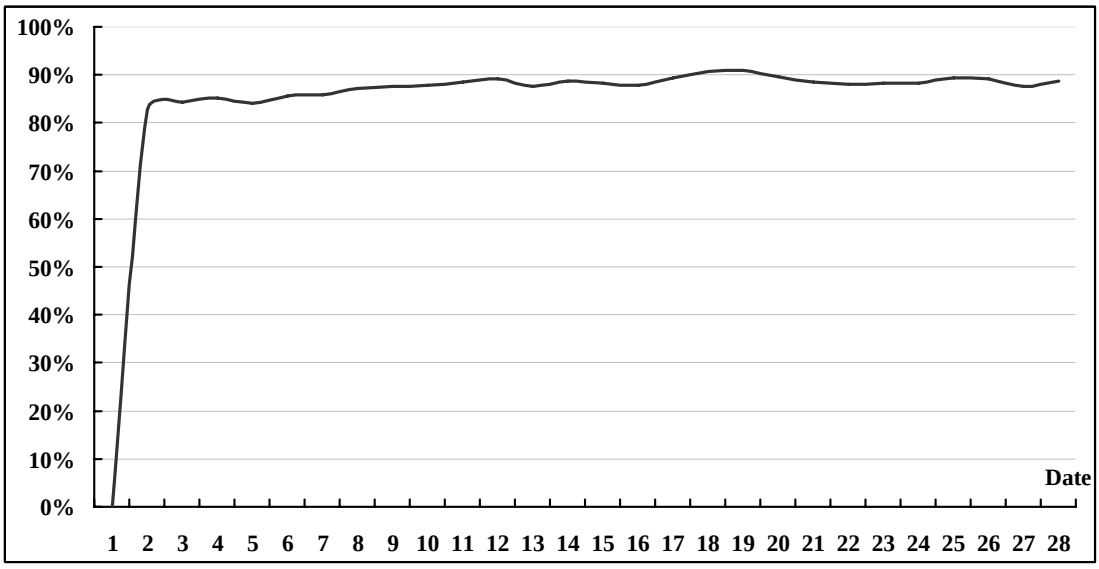

图 3-5 用户查询历史特征的可用性分析

一方面，前人工作说明查询历史特征对于查询需求分类工作有较大帮助，另一方面，针对真实搜索引擎日志的分析又说明查询历史特征对绝大多数的查询是可用的。因此我们的查询分类工作将重点围绕查询历史特征的分析进行。

3.5 面向查询分类的用户行为特征提取

上节中，我们将可能用于搜索引擎用户需求分类的查询特征进行了可靠性、分类能力和可用性的分析，分析结果证明，查询历史特征是一种分类能力强、可以用于绝大多数用户查询的分类特征。但 Lee et al（2005）工作中提到的用户点击集中度特征仅仅挖掘了查询历史信息中的少数内容，其思路重点集中在用户“是否点击”了某个结果上，而缺乏对于用户“如何点击”查询结果的深入分析。

我们对 Lee 的工作进行了进一步的发展，基于用户群体行为分析的方法对用户的查询历史特征进行了深入的挖掘和分析，提出了以下两个可以用于查询需求分类的新行为特征。

3.5.1 前 N 次点击满足用户需求率

“前 N 次点击满足用户需求率”集中在分析用户点击次数特征，它的设计思路基于如下假设：

假设 1（懒鬼假设）：用户的查询信息需求是导航类型时，一般他只会点击很少数的几个答案。

假设的合理性源自用户进行导航类检索时的信息需求特征，由于用户意识中有一个比较明确的查找目标，因此他只会在结果页面中重点浏览与这个查找目标非常相关的 URL 或摘要内容，而不会点击其他的结果。

具体到数值特征而言，某个查询的“前 N 次点击满足用户需求率”定义为只需对搜索引擎返回的结果进行小于或等于 N 次点击就满足其信息需求的用户比例。对于某个查询 *Q* 而言，其具体的计算公式是：

$$\text{前N次点击满足用户需求率} = \frac{\text{查询Q时点击小于或等于N次的用户数}}{\text{查询Q的总用户数}} \tag{3-3}$$

按照上述定义，由于“查询 *Q* 时点击次数小于或等于 N 次的用户”必然是“查询 *Q* 的用户”的一部分，因此“前 N 次点击满足用户需求率”的取值范围必然在 0 至 1 之间。我们进一步对查询样例训练集合内的“前 N 次点击满足用户需求率”取值分布情况进行了统计，如下图所示。

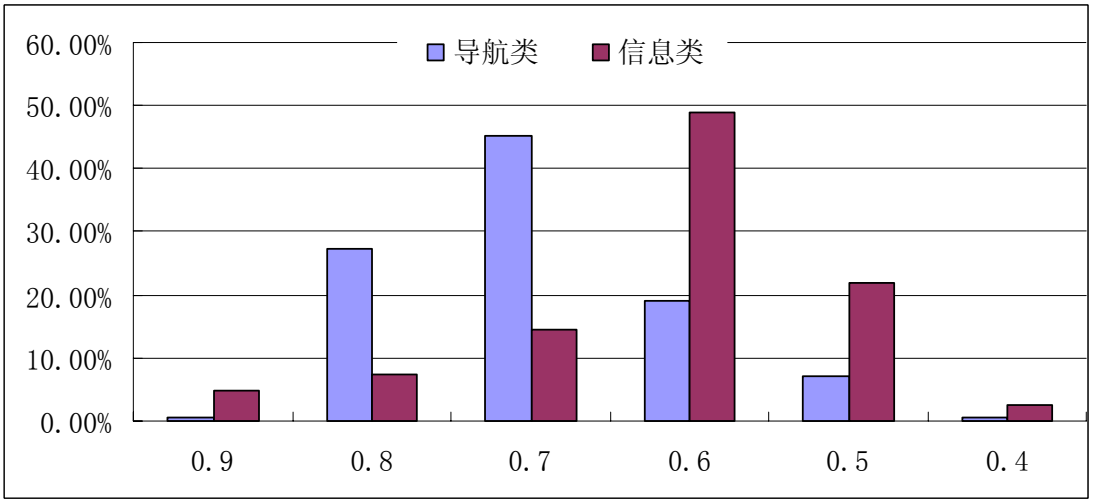

图 3-6 查询样例训练集合中的“前 1 次点击满足用户需求率”分布

图中，横坐标为“前 1 次点击满足用户需求率”的数值，即仅进行了一次点击的用户占总用户数的比率，纵坐标是对应该比率的查询在查询总数中的比例。如图 3-6 所示，我们发现这个特征能较好的区分导航类和信息事务类查询需求，正如假设 1 所提到的，用户在进行导航类查询时，其点击次数较少。实验数据说明，有 73%以上的导航类查询的“前 1 次点击满足用户需求率”超过 0.7；

而同样超过 73%的信息事务类查询的相应比率小于 0.7。

3.5.2 前 N 位结果满足用户需求率

“前 N 位结果满足用户需求率”集中在分析用户所点击结果在搜索引擎结果列表中的排序特征，它的设计思路基于如下假设：

假设 2（封面假设）：用户的查询信息需求是导航类型时，一般他只会点击排名最靠前的几个答案。

假设的合理性源自网络信息检索系统处理不同类型查询时的性能差异，如第 3.4.2 节所述，检索系统处理导航类查询的能力要远远强于其处理信息事务类查询的能力。由于检索系统处理导航类信息需求的性能一般都较高（MRR 在 80%以上），因此用户很少有必要点击前几位之后的答案。

具体到数值特征而言，某个查询的“前 N 位结果满足用户需求率”定义为只需点击前 N 位搜索引擎返回的网页结果就满足其信息需求的用户比例。

对于某个查询 Q 而言，具体的计算公式是：

$$\text{前N位结果满足用户需求率} = \frac{\text{查询Q时只点击前N个结果的用户数}}{\text{查询Q的总用户数}} \tag{3-4}$$

按照上述定义，由于“查询 Q 时只点击前 N 个结果的用户”必然是“查询 Q 的用户”的一部分，因此“前 N 位结果满足用户需求率”的取值范围必然在 0 至 1 之间。我们进一步对查询样例训练集合内的“前 N 位结果满足用户需求率”取值分布情况进行了统计，如下图所示。

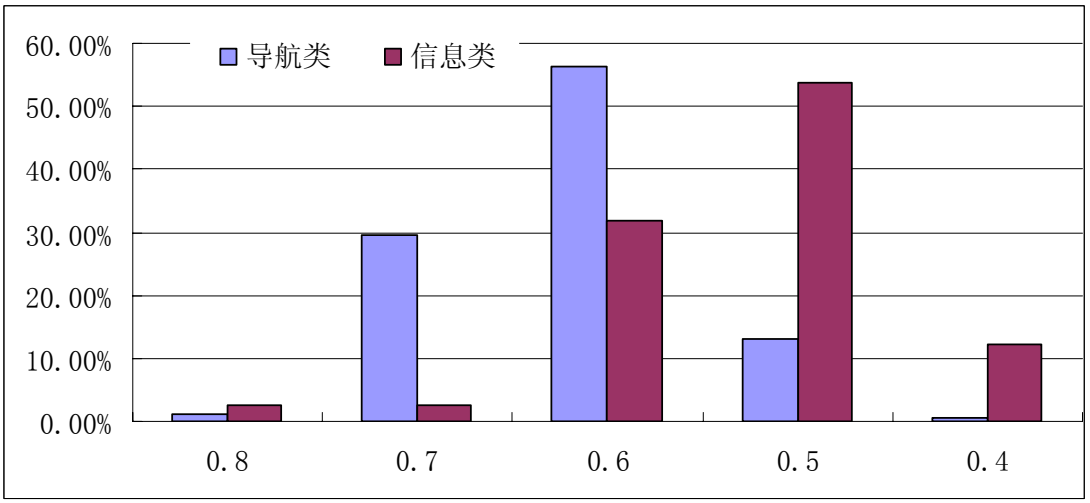

图 3-7 查询样例训练集合中的“前 5 位结果满足用户需求率”分布

图中，横坐标为“前 5 位结果满足用户需求率”的数值，即仅点击前 5 位结果的用户占总用户数的比率，纵坐标是对应该比率的查询在查询总数中的比例。如图 3-7 所示，我们发现这个特征能较好的区分导航类和信息事务类查询需求，正如假设 2 所提到的，用户在进行导航类查询时，其很少需要点击排名比较靠后的结果。实验数据说明，“前 5 位结果满足用户需求率”超过 0.7 的查询基本全部是导航类查询，而相应比率低于 0.6 的查询则绝大多数是信息事务类查询。

3.6 基于用户群体行为分析的查询分类算法

上节中，我们基于用户行为特征的挖掘方法，在“用户点击集中度”之外提出两个用于用户查询信息需求分类的新特征：“前 N 次点击满足用户需求率”和“前 N 位结果满足用户需求率”。基于查询样例训练集合的实验同时说明，这两个特征对于查询需求分类具有较高的效果，因此我们考虑使用机器学习方法将以上三个用户行为特征进行综合，以得到用于查询需求分类的更好性能的算法。

3.6.1 选择决策树算法进行查询分类的原因

我们选用了决策树学习（decision tree learning）的方法进行用户行为特征的综合，这是由于决策树算法自身的一些特点所决定的：首先，决策树学习适合解决目标函数具有离散输出值的问题，而且往往是特征数目较少时，解决此类问题的最简单有效的途径之一；此外，决策树学习中从根节点到下级节点选择特征的先后顺序，自然给予了这些特征排序关系，这也为评价用户行为特征分类能力的高低提供了重要的参考；最后，决策树算法判定样例类别的复杂度只与树的节点数相关，由于我们所采用的特征数目较少，因此算法在实际应用中的复杂度较低，适用于真实网络信息检索系统的应用需要。

根据 Mitchell（1997）的论述，决策树通过把实例从根结点排列到某个叶子结点来分类实例，叶子结点即为实例所属的分类。树上的每一个结点指定了对实例的某个属性的测试，并且该节点的每一个后继分支对应于该属性的一个可能值。决策树中，从树根到树叶的每一条路径对应一组属性测试的合取，而树本身对应于这些合取的析取。

决策树算法的核心问题是选取在树的每个结点要测试的属性，争取能够选择出最有助于分类实例的属性。为解决这个问题，ID3 算法引入了信息增益的概念，并使用信息增益的多少来决定树的不同结点需要测试的属性（Quinlan 1986）。但这样做存在着会对取值情况较多的属性有所偏袒的问题。因此 C4.5 算法针对这个问题用信息增益率代替信息增益作为评价属性分类能力的度量，从而避免使算法倾向于优先选择分支多的属性（Quinlan 1993）。C4.5 还可以通过自动离散化的方式处理取值连续的属性，并借助决策树修剪的方式减少过学习情况的出现。

除 ID3 和 C4.5 之外，比较著名的决策树学习还包括 Breiman et al（1984）提出的 CART 算法和 Kononenko et al（1984）提出的 ASSISTANT 算法，此外还包括 Mingers（1989）做出的关于属性选择和决策树修剪策略的详细比较研究等等。这些算法可能对某些特定的学习问题进行了优化，但其主要架构都与 ID3 及 C4.5 算法基本保持一致，因此不再赘述。

3.6.2 查询信息需求分类算法设计

具体到查询需求分类的任务，我们首先进行了特征参数选择的工作，针对“前 N 次点击满足用户需求率”和“前 N 位结果满足用户需求率”两个特征，其参数 N 仍需确定。我们根据每个参数计算其对应的特征分布情况，并结合以下的特征值离散化操作计算其信息增益，选择带来最大信息增益的特征参数。最终选择的参数对应的特征为“第 1 次点击满足用户需求率”和“前 5 位结果满足用户需求率”。

随后，我们利用 C4.5 算法进行了属性值离散化的操作。离散化属性取值的目的，是为了适应决策树算法处理的要求；而将取值类别局限在布尔变量上，则是出于减少算法复杂度的需要。事实上，将连续特征取值离散化是使用决策树进行学习的一般通用步骤。进行离散化后，各个属性分支的数目相同，信息增益率的比较结果与信息增益的比较结果类似，因此可以只进行运算量较小的信息增益的比较（即选用 ID3 算法进行决策树学习）。

离散化的方法是采用选取阈值的方法，阈值的选取要使 Web 页面全集在该属性进行离散化后可能的信息增益值最大。在本文的实验研究中，阈值选取按照以下方法实现：先按照属性值的分布特征选区若干可能成为阈值的取值点，

而后逐一计算比较这些待选点作为阈值的信息增益，最后选择离散化后信息增益最大的一个待选点作为离散化的阈值点。

进行特征参数选择与特征值离散化后，我们进行了针对查询需求分类任务的决策树学习，生成的决策树结果如图 3-8 所示。

从决策树生成的结果中可以得到如下结论：

1．按照特征在决策树中所处节点位置的不同，可以对特征对应的分类能力进行比较。根据我们的决策树学习结果，处于根节点的“前 5 位结果满足用户需求率”具有最强的分类能力，而随后是“前 1 次点击满足用户需求率”，用户点击集中度则在此三个特征中分类能力最差。这充分说明了我们所提出的两个查询需求分类特征具有较强的分类能力。

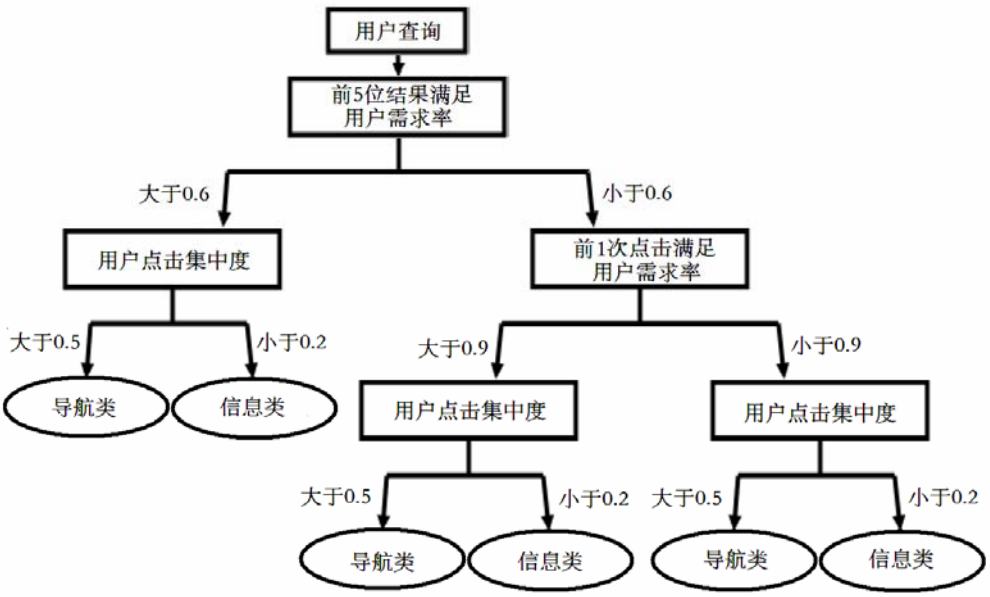

图 3-8 查询需求分类决策树

2．决策树被设计为带有拒识判断，由决策树可见，用户集中度大于 0.2 而小于 0.5 的查询将被拒识。这是与查询需求分类工作的实际情况相关的，有相当数量的用户查询难以进行用户需求的判断。这是与如下现象相关的：尽管不同用户均使用某个查询进行检索，但其查询需求是不同的。例如：对于使用查询“魔兽争霸”（著名电脑游戏名称）进行查询的用户，其可能是想定位这个游戏的官方站点（导航类需求），也可能是想获取关于这个游戏的最新资讯（信

息事务类需求）；还可能是想查找游戏补丁下载（事务类需求）。具有不同查询需求的用户使用同样的查询词进行检索，这使我们无法对某些查询的信息需求进行判断，因此采取拒识操作。

3.7 实验结果与讨论

上节中，我们依据用户群体行为分析特征和机器学习算法对用户查询需求进行了分类，本节则重点对这种分类算法的效果进行评价。

评价方式的选择上，我们挑选了考察分类算法有效性的精确率—召回率模型，其中，召回率可以通过查询信息需求测试集在实验所得到的结果集合中的覆盖度来估计，即：

$$\begin{aligned} recall(\text{导航类}) &= \frac{\#(\text{分类为导航类查询的集合} \cap \text{导航类测试集合})}{\#(\text{导航类测试集合})} \\ recall(\text{信息类}) &= \frac{\#(\text{分类为信息类查询的集合} \cap \text{信息类测试集合})}{\#(\text{信息类测试集合})} \end{aligned} \tag{3-5}$$

而精确率表达式则表示为：

$$\begin{aligned} precision(\text{导航类}) &= \frac{\#(\text{分类为导航类查询的集合} \cap \text{导航类测试集合})}{\#(\text{分类为导航类查询的集合})} \\ precision(\text{信息类}) &= \frac{\#(\text{分类为信息类查询的集合} \cap \text{信息类测试集合})}{\#(\text{分类为信息类查询的集合})} \end{aligned} \tag{3-6}$$

为了在分类精确率与召回率之间找到较好的平衡点，我们利用通常使用的均衡评价精确率与召回率的 *F-measure* 评价，它的定义为：

$$F_\alpha(r, p) = \frac{1}{\frac{1}{\alpha+1}(\frac{\alpha}{recall} + \frac{1}{precision})} = \frac{(1+\alpha)recall * precision}{recall + \alpha * precision} \tag{3-7}$$

其中精确率 *precision* 的权重为1，而召回率 *recall* 的权重为 α。在我们的实验中，我们为精确率和召回率设定同样的权重，即设定 $\alpha = 1$。

根据这一评价指标的设定，我们基于表 3-5 的训练集合和测试集合进行的试验结果如表 3-6 所示。

表中，我们可以发现，在训练集和测试集上的综合分类性能（召回率、精确率）都超过了 80%。这说明我们的算法能有效地区分大多数查询信息需求样例。

表 3-6 查询信息需求分类实验结果

	训练集合			测试集合		
	信息事务类	导航类	综合	信息事务类	导航类	综合
精确率	76.00%	91.07%	87.65%	73.74%	85.62%	81.49%
召回率	66.67%	90.71%	85.25%	72.84%	86.18%	81.54%
F-measure	0.71	0.91	0.86	0.73	0.85	0.81

同时我们注意到，信息事务类查询的分类效果要低于导航类查询的分类效果，其 F-measure 评价的差距达到 28%（训练集）和 16%（测试集）。分类效果的差异是与我们的算法设计直接相关的：由于导航类查询具有比较明显的用户行为特征，我们所使用的三个特征均是为辨识导航类查询而设计的，而信息事务类则是根据“导航类之外就是信息事务类查询”的思路来进行辨识，因此信息事务类查询与部分被拒识的导航类查询混杂在一起较难区分，造成了信息事务类分类性能稍低。

我们进一步考察了一部分样例分类错误的原因，我们发现较大量出现、以导致信息事务类查询分类召回率较低的样例都是属于一类较特殊的信息事务类查询：典型的样例包括“禽流感”、“姚明”等。这类查询虽然是信息事务类查询需求（没有确定的查询目标页面），但由于网络上客观存在一部分权威性的站点（如 http://www.39.net/disease/qlg/，http://sports.sohu.com/yaoming.shtml 等），因此绝大多数用户在进行查询后会将点击集中在这一小部分站点上。这就造成了这部分查询所对应的“前 N 次点击满足用户需求率”和“前 N 位结果满足用户需求率”都较高，进而容易被错误辨识为导航类查询。但对于这部分查询而言，检索系统查询处理的方式也可以与导航类查询类似，如较多采用链接文本内容、URL 类型等协助排序，因而这类错误并不会对检索系统查询处理的性能产生明显的影响。

我们将基于决策树的查询需求分类方法与 Lee et al（2005）提出的基于点击集中度的查询需求分类方法进行了比较。由于 Lee 等人的方法与我们的方法均为使用用户点击信息进行分类，因此这种比较有较强的可比性。实验分别在训练集和测试集平台上进行，考察两种不同方法进行查询需求分类实验的具体结果，实验结果如下图所示：

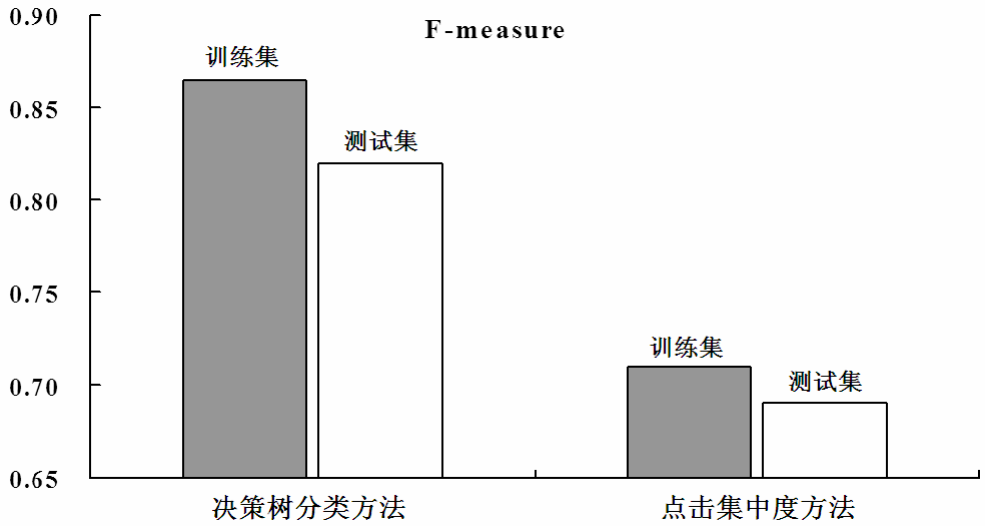

图 3-9 查询需求分类实验结果比较

由图 3-9 可知，在训练集和测试集上，我们所提出的基于决策树分类的方法都优于基于点击集中度的方法。训练集的性能提高为 21%，而测试集提高 18%。这说明，一方面，我们所提出的“前 N 次点击满足用户需求率”和“前 N 位结果满足用户需求率”两个用户行为特征能够有效地区分导航类与信息事务类查询需求；另一方面，决策树学习的方式是合适的综合用户行为特征进行查询分类的方法。

3.8 本章小结

对用户信息需求的正确理解与否，直接关系到网络信息检索系统查询处理能力的高低。由于当前的检索系统交互方式造成用户的信息需求传递瓶颈，因此利用用户群体行为分析的方法，通过用户群体的查询点击行为分析当前查询对应的信息需求类别成为一个可行的解决途径。

为了实现提高检索系统查询处理能力的目的，我们设计了基于用户群体行为分析的查询需求分类算法。首先，我们分析了查询需求分类工作可能依据的查询和用户行为特征，并对这些特征的可用性和分类能力进行了比较。随后，我们选择用户查询和点击行为的历史信息作为分类工作的主要依据，提出了“前N次点击满足用户需求率”和“前N位结果满足用户需求率”两个用户行为特征并考察了此两个特征的分布特性。最后，我们基于决策树学习的方法将用户行为特征加以综合，实现了基于用户群体行为分析的查询需求分类算法。

利用海量规模搜索引擎网络日志数据的实验结果证明，上述基于用户群体行为分析的查询需求分类算法能够辨识大部分用户查询信息需求，其辨识能力超过以往利用用户行为信息进行查询需求分类的工作成果。由于我们所采用的用户行为特征可以作为绝大多数用户查询的分类依据，因此算法也适用于真实网络信息检索系统的查询处理流程。比较前人所提出的基于链接文本信息、查询内容信息、查询反馈信息的查询需求分类算法，其算法效率、性能和适用性都有较大的提高。

综上所述，我们基于用户群体行为分析的思路实现了一种有效的查询信息需求分类算法，此算法适应大规模网络信息检索系统的应用需求，同时能够准确辨识用户的查询信息需求，以提高检索系统查询处理的性能。

第四章 基于用户群体行为分析的搜索引擎自动性能评价研究

4.1 引言

上两章，我们分别利用用户群体行为分析的方法尝试改进网络信息检索系统中的数据处理与查询处理过程，以缓解搜索引擎用户信息需求传递不畅造成的瓶颈问题。本章我们将考察的重点转移到检索系统的评价问题，尝试基于用户群体行为分析的方法实现搜索引擎的自动性能评价。

检索系统的评价问题一直是信息检索研究中的最核心问题之一， Saracevic（1995）指出：“评价问题在信息检索研发过程中处于如此重要的地位，以致于任何一种新方法与他们的评价方式是融为一体的”。Kent et al（1955）首先提出了精确率—召回率的信息检索评价框架（如第 1.1.3 节所述），随后，美国政府所属的研究机构开始大力支持关于检索评价方面的研究，而英国 Cranfield 工程在上世纪五十年代末到六十年代中期所建立的基于查询样例集、标准答案集和语料库的评测方案，则真正使信息检索成为了一门实证性质的学科，也由此确立了评价在信息检索研究中的核心地位（Cleverdon et al，1966），其评价框架一般被称为 Cranfield 方法（A Cranfield-like approach）。

Cranfield 方法指出，信息检索系统的评价应由如下几个环节组成：首先，确定查询样例集合，抽取最能表示用户信息需求的一部分查询样例构建一个规模恰当的集合；其次，针对查询样例集合，在检索系统需要检索的语料库中寻找对应的答案，即进行标准答案集合的标注；最后，将查询样例集合和语料库输入检索系统，系统反馈检索结果，再利用检索评价指标对检索结果和标准答案的接近程度进行评价，给出最终的用数值表示的评价结果。

Cranfield 方法一直到今天也被广泛的应用于包括搜索引擎在内的大多数信息检索系统评价工作中。由美国国防部高等研究计划署(Defense Advanced Research Projects Agency，简称 DARPA)与美国国家标准和技术局(National Institute of Standards and Technology，简称 NIST)共同举办的 TREC（文本信息检索会议，http://trec.nist.gov/）就是一直基于此方法组织信息检索评测和技术交流论坛。除 TREC 之外，也有一些针对不同语言设计的基于 Cranfield 方法的检索

评价论坛开始尝试运作，如 NTCIR (NACSIS Test Collection for IR Systems)计划与 IREX (Information Retrieval and Extraction Exercise)计划等。

随着万维网的不断发展与互联网信息量的增加，如何评价网络信息检索系统的性能逐渐成为近年信息检索评价中的热点关注方向，而进行这方面评价时，Cranfield 方法遇到了巨大的障碍。困难主要反映在针对查询样例集合的标准答案标注上，根据 Voorhees et al（2001）的估计，对一个规模为 800 万文档的语料库进行某个查询样例的标准答案标注需要耗费 9 个评测人员一个月的工作时间。尽管 Voorhees 提出了诸如 Pooling（如第 1.3.2 节所述）这样的标注方法来缓解标注压力，但当前针对海量规模网络文档的答案标注仍是十分困难的。如 TREC 海量规模检索任务（Terabyte Track）一般需要耗费十余名标注人员 2-3 个月的时间进行约几十个查询样例的标注，而其语料库数据规模不过 1000 万文档左右。考虑到当前搜索引擎涉及到的索引页面都在几十亿页面以上（Yahoo!报告为 192 亿网页，中文方面 Sogou 声称的索引量也超过百亿），利用手工标注答案的方式进行网络信息检索系统的评价会是一个既耗费人力、又耗费时间的过程。由于搜索引擎算法改进、运营维护的需要，检索效果评价反馈时间需要尽量缩短，因此提高搜索引擎性能评价的自动化水平是当前检索系统评价研究中的热点。

4.2 相关工作

为了解决 Cranfield 方法在网络信息检索系统评价中所面临的困境，不少研究人员提出了一些自动进行搜索引擎性能评估的方案，其工作集中在两个方面：基于 Cranfield 框架，只是使用自动化方法进行答案自动标注；或采用不同于 Cranfield 方法的评价框架进行自动化评价。

4.2.1 基于 Cranfield 方法的搜索引擎自动性能评价

前一方面的研究工作中，研究者尝试使用检索系统反馈的结果信息进行自动标注。Soboroff et al（2001） 在基于 TREC 实验平台的研究中发现：评价人员对于结果池内文档的标注结果差异基本不影响检索系统性能排序的结果，因而随机挑选结果池内文档作为标准答案也有可能达到评价检索系统性能的作用。他因而提出可以在检索系统结果池中，随机挑选一定数量的结果作为答案

集合进行评价。实验效果证明，按这种方式实现的检索系统评价结果与基于手工标注集合的评价结果正相关，但对于检索系统性能排序的影响较大因而难以投入使用。Nuray et al（2003）提出对 Soboroff 方法的修正方案，即选择结果池中原本在搜索引擎结果序列中排序较前的文档作为标准答案，他们的方法也没有取得与手工评价方法相类似的评价结果。

我们认为，这类基于搜索引擎结果反馈信息（伪相关反馈信息）进行搜索引擎评价的尝试很难获得成功。这是由于伪相关反馈信息本身就是一种不可靠的信息源，它只能对搜索引擎处理性能较高的查询进行正确的结果标注，而事实上由于针对这部分查询的评价不会对搜索引擎的性能提高起到指导作用，因此很少需要对其进行性能评价。这就形成了需要进行评价的查询标注不好，不需要进行评价的查询反而标注的较好的情况，因此这种自动标注的思路很难应用于实际搜索引擎评价中。

也有部分研究人员基于已有的网页目录资源进行结果的自动标注，如 Chowdhury et al（2002）和 Beitzel et al（2003）提出的利用开放目录计划（ODP 计划）所整理的网页目录和对应的网页摘要资源进行性能评测的工作。其方法的优势在于答案标注的正确性比较单纯使用搜索引擎结果反馈信息更高，但使用网页对应的摘要信息作为用户查询的模拟还是一个不合理的假设。尽管针对这种模拟的查询集合的结果标注比较可信，但模拟查询集合自身并不能反映真实的用户行为，因而其工作也没有得到大规模的普及应用。

4.2.2 搜索引擎自动性能评价的其他尝试

第二方面的研究工作中，比较有代表性的有 IBM Haifa 研究院研发的“相关词集合评价方法”与 Joachims 提出的基于用户点击行为的评价方法等。

Amitay et al（2004）提出了“相关词集合评价方法”（*Term Relevance Sets*, 简称 Trels 方法）。方法的核心思想是对词与词之间的相似度评价代替传统评价方式中词与文档之间的相似度评价。在 Trels 方法中，首先选择一定量的代表用户查询需求的查询词；随后针对每一个查询词，手工标注尽量多的与此查询词相关联的词项；施行评价时，通过待评测文档中关联词项的分布情况判定文档的相关程度及检索结果的可靠性。这种方法将大量手工工作从收集检索结果的过程之后转移到收集结果之前，作者也认为其标注的关联词项能够较长时间发挥稳定的评价作用。

Trels 方法一定程度上解决了评价结果反馈时间过长的问题，但丝毫没有减少甚至增加了相关性标注的难度，对于网络信息检索中的海量用户需求，标注词与词的相关度依然会变得不可行。同时，词与词的相关程度本身就是一个难以界定的问题，随着时间、地域的不同，这类相关的概念会有较大的差别。作者基于 TREC 小规模数据的实验取得了一定的效果，但并没有将之使用在大规模的网络信息检索系统评价中。

Joachims（2002）第一次提出了使用用户点击行为信息评价搜索引擎性能的思路。他设计了一个元搜索引擎，用户输入查询词后，将查询词在几个著名搜索引擎中的查询结果随机混合反馈给用户，并收集随后用户的结果点击行为信息。根据用户不同的点击倾向性，就可以判断搜索引擎返回结果的优劣，Joachims 同时证明了这种评价方法与传统 Cranfield 方法评价结果具有较高的相关性。

Joachims（2002）的方法引入了用户点击行为分析的研究思路，记录用户选择检索结果的行为是一个不耗费人力的过程，因此可以避免传统 Cranfield 方法反馈过慢的问题。但这之前，必须首先评判用户点击行为的可靠性，即用户的点击是否意味着其认为被点击的结果与查询相关。Joachims 在这方面并没有给出一个完善的解决方案，其随机混合答案的方式尽管避免了所谓的“排序偏置”（即减少用户因为结果排列在前面就点击它的可能性），但也与用户正常使用搜索引擎的体验产生差异，因此收集到的用户行为可信程度降低；同时，使用这个元搜索引擎本身并无法为用户带来更加快捷方便的搜索体验，因此其必然无法吸引足够多的用户提供点击信息，进而影响到评价结果的可信程度。Joachims 的实验结果仅仅是建立在不足几百条点击记录的基础上的，因此其可信程度很难保证。

4.2.3 搜索引擎自动评价已有工作总述

综上所述，研究人员基于 Cranfield 框架进行了自动结果标注的尝试，但由于选择的标注方式不可靠而没有获得成功；在 Cranfield 框架之外进行的各种尝试，尽管自动化程度都较高，但其评价方法的可靠性问题还有待商榷。

我们认为，Cranfield 的检索系统评价方式是经过相当程度的理论和实践检验，因而在其面临搜索引擎评价的困境时将其抛弃是不明智的选择。而发展 Joachims 的用户点击行为分析方法，将其扩展到查询样例集合的结果自动标注过程中，是一个可行的解决方案。

4.3 查询信息需求与自动性能评价

上一节，我们对搜索引擎自动评价的研究成果进行了综述，并提出了使用用户点击行为分析的方法进行答案自动标注的问题。这个想法的出发点在于：由于现有的绝大多数搜索引擎用户还是能够通过搜索引擎找到满足其查询需求的答案（尽管可能需要花费较多的精力），因此用户的点击行为中肯定蕴含了其对检索结果相关性的评价。

从个体用户的行为上讲，有可能由于个人知识水平、网络使用习惯的不同而点击某些与查询需求无关的页面，甚至有可能被垃圾页面，SEO 页面等所欺骗；但从用户群体的宏观行为规律上讲，这些无关点击可以被认为是随机噪声而滤除掉。因而当用户群体足够大，收集到的点击信息足够完善时，点击信息的可靠程度还是能够得到一定的保证。

对于搜索引擎而言，其网络服务供应商的身份同时也为其收集了海量规模的用户日志信息。第三章中，我们利用这部分用户日志信息实现了用户查询信息需求的分类，那么，利用这些信息中蕴含的用户群体点击行为信息实现答案自动标注也是一个自然的解决问题的思路。

然而，用户群体行为的可靠性尽管可以得到保证，但对于性能评价中的答案标注而言，标注出正确的结果并不是唯一需要考虑的问题，是否标注出了所有正确的结果同样值得考虑，这就需要具体考虑用户查询信息需求的问题（参见第 3.3 节对用户查询信息需求的表述）。

对于导航类查询而言，其正确答案唯一，因而不需考虑答案全面性的问题；其对应的搜索引擎检索性能也较高，因此用户点击行为的可靠性也比较容易保证。即：用户在进行导航类查询时，较容易发现并点击结果列表中对应的答案，因而我们所进行的主要工作，只是将用户点击行为中反映出的答案挑选出来。

对于信息事务类查询而言，情况则要复杂的多，其正确答案不唯一，因此必须考虑答案全面性的问题；而其对应的搜索引擎检索性能相对较低，用户能否点击到即使是正确的答案也较难保证。

为了考察用户点击行为是否适用于进行信息事务类查询的答案标注，我们考察了提交查询词“电影”的四个常用中文搜索引擎（百度，谷歌，雅虎，搜狗）用户在 2006 年 12 月 10 日的点击情况，如下图所示：

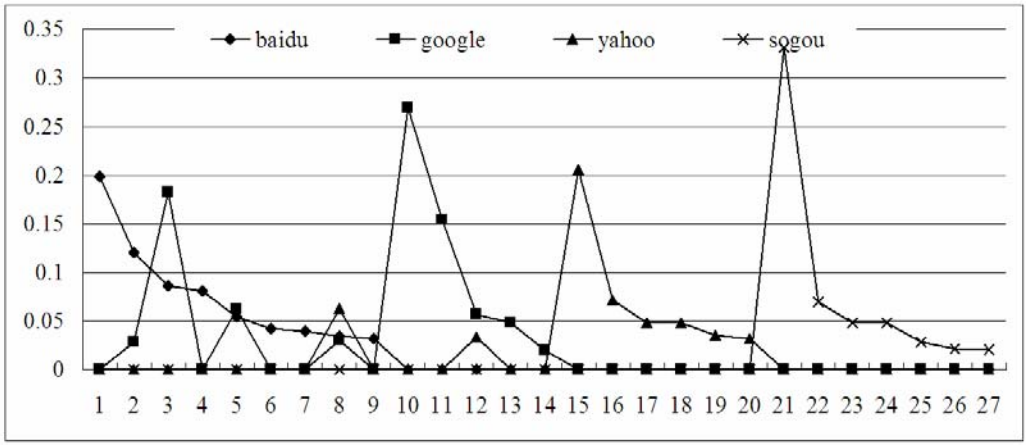

图 4-1 针对查询词“电影”的不同搜索引擎用户点击情况

实验收集了四个搜索引擎针对查询词返回的前 10 位结果，取并集后共 27 个结果，图 4-1 中的横轴对应这 27 个结果，而曲线上的点则是结果对应的不同搜索引擎的用户点击频度信息。如第 21 号结果对应的搜狗搜索引擎曲线（用“×”表示）上的数值约为 34%，即代表第 21 号结果在搜狗搜索引擎上被 34%的查询“电影”的用户所点击。本实验数据的获得是通过搜狗公司采集的用户搜索反馈信息，共涉及了近 200 名用户的搜索引擎访问信息。

图中，我们可以发现，不同搜索引擎用户针对这个查询的点击情况差异非常大，如百度用户的点击多集中在第 1 号结果上，而谷歌用户点击第 3 号和第 10 号的最多；各个搜索引擎结果尽管有一定交集，如第 3、5、8 号结果均被多个搜索引擎用户所关注，但关注程度却有较大差异。

尽管“电影”这个查询词仅仅是信息事务类查询的一个简单样例，但它可以反映出这种类型的查询需求对应的检索结果反馈现象：提交同一个信息事务类查询需求时，用户在不同搜索引擎上得到的结果是不同的。这种差异既来源于搜索引擎的页面索引差异（即不同搜索引擎索引到的页面集合不同），也来源于搜索引擎的结果排序策略差异，因而对于查询目标页面不唯一的信息事务类查询是难以避免的。

这说明，对于信息事务类而言，用户期望的正确答案可能有多个，但某单个搜索引擎则很难反馈全所有的结果，因此使用某个搜索引擎的用户行为信息去评价其他搜索引擎信息事务类查询的性能是不合理的。

对于研究人员而言，获取多家搜索引擎的用户日志有较高的难度，对于搜索引擎自身来讲，获取其它供应商的日志更是难上加难，因此在现有的实验环境和商业运行模式下，实现信息事务类查询的自动评价可能是不现实的选择。

4.4 导航类自动性能评价算法设计

上节的论述中，我们明确了在当前的实际应用条件限制下，搜索引擎性能自动评价的对象只能限制于导航类检索，因此本节我们来讨论导航类自动性能评价系统的算法设计。依照 Cranfield 方法框架，查询样例集合、标准答案集合和语料库是性能评价必备的三要素，对于网络信息检索系统而言，Web 数据集合即其面对的语料对象，因此实现查询样例集合和标准答案集合的自动生成，就成了我们所主要关心的问题，包括这两个环节在内的搜索引擎自动评价方法的整体运行流程如下图所示。

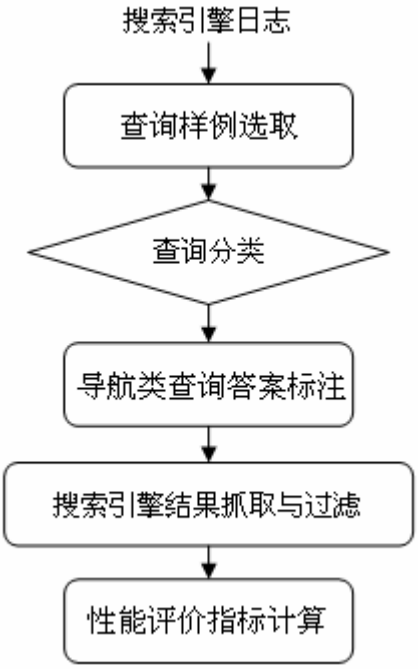

图 4-2 搜索引擎自动评测方法流程

搜索引擎日志首先经过数据预处理，获得必需的用户点击行为特征，随后进行查询样例集合的自动选取，并依据第三章所述的搜索引擎用户查询信息需求分类方法进行查询需求分类，其中的导航类需求被挑选进行自动标准答案标

注，此后进行搜索引擎结果的抓取和性能评价指标的计算。

在上述评测方法流程中，搜索引擎结果的抓取与过滤是指将查询样例集合中的样例提交给搜索引擎进行查询，并收集其结果页面，过滤出结果 URL 列表。而搜索引擎的性能评价指标计算则是指根据搜索引擎返回的结果 URL 列表与自动标注出的答案集合，计算性能评价指标的过程。对于导航类查询需求而言，性能评价指标使用“首现正确结果排序倒数”（RR，参见第 1.3.2 节）进行计算。

4.4.1 查询样例集合的自动生成

构建有合适代表性的查询样例集合对于搜索引擎评价结果的可靠性也是至关重要的。在传统的性能评价研究如 TREC 相关工作中，查询样例集一般是由评测人员专门挑选出的，部分任务的查询主题可能来自于对搜索引擎日志的筛选，但大部分是专门设计的用于评测系统性能水平的特定查询。此外，由于手工标注工作量的限制，因此查询样例集合的规模一般较小，每单个 TREC 检索任务的查询样例集合约包括几十到一、二百查询不等。

由于我们所进行的是自动性能评测系统的查询样例集合设计，因此可以较少考虑人工标注所导致的查询数量限制，因此我们重点考察查询样例集合的代表性问题，即多大规模的样例集合足够代表搜索引擎用户的实际查询情况。为此，我们对第 3.1 节所述的 Sogou 搜索引擎日志集合进行了查询频度分析，分析结果如图 4-3 所示。

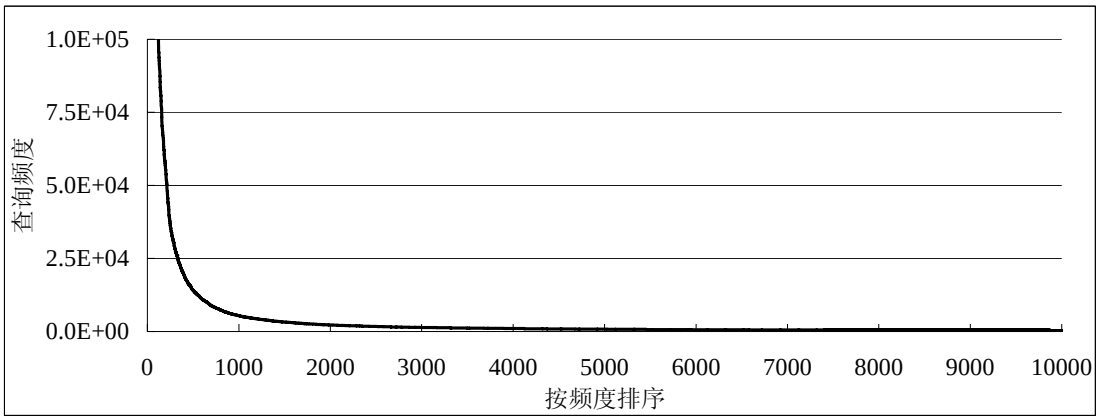

图 4-3 查询日志中的查询频度分布情况

图 4-3 中，我们选择了查询频度最高的 10000 个查询词，并观察其频度的分布情况。图中的横坐标为按频度进行排序的序号，而纵坐标为对应排序的查询的查询频度多少。从图中，我们可以发现频度绝对数值随排序增加下降的非常迅速，这意味着少数查询即可能代表相当大一部分的用户查询需求。根据统计，此查询词集合中频度高于 100 的查询仅有 35177 个，占查询总数目的不足 1%，但此 1%的查询却覆盖了 69%的用户查询需求。这说明使用一个较小规模的查询样例集合代表搜索引擎大部分用户的信息需求是完全可行的。

尽管标准答案集合的标注将自动完成，但由于搜索引擎结果抓取速度受到网速、搜索引擎服务策略等多方面的限制，因此查询样例集合的整体规模不宜过大。考虑到实际施行难易程度和用户需求代表性两方面因素，我们认为选择约 10000-15000 查询词作为查询样例集合较为合适。这个规模的样例集合能够代表相当大比例的用户需求（一个月用户需求总数的约 50%），处理时间也可以接受（当程序运行硬件环境为 1.8G 主频 CPU、1G 内存与 100MLAN 网络时，约需 1 天时间完成性能评估）。

因此，选取一段时间内搜索引擎用户查询频度最高的一定数量查询，是我们构建查询样例集合的核心策略。

4.4.2 标准答案集合的自动标注

比较起查询样例集合的自动生成而言，标准答案集合的标注是更为困难的研究课题。在 4.1 节的论述中，我们也指出这个标注过程是 Cranfield 方法在评价搜索引擎性能时面临的最大困难所在。上节中我们将查询样例集合的规模控制在 10000-15000 查询左右，这个规模的样例集合对于手工标注答案而言是不可完成的任务，因此构建快捷准确的答案标注算法势在必行。

第 3.4 节中，我们给出了点击集中度的定义，点击集中度首先由 Lee et al（2005）给出，对于某个查询 Q，定义 R_{most} 为查询 Q 的搜索引擎用户点击的最多的一个结果，而点击集中度则为对应 R_{most} 的点击数与针对 Q 的总点击数的比例值。Lee 提出点击集中度的概念，更多的是从查询信息需求分类的角度进行考虑，而我们把注意力转向对于 R_{most} 的考察。

对于导航类查询而言，由于用户的信息需求唯一，因而用户的点击一般会集中在其查询目标页面上，即 R_{most} 有很大可能性成为查询目标页面。但当搜索引擎无法把查询目标页面反馈在结果序列中较靠前的位置时，由于 Silversteinet

al（1998）和余慧佳 等（2007）指出的“绝大部分用户只点击第一页搜索结果”的情况，上述推断也可能出现错误。由于搜索引擎针对导航类查询的检索效果较好，在前一、两页结果中没有返回查询目标页面的概率很小，因此这种推断错误的情况应该认为很少发生。

如果定义网页 r 针对查询 Q 的“点击比率”为：

$$\text{点击比率}\ (Q, r) = \frac{\text{查询Q的用户中，点击}\ \ r\ \text{的次数}}{\text{查询Q的用户进行点击的总次}\ \ \text{数}} \tag{4-1}$$

则我们有下式成立：

$$\text{点击比率}\ (Q, R_{most}) = \text{点击集中度}\ (Q) \tag{4-2}$$

即点击集中度等于点击比率的最大值，而点击比率最大的 r 则很有可能成为导航类查询 Q 的查询目标页面。

依照上述推断，我们可以设计下面的标准答案自动标注算法：

对待标注的查询样例集合中给定的查询 Q 及其被点击过的查询结果 r_1, r_2, …, r_M：

IF Q 为导航类查询
    在 $r_1, r_2, \ldots, r_M$ 中定位 R，使之满足点击比率(Q,R) = 点击集中度(Q);
    IF 点击集中度(Q) > T
        标注 R 为 Q 的标准答案;
    EXIT;
    ELSE
        Q 不可被标注;
    END IF
ELSE //Q 不是导航类查询
    Q 不可被标注;
END IF

在算法过程中设计点击集中度最小阈值 T 的目的，在于避免出现前文提到的搜索引擎无法把查询目标页面反馈在结果序列中较靠前的位置的情况，在这种情况下，无法依靠用户行为日志实现答案自动标注，因此算法需要给出对应的提示信息。

4.5 实验结果与讨论

本节针对上述提出的搜索引擎自动性能评价算法，利用实验方法验证其可靠性。实验数据采集自 Sogou 搜索引擎 2006 年 6 月至 2007 年 1 月半年多的查询、点击日志，每日的查询和点击行为数量约为 150 万条。采用海量真实规模搜索引擎数据进行算法有效性的验证，可以充分考察算法的可靠性及施行效率。

如上文所述，实验所采用的硬件平台是一台普通 PC 级别计算机，整体花费约 6000 元人民币，实验在 100M 局域网内进行。算法的时间花销主要集中在待评测搜索引擎结果序列的抓取上，每小时处理的查询个数约为 400 个。比较传统性能评价工作中十几个标注人员工作数月的效率而言，有了质的飞跃性提高。

下面我们将分别从标准答案标注的正确性与性能评价指标的准确性两方面来衡量搜索引擎自动评价算法的性能。

4.5.1 答案自动标注实验结果

为控制查询样例集合的规模、同时考察算法针对用户行为时间变化的鲁棒性，我们没有整体使用上述日志数据，而是将实验所用的用户行为日志数据按时间段分成三部分，分别施行查询样例提取和标准答案标注。针对每一个部分的答案标注结果，我们随机抽取约 5%的数据进行手工验证，以检查其标注正确性。实验结果如下表所示：

表 4-1 答案自动标注实验结果

时间段	查询样例集合规模	手工验证数据规模	精确率
2006.6 - 2006.8	13,902	695	98.13%
2006.9 - 2006.11	13,884	694	97.41%
2006.12 - 2007.1	11,296	565	96.64%

每个时间段导航类查询样例集合的规模都控制在了 10000-15000 个查询，最后一个时间段的时间跨度略短，因此集合规模也略小。而手工验证的实验结果则说明，答案标注的准确度相当高，每个样例集合的标注精确率都超过了 95%，考虑到即使手工标注也很难避免错误，自动答案标注可以说满足了搜索引擎性能自动评价的需要。

我们进一步分析了若干标注错误的样例，发现绝大部分的错误都是如下情况导致的：正确结果应当是某个站点的主页，而自动标注的结果为此站点的某个子站点主页。例如查询“163”对应的标准答案应当是 http://www.163.com，而自动标注的答案是 http://mail.163.com；查询“搜狗”对应的标准答案应当是 http://www.sogou.com，而自动标注的答案是 http://d.sogou.com。

通过进一步分析，我们发现，这种自动标注答案发生错误的情形是由用户点击行为的倾向性造成的。由于绝大部分用户事实上并非想访问其查询词对应的站点主页，而是这个站点最具吸引力的某个子站点首页，因此主页的用户点击集中度反而不及子站点首页高。查询“163”的用户大都点击了其邮箱主页是因为 163 的免费邮箱服务是中文网络环境中最受欢迎的，而搜狗 mp3 搜索比搜狗网页搜索产品质量更高也是一般网络用户的共识。

因此，这种错误的产生事实上并不是自动标注算法的谬误，而是用户真实行为和选择倾向性的体现。从这个角度讲，是否只有标注站点主页才算标注正确，也是值得商榷的问题，但这并非本文讨论的重点，在下文的实验中，我们还是以传统意义上的主页作为查询对应的标准答案。

4.5.2 性能评价实验结果

依照 4.4 节中所述的搜索引擎自动性能评价方法，即可以自动生成性能评价所需的查询样例集合和标准答案集合，并使用 Cranfield 方法对搜索引擎的处理导航类查询需求的性能进行评价。

我们选用了搜狗搜索引擎之外的五个中文搜索引擎作为性能评价的对象，它们是：百度、谷歌、雅虎中文、新浪爱问搜索与中国搜索。之所以没有选择提供用户行为日志的搜狗搜索作为被评价的目标，是考虑到用户行为日志的提供者有可能在评价的过程中被给予不应有的偏向（这部分内容将在 4.5.3 节中详细分析）。

评价指标方面，我们选择了“首现正确结果排序倒数”（RR，参见第 1.3.2 节）指标进行性能计算与比较。而为了验证评价实验效果的正确性，我们也建立了一个手工评价集合与自动评价结果进行比较。

手工评价的查询样例集合由评价人员从查询日志中手工挑选的 320 个导航类查询词组成，查询样例集合被投入上述五个搜索引擎进行查询并收集结果，

将结果取并组成结果池，并使用 pooling 方法由评价人员进行了答案标注，以此获得了查询样例对应的标准答案集合和进一步的搜索引擎评价结果。

具体的评价实验结果如下图所示：

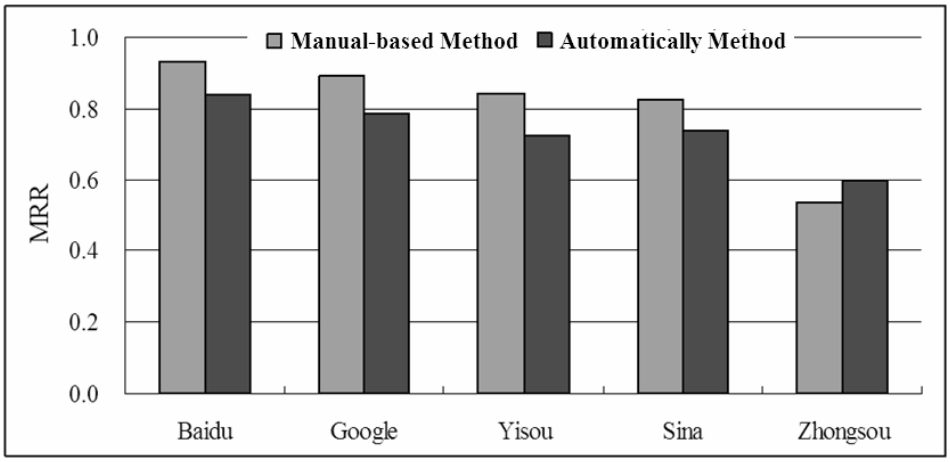

图 4-4 搜索引擎自动性能评价结果与手工评价结果的比较

图 4-4 介绍了搜索引擎自动评价的实验结果，并将其与手工评价结果进行了比较。由图可见，自动评价的搜索引擎性能排序结果与手工评价的结果完全相同，两种方法给出的 MRR 值之间的相关系数为 0.965，这意味着两种方法给出的评价结果非常相似。

从图中我们可以发现，尽管从总体性能评价排序而言，手工方法与自动评价方法保持一致，但两种评价方式对应的 MRR 绝对数值则有一定差别。这是由所采用的查询样例集合不一致造成的，手工评价的 MRR 数值较高，是由于其手工选择的样例较少，可能都是搜索引擎处理的较好的样例，因而性能绝对数值都较高；而自动评价所选用的样例数量多（是手工评价样例的 40 倍以上），代表性好，其数值更能反映搜索引擎的实际处理能力的高低。

4.5.3 自动性能评价的实验结果讨论

正如我们在第 4.2 节中所介绍的，我们的研究工作并不是首次把用户行为信息应用到搜索引擎评价中的尝试， Joachims（2002）就提出了利用元搜索引擎接口评价搜索引擎性能的思路。但我们对以往工作中的核心假设，即用户点击的结果即为相关结果产生了怀疑，这促使我们从宏观而不是个体的角度考察用户行为，并提出了本章中上述的搜索引擎自动评价方法。

利用用户行为信息进行搜索引擎评价的工作容易被两方面的问题所困扰，其一是我们已经提到的用户点击的可信度问题，其二则是行为信息提供者（在本文的工作中即为搜狗搜索）本身的评价问题。

对于第一个问题，用户个体的点击行为确实容易受到多种因素的干扰而发生偏移。这是由于用户点击行为的作出是基于搜索引擎返回结果的标题与部分摘要文字作出的，但标题和摘要文字并无法完整的代表网页全貌，甚至有部分标题与摘要被专门设计用于欺骗用户点击（这在 SEO 操作中屡见不鲜）。对于中文搜索引擎用户，在第 3.1 节中我们也曾提到其行为具有“懒于查询勤于点击”的特点，这就使个体用户点击行为的质量变得更不可信。

但从宏观而言，用户群体的点击行为还是可以信任的，如果某个页面吸引大量用户进行点击，但其并不是真正的查询目标页面，则这个页面要么是我们上文提到的与查询目标页面相关的同样吸引用户的一个页面；要么则是设计的极为出色的垃圾页面。从搜索引擎设计的角度，这样的垃圾页面会对用户体验造成毁灭性的打击，因此一般会被及时处理，而不会对用户的行为产生过大的影响。因此，用户群体在一定长度时间段内的行为特征是值得信任的。我们在 4.5.1 和 4.5.2 节的实验结果也充分验证了这一点。

对于第二个问题，我们认为用户行为信息的提供者不适宜同时作为被评价的对象。这是因为作为日志提供者的搜索引擎在评价过程中会具有评价偏向：由于只有出现在日志提供者索引中的网页才有可能被选择为标准答案，因此未出现在日志提供者索引中的正确答案也不可能被标注，即使其他搜索引擎返回了这样的答案，自动评价方法也无法辨识。例如 http://www.sysu.edu.cn/和 http://www.zsu.edu.cn/是中山大学主页的两个不同镜像，因此均应当被标注为“中山大学”的标准答案，但由于只有后者出现在了搜狗搜索引擎的索引中，因此只把前者作为查询结果返回的搜索引擎自然会被给予不合理的评价。这也正是我们没有在 4.5.2 节的实验中纳入搜狗搜索引擎作为被评价对象的原因。

4.6 本章小结

性能评价一直是信息检索系统研究中的核心问题之一，传统的基于标准语料库、查询样例集合和标准答案集合的Cranfield方法，在处理搜索引擎的性能评价问题上面临着巨大的困难。基于用户查询点击行为挖掘的方法自动化的评价

搜索引擎的查询性能，是本章要解决的主要问题。

为了实现提高检索系统查询处理能力的目的，我们提出基于用户群体行为分析的搜索引擎自动评价方法。该方法利用搜索引擎用户查询、点击行为的宏观分析，自动挑选适用于搜索引擎评价的查询集合，并进一步自动定位对应这些查询的标准答案。由于挑选查询集合和标准答案的过程由计算机自动完成，因此可以及时、准确、客观的反映搜索引擎的真实性能。

利用海量规模搜索引擎网络日志数据的实验结果证明，上述基于用户群体行为分析的搜索引擎自动性能评价方法能够自动提取用户查询需求样例，并准确标注超过95%的导航类查询答案。针对中文搜索引擎性能评价的实验也发现，该自动方法性能评价的结果与手工评价的结果具有很高的一致性。比较传统的手工评价方法而言，自动评价方法具有及时、客观、准确、全面的特点。

综上所述，我们基于用户群体行为分析的思路实现了一种有效的搜索引擎性能自动评价方法，该方法适应大规模网络信息检索系统的应用需求，同时能够客观准确的实现搜索引擎系统的自动性能评价。

第五章 网络信息检索系统架构设计的改进

在第二至四章中，我们分别对网络信息检索系统的数据处理、查询处理和性能评价技术进行了研究和改进。在本章中，我们尝试将本文已述的研究成果加以综合，应用基于用户群体行为分析的数据质量评估技术、查询需求分类技术和效果自动评价技术改进现有网络信息检索系统架构，以达到整合研究成果、提高检索性能的目的。

5.1 引言

第1.2.1节讨论了网络信息检索系统的组成与运行原理，图5-1也给出了网络信息检索系统的一般组成架构。在本文的前几章中，我们已经基于用户群体行为分析的思路，将其中的网络数据处理与查询处理方法进行了改进。在此基础上，对网络信息检索系统的组成架构进行重新审视，我们发现可以依据本文介绍的研究成果对系统架构中的某些模块和模块整合方式进行调整，以使其更好的服务与网络信息检索用户的信息需求。

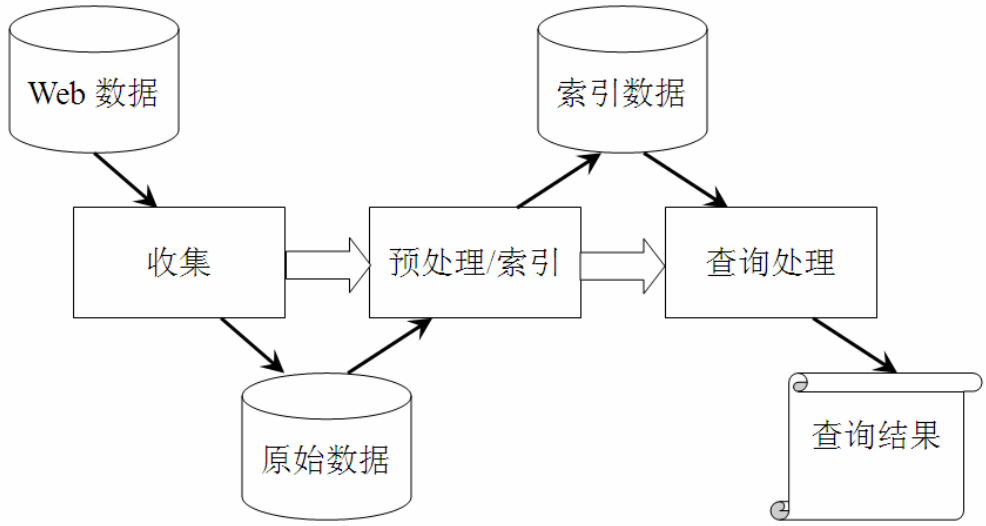

图 5-1 网络信息检索系统的模块组成示意图（引自王斌（2006），有改动）

首先，依据网页质量评估算法对检索系统的索引模块结构进行改进。当前对网络信息检索系统索引模块的研究多集中在对其倒排索引结构、词项组成方

式、词项位置信息利用等方面。但对于网络信息检索系统而言，如何有效地组织网络数以十亿计的网络数据、高效准确的反馈其他模块对于文本统计信息的使用需求才是索引模块的核心任务所在。为此，利用网页质量评估方法将检索系统索引到的网络数据的数据价值进行评估，进而合理的组织不同质量网络数据，以达到同等硬件水平下的最高索引吞吐性能是值得关注的研究方向。

其次，依据用户查询需求分类算法对检索系统的查询处理模块进行改进。查询处理模块的核心任务是将检索系统所返回给用户的相关文档进行合理排序，以方便用户获取所需信息。但正如本文研究的核心内容所指出的，用户查询信息需求传递的瓶颈使得查询处理模块难以获得真实的用户需求信息，而只能把信息需求的检索矮化为利用关键词进行的文本匹配计算。为改进这种情况，我们提出了用户查询需求分类算法，能够根据用户群体的行为信息确定用户信息需求的大致类别。由此，查询处理模块需要进行改进，以便针对不同的检索需求使用不同的检索模型、参数来进行文本相关度计算。

最后，对索引模块和查询处理模块的改进方式加以整合，并引入检索性能监控模块，以用户群体行为分析方面的研究成果对网络信息检索系统架构进行整体改进。对应不同的查询信息需求，查询处理模块不仅需要选择不同的检索模型和检索策略，而且也需要考虑在具有特定特征的网页上进行检索，以提高检索效率、避免计算资源浪费，这就需要对索引模块和查询处理模块从系统架构层次进行整合。而引入基于搜索引擎自动性能评价方法的检索性能监控模块，则是出于能够及时了解检索系统的整体服务性能水平，以便发现和解决运营层面的问题。

下面，我们就分这三个方面介绍对网络信息检索系统架构设计进行改进的具体方式。

5.2 基于网页质量评估的索引模块设计

第二章中提出了一种基于用户群体行为分析的网页质量评估算法，算法能够准确的评估页面成为查询目标页面的可能性，而且能够在减少索引规模的前提下实现网络信息检索系统的检索性能提高。实验证明，网络信息检索系统收集到的数以十亿计的网页中的相当一部分对于绝大部分查询用户的信息获取没有帮助。从实验研究的角度，我们可以尝试摒弃这些对大部分用户无用的页面，

以便在保持检索效果的前提下尽量减少索引规模；但对于实际的网络信息检索应用系统（搜索引擎）而言，丢弃质量评估指标低的网页是非常危险的做法，它会导致这部分网页可能的用户（即使很少）的需求无法满足，而造成搜索引擎潜在用户的流失。这就是搜索引擎即使了解到大部分索引数据被长期闲置，也要花费大量资源保存这部分网页，并提供面向它们的索引服务的原因。

然而，网页之间数据质量和用户需求量的不同造成了其索引访问量的差异，对所有数据都维护同等水平的索引服务效果是没有必要的，也会造成巨大的硬件和内存空间浪费。为了在保证最大范围用户群体查询需求可以得到满足的前提下尽量把有限的资源集中到少数高质量网页索引上，以提高索引系统的整体服务效果，我们提出了基于层次结构的索引模块设计方案。

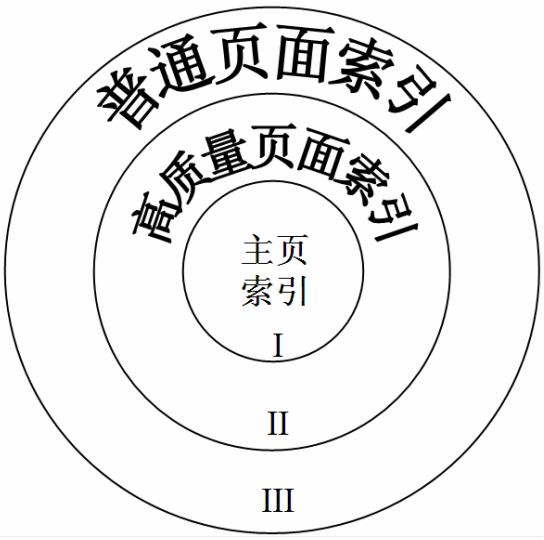

图 5-2 基于网页质量评估的层次索引结构设计示意图

如图5-2所示，可以基于第二章所述的网页质量评估方法设计网络信息检索系统的层次式索引结构。第I级索引由网页中的权威站点首页组成，权威站点一般是指各种网页目录和目录式搜索引擎倾向于收录的站点，其在中文网页中的总体数量按照各网页目录站点（如http://dir.sohu.com/，http://www.hao123.com/等）估计，约为十几万到几十万个页面组成。第II级索引由网页中的高质量页面组成，按照第二章中对中文网页集合的实验结果，这部分网页约占中文网页总数的5%左右，即由约几千万网页组成。第III级索引则是原有的网页索引，即包括所有网页内容的索引。

三级索引之间具有近似的内容蕴含关系，即第II级索引基本包括第I级索引所收录的页面，而第III级索引完全包括了第I级和第II级索引收录的页面。从页面数量上来讲，第III级索引收录的网页远远大于前两级索引，因此其倒排索引条目的数目，以及每个索引条目所对应的存储规模也要远远多于第I级和第II级索引。

检索系统处理用户查询需求时，查询需求首先递交低级别索引，只有当低级索引没有对应的索引条目、或返回的索引条目规模不足（由经验值确定）时，系统方进行高级别索引的访问。由于大部分用户需求（按照第二章实验结论，超过90%的用户需求）可以被第I级和第II级索引所满足，因此较少有机会访问第III级索引。由于访问索引时需要耗费大量的内存空间载入索引条目列表等内容，因此在绝大部分运行时间中，第III级索引可以不保存在内存空间中，这就大大减少了系统运行的空间代价，从而能够提升整体的索引性能。

5.3 基于用户需求分类的查询处理模块设计

第三章中，我们指出了信息事务类与导航类查询的不同信息需求特性，信息事务类一般没有确定的用户查询目标页面，而导航类查询则通常具有确定的目标页面。这种信息需求特性差异也造成了网络信息检索系统在处理两种类型查询时需要采用不同的检索模型和排序策略。精确匹配的检索策略（如基于二元组关系的检索）、链接文本信息和 URL 信息对提高导航类检索的性能有利；而相对模糊的检索策略、查询扩展方法和链接结构信息则会提高信息检索系统处理信息事务类查询的能力。

基于第三章中提出的用户查询信息需求分类方法，我们可以利用用户群体行为信息辨识个体用户查询对应的信息需求类别。这种用户需求类别辨识信息对于网络信息检索系统处理用户查询至关重要，它可以使检索系统针对不同的查询类型使用不同的检索模型，进而提高检索效果。

查询需求类型信息除了可以指导检索系统使用特定的检索模型和策略提高检索效果外，还可以进一步作为选择层次设计的索引结构中索引级别的依据。如导航类查询需求的目标页面唯一，且有较大可能性是站点主页，因此基本可以确定在主页级别索引或高质量页面级别索引中能够找到目标页面；而信息事务类查询需求对应的目标页面较多，主页索引基本不可能满足其需求，因此可

以略过主页级别索引，而直接到高质量页面索引或普通页面索引中进行查询。

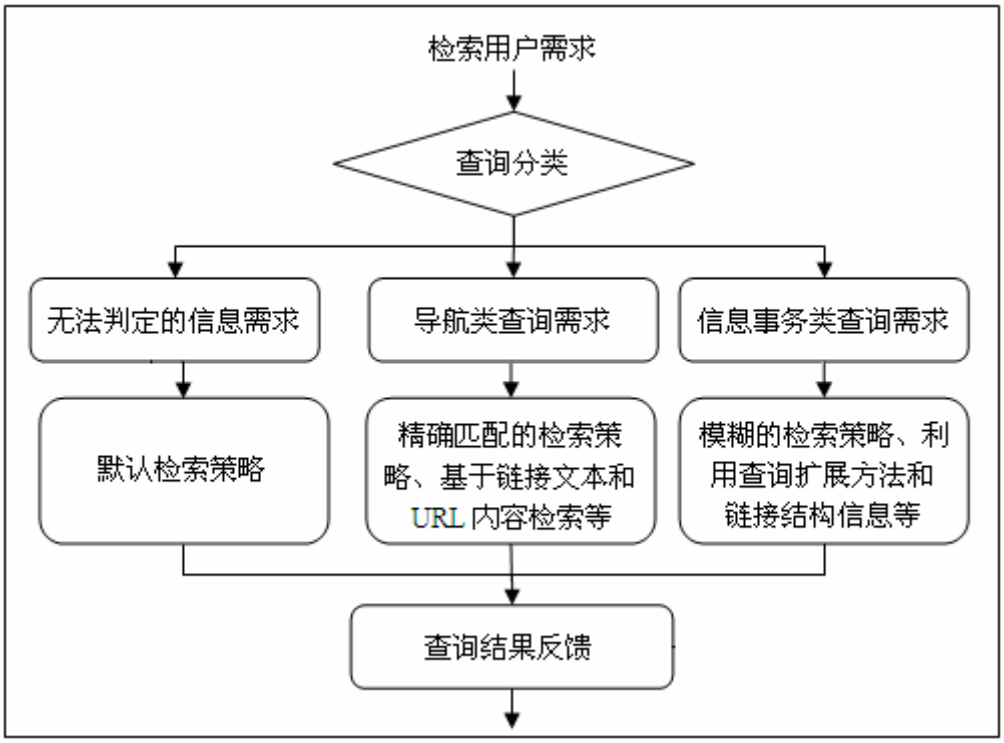

图 5-3 基于查询需求分类方法的查询处理模块设计示意图

如图5-3所示，将用户查询需求分类方法引入查询处理模块之后，查询处理流程以分类结果驱动，对于能够明确区分查询需求类别的查询，按照需求分类结果的不同决定其采用的检索模型与策略；对于其它查询，则按照默认的查询策略进行处理。由于查询需求分类的性能较高（超过80%），因此针对不同类型的查询会进行有的放矢的处理；而对于算法拒识的查询采取了传统的不区别查询需求的检索策略也可以保证其检索效果，因此整体的查询处理效果会比不利用查询需求分类有提高。

5.4 改进的网络信息检索系统架构

上文中，我们根据用户群体行为分析的方法改进了检索系统索引模块和查询处理模块的设计方案，将改进方案加以整合，并引入自动性能评估模块的检索系统整体架构如下图所示：

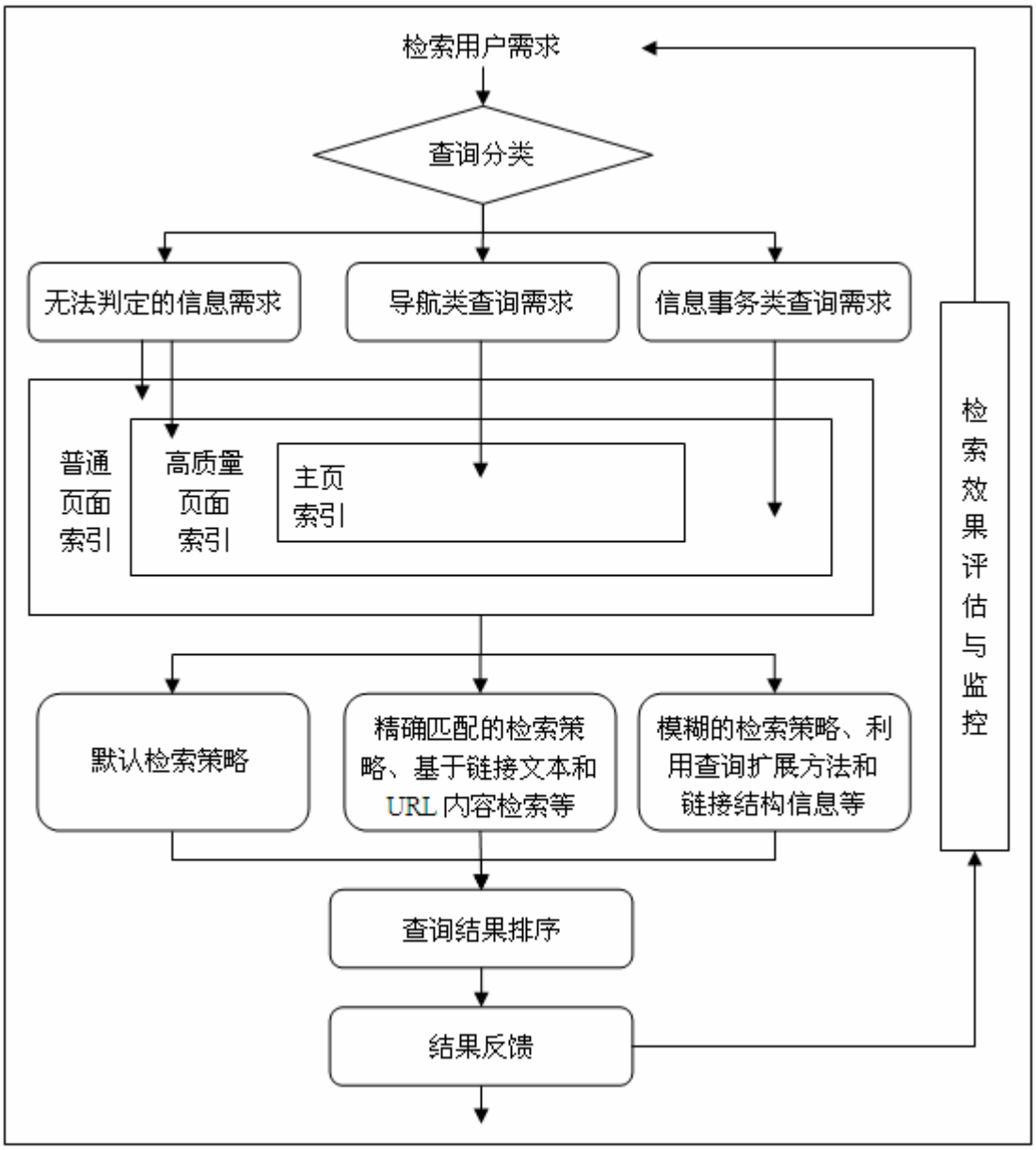

图 5-4 改进的网络信息检索系统架构示意图

图5-4将改进的索引模块与查询处理模块的整合方式做了图示，由于篇幅限制，没有在此架构图中绘入网络信息检索系统的其他运行模块如网络爬虫模块、链接结构分析模块等。

改进的系统架构除了包括上文提到的层次索引结构和以用户查询需求为指导的查询处理模块之外，还包括了检索效果评估与监控模块。这一模块是依据本文第四章所述的搜索引擎自动评估方法设立，其目的在于对网络信息检索系

统的检索质量进行实时评估，进而监控系统运行情况。网络信息检索系统作为一个完整的网络信息服务系统，由于面临着巨大的网络数据和用户查询处理需求，因而是一个庞大的硬、软件服务综合系统。其运营维护方式本身也是一个应用层面需要解决的问题，引入检索效果评估与监控模块后，系统可以利用查询用户的行为信息（表现在用户日志中）对搜索结果的质量进行自动评估，并及时把检索性能的变化加以反馈，以便运营维护人员了解系统运行情况，对可能出现的运行问题加以修正。

改进的系统架构运行的流程可以表述为：用户向检索系统提交查询，系统接到查询请求后，首先判定查询对应的信息需求种类，如果是导航类信息需求，则首先提交I级索引（主页索引）进行检索，并使用精确匹配的检索策略，并使用链接文本、URL内容检索等方法辅助检索；如果是信息事务类信息需求，则首先提交II级索引（高质量页面索引）进行检索，并使用模糊匹配的检索策略，并使用查询扩展、链接结构分析等方法辅助检索；如果无法判定信息需求类型，则同时将查询提交II级索引和III级索引（普通页面索引）进行检索，同时采用一般检索策略进行检索。

与传统信息检索系统架构相比，改进后的架构具有以下特点：

1．充分利用用户群体行为分析得到的研究成果，以用户信息需求为中心，一定程度上解决了网络信息检索系统面临的信息需求传递瓶颈问题。改进后的架构将本文针对信息需求传递瓶颈问题设计的页面质量评估算法、查询需求分类算法和自动性能评估算法加以整合，利用群体用户的宏观行为理解个体用户的信息需求，并在查询处理过程中有的放矢，从而可以从很大程度上缓解检索系统交互方式带来的信息需求传递瓶颈。

2．对原有检索系统架构改变不大，便于充分利用现有架构模块加以改进实现。改进后的层次索引结构比较原有的索引结构扩展了第I级和第II级索引，但这两级索引所包括的文档数目比原有索引规模小得多（分别是索引总量的0.01%和5%左右），因此实现的难度不大。改进后的查询处理算法与原有算法相比使用的内容、结构特征都没有增加，原有的处理流程依然得到了保留，新添加的处理流程主体也可以沿用原有流程。而系统改进所涉及到的索引模块和查询处理模块接口也均可以沿用原有接口。

3．重视对真实规模网络信息检索服务应用环境的考虑，在保证效率的前提下提升性能。系统所引入的页面质量评估模块和查询需求分类模块对应的算法

在实际运行中均为线性复杂度，因而对系统整体运行的效率影响不大，层次索引结构更是能够在保证系统检索性能的前提下提高索引服务效率。此外，自动性能评估模块在检索系统的运营过程中引入了自动效果反馈机制，也大大提高了检索系统运营维护的效率。

5.5 本章小结

本章中，我们提出了一种改进的网络信息检索系统架构，此架构整合了本文所提出的网页质量评估算法、查询需求分类算法和自动性能评估算法，基于用户群体行为分析的思路整合检索系统架构，以缓解网络信息检索中面临的用户信息需求传递瓶颈问题。

网络信息检索系统是当今网络环境中最复杂的应用系统之一，而其巨大的数据处理需求和用户访问量更是对组成的各个环节提出了极高的要求。网络信息检索系统的设计在理论和应用上都是一个繁杂的课题，本文不可能也没有必要在系统设计的每一个环节进行研究与探讨。因此我们将研究的思路重点集中在当前产业界和研究界设计、实现过程中较为欠缺、有待进一步提高的方面，希望能够在这个层面对网络信息检索系统结构的改进发挥自己的作用。

第六章 结论

随着网络信息的爆炸性增长和国家信息化过程的不断推进，网络信息检索工具越来越成为生产和生活中不可缺少的组成部分，在协助人们充分利用网络信息和服务的过程中发挥着重要作用。然而，检索系统现行采用的“关键词查询+选择性浏览”的信息交互方式引起了用户信息需求传递的瓶颈问题，阻碍了网络信息检索系统性能的进一步提高，造成了人们当前进行检索过程中的种种不便。利用对检索用户群体行为的深入挖掘与分析，对互联网的数据环境和用户查询需求进行重新认识，从而使网络信息检索工具能够更加有效的处理检索需求，是本文研究关注的重点。

用户行为分析的方法一直是网络信息检索技术前进的推动力，检索用户的信息需求直接反映在其面向检索系统的查询、浏览、点击行为中，而针对用户群体行为的分析，则可以有效地加深系统对于个体用户信息需求的理解，进而提高系统处理用户查询的性能。本文基于用户群体行为分析的方法，从网络信息检索系统面临的网络数据处理、用户查询处理和检索性能评价问题着手研究，以提出在现有交互方式限制下缓解用户信息需求传递瓶颈问题的解决方案。

首先，针对检索系统面临海量繁杂网络数据收集需求的问题，提出网页数据质量评估算法，综合利用网页查询无关特征和机器学习方法，判定网页满足用户查询需求的可能性，借以保证系统优先的存储、计算能力服务于真正对检索用户信息需求有帮助的网页。随后，针对检索系统无法判断用户查询信息需求的问题，提出基于用户群体行为信息的查询信息需求分类算法，根据用户群体查询的宏观信息判定个体用户当前的信息需求，以保证系统采用有针对性的检索模型和策略处理查询，提升检索系统效果。接下来讨论网络信息检索系统的自动评估问题，以达到客观、准确、实时监控检索效果的目的。最后整合网页质量评估算法、查询需求分类算法和自动性能评估算法，提出一种基于用户群体行为分析方法的检索系统架构设计方案。

本文的主要创新点与主要贡献包括如下方面：

1．提出网络信息检索系统评价网页质量的标准，即按照满足用户信息需求程度的高低来评判网页的有用程度。页面质量评估不能脱离用户的查询信息需求存在，对于检索系统的应用而言，有可能成为用户查询目标的网页才是高质量的网页。

2．基于海量规模网页数据，分析高质量网页（查询目标页面）与普通网页的页面特征差异。发现查询目标页面（包括关键资源页面与特殊需求页面）具有不同于普通页面的查询无关特征，这些特征既包括链接结构特征（如入链接个数特征、*PageRank* 特征），也包括其他信息来源的特征（如 URL 长度特征、页面编码特征）。

3．提出一种基于贝叶斯学习的页面质量评估算法，网页质量由其成为查询目标页面的可能性来衡量，算法基于语料库统计方法计算网页成为查询目标页面的概率，并依照其大小对网页质量进行评估。算法在海量规模中文网页数据上的质量评估实验中取得了优于链接分析算法的评估效果，基于 TREC2004 实验数据的检索实验也证明进行页面质量评估有助于检索系统在控制索引规模的前提下提升性能。

4．对网络信息检索用户的行为特征进行分析和挖掘，发现中文网络信息检索用户具有“懒于输入，勤于点击”的行为习惯。这种习惯加重了搜索引擎获取用户信息需求的瓶颈问题，但也带来了宝贵的用户点击行为信息，有助于利用点击行为分析提高搜索引擎的查询处理能力。

5．对可能用于搜索引擎用户需求分类的查询特征进行了可靠性、分类能力和可用性的分析。分析结果证明，查询历史特征是一种分类能力强、可以用于绝大多数用户查询的分类特征。而查询需求分类传统方法经常采用的查询自身特征、查询反馈特征、链接文本特征等则由于实用性、分类能力等方面的原因不适于网络信息检索系统的大规模应用。

6．基于用户群体行为分析的方法对用户的查询历史特征进行了深入的挖掘和分析，提出两个可以用于查询需求分类的新行为特征：“前 N 次点击满足用户需求率”和“前 N 位结果满足用户需求率”。基于查询样例训练集合的实验说明，这两个特征对于查询需求分类具有较好的效果。

7．提出一种基于决策树学习的用户查询需求分类算法，利用海量规模搜索引擎网络日志数据的实验结果证明，算法能够正确辨识 80%以上的用户查询信息需求，其辨识能力超过查询需求分类传统方法的分类效果，这使得通过用户群体的查询点击行为分析当前查询对应的信息需求类别成为可能。

8．分析了利用 Cranfield 方法进行搜索引擎性能评估的可能性，同时考察用户查询信息需求与性能评价之间的关系，得出了在当前的实际应用条件限制下，搜索引擎性能自动评价的对象只能限制于导航类检索的结论。

9．提出一种基于用户群体行为分析的搜索引擎自动评价方法，该方法基于传统的 Cranfield 评价框架，实现了查询样例集合和标准答案集合的自动化生成。利用海量规模搜索引擎网络日志数据的实验结果证明，此方法能够自动提取用户查询需求样例，并准确标注超过 95%的导航类查询答案。针对中文搜索引擎性能评价的实验也发现，该自动方法性能评价的结果与手工评价的结果具有很高的一致性。

10．提出了一种改进的网络信息检索系统架构，此架构整合了本文所提出的网页质量评估算法、查询需求分类算法和自动性能评估算法，基于用户群体行为分析的思路整合检索系统架构，以缓解网络信息检索中面临的用户信息需求传递瓶颈问题。

网络信息检索是一个实证意味很强的新兴研究方向，任何一种算法的有效性最终都要经受真实网络数据环境和用户需求的考验。在研究过程中，我们一直尽量保证研究所使用的实验数据与真实应用环境保持高度的一致性。网页质量评估实验过程中所采用的海量规模网页数据，包括了中文网络环境中近 5%的网页；查询需求分类实验和搜索引擎性能评估中所采用的用户查询日志数据，是采集于真实中文搜索引擎的近半年的超过 5 亿条查询、点击日志。算法和系统结构设计的每一个方面，都尽量考虑网络应用环境的需求，以保证我们的研究结果能够为网络信息检索的实际应用水平的提高做出贡献。

本文的相关工作证明，利用网络用户群体的行为分析进行信息检索方面的研究，是一个可行的途径。在未来的工作中，我们将进行更深层次的用户行为分析和挖掘，以解决网络信息检索中面临的各种问题。我们将尝试对用户查询的语言模型进行分析，借以更好的指导检索系统处理用户查询的过程，以及进行查询纠错、查询提示方面的研究；我们将尝试利用用户行为和网页特征分析的方法对网络上充斥的网页垃圾进行辨识，借以保证网络信息检索系统查询结果的质量；我们还将尝试对用户群体访问网络页面的行为特性进行分析，对整个网页的用户关注程度和生命周期变化方式加以了解，借以加深我们对互联网络的认识，并指导网络信息检索系统的数据收集与处理过程。总之，利用用户群体行为分析的方法进行完善网络信息检索系统的研究依然大有可为。

网络信息环境已经成为历史上所存在过的最大规模的知识宝库，人类几千年文明所积累的信息财富大都蕴含其中。网络信息本身的繁杂特性决定了信息和协助人们利用信息的工具必定形影相随，这是网络信息检索工具担负的重大

责任，也是相关研究人员的重大机遇与挑战。我们在本文中利用用户群体行为分析方法所进行的研究，仅仅是向发展、完善网络信息检索工具的道路上迈出的微不足道的一步，希望能够为这项宏伟的事业尽一点绵薄之力。

参考文献

Amento B, Terveen L and Hill W. 2000. Does authority mean quality? Predicting expert quality ratings of Web documents. In: Nicholas J. Belkin, Peter Ingwersen and Mun-Kew Leong, eds. SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2000. New York: ACM Press. 296-303.

Amitay, E., Carmel, D., Lempel, R., and Soffer, A. 2004. Scaling IR-system evaluation using term relevance sets. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 10-17.

Bar-Yossef, Z. and Rajagopalan, S. 2002. Template detection via data mining and its applications. In Proceedings of the 11th international Conference on World Wide Web (Honolulu, Hawaii, USA, May 07 - 11, 2002). WWW '02. ACM Press, New York, NY, 580-591.

Benczur A A, Csalogany K, Sarlos K, and Uher M. 2005. Spamrank - fully automatic link spam detection. In First International Workshop on Adversarial Information Retrieval on the Web.

Beitzel S. M., Jensen E. C., Chowdhury A., and Grossman D. 2003. Using titles and category names from editor-driven taxonomies for automatic evaluation. In Proceedings of the twelfth international conference on Information and knowledge management, 17-23.

Bharat, K. and Henzinger, M. R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 104-111.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C.J. 1984. Classification and Regression Trees, Wadsworth(California): Belmont.

Brewington, B. E. and Cybenko, G. 2000. How dynamic is the Web?. In Proceedings of the 9th international World Wide Web Conference on Computer Networks : the international Journal of Computer and Telecommunications Netowrking (Amsterdam, The Netherlands). North-Holland Publishing Co., Amsterdam, The Netherlands, 257-276.

Brin, S., Davis, J., and García-Molina, H. 1995. Copy detection mechanisms for digital documents. SIGMOD Rec. 24, 2 (May. 1995), 398-409.

Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the Seventh international Conference on World Wide Web 7 (Brisbane, Australia). P. H. Enslow and A. Ellis, Eds. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 107-117.

Broder, A. Z., Glassman, S. C., Manasse, M. S., and Zweig, G. 1997. Syntactic clustering of the Web. Computer Network ISDN Systems 29, 8-13 (Sep. 1997), 1157-1166.

Broder, A. 2002. A taxonomy of web search. SIGIR Forum 36, 2 (Sep. 2002), 3-10.

Broder, A. Z., Lempel, R., Maghoul, F., and Pedersen, J. 2006. Efficient PageRank approximation via graph aggregation. Information Retrieval 9(2):123-138.

Buchner, A.G., Baumgarten, M., Mulvenna, M.D., Böhm, R., Anand, S. S. 2000. Data Mining and XML: Current and Future Issues, International Conference on Web Information System Engineering, 131-135.

Buckley, C. and Voorhees, E. M. 2004. Retrieval evaluation with incomplete information. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 25-32.

Byrne, M. D., John, B. E., Wehrle, N. S., & Crow, D. C. 1999. The tangled web we wove: A taskonomy of WWW use. In Human Factors in Computing Systems: Proceedings of CHI 99. 544-551.

Cai, D., Yu, S., Wen, J., and Ma, W. 2004. Block-based web search. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 456-463.

Cai, D., Yu, S., Wen, J., and Ma, W. 2004b. Block-level Link Analysis. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 440-447.

Campbell, D. M., Chen, W. R., and Smith, R. D. 2000. Copy Detection Systems for Digital Documents. In Proceedings of the IEEE Advances in Digital Libraries 2000 (May 22 - 24, 2000). ADL. IEEE Computer Society, Washington, DC, 78.

Catledge, L. D., Pitkow, J. E. 1995. Characterizing Browsing Strategies in the World-Wide Web. Computer Networks and ISDN Systems, 27, 1065-1073.

Chakrabarti, S., Byron E. Dom, Gibson, D. Kumar, R. Raghavan, P. Rajagopalan, S. and Tomkins, A. 1998. Experiments in topic distillation. In Proceedings of the ACM

SIGIR Workshop on Hypertext Information Retrieval on the Web (Melbourne, Australia). ACM, New York.

Cho, J. and Garcia-Molina, H. 2000. The Evolution of the Web and Implications for an Incremental Crawler. In Proceedings of the 26 th International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt. 527-534.

Cho, J. and Garcia-Molina, H. 2000b. Synchronizing a database to improve freshness. In Proceedings of the 2000 ACM SIGMOD international Conference on Management of Data (Dallas, Texas, United States, May 15 - 18, 2000). SIGMOD '00. ACM Press, New York, NY, 117-128.

Cho, J. and Garcia-Molina, H. 2003. Estimating frequency of change. ACM Transactions on Internet Technology 3, 3 (Aug. 2003), 256-290.

Chowdhury, A., and Soboroff, I. 2002. Automatic Evaluation of World Wide Web Search Services. In Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 421-422.

Cleverdon, C. W., Mills, J. & Keen, E.M. 1966. An inquiry in testing of information retrieval systems. (2 vols.). Cranfield, U. K.: Aslib Cranfield Research Project, College of Aeronautics.

Cockburn, A., & Jones, S. 1996. Which way now? Analysing and easing inadequacies in WWW navigation. International Journal of Human-Computer Studies, 45, 105-129.

Craswell, N., Hawking, D., and Robertson, S. 2001. Effective site finding using link anchor information. In Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 250-257.

Craswell, N., Hawking, D. 2004. Overview of the TREC-2004 Web track. In NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004).

Craswell, N., Robertson, S., Zaragoza, H., and Taylor, M. 2005. Relevance weighting for query independent evidence. In Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM Press, New York, NY, 416-423.

Davison, B. D. 2000. Topical locality in the Web. In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Athens, Greece, July 24 - 28, 2000). SIGIR '00. ACM Press, New York, NY, 272-279.

Davision, B.D. 2000b. Recognizing Nepotistic Links on the Web. Presented at the AAAI-2000 Workshop on Artificial Intelligence for Web Search, and published in Artificial Intelligence for Web Search, Technical Report WS-00-01, pp. 23-28.

DeComite F., Denis F., and Gilleron R. 1999. “Positive and Unlabeled Examples Help Learning,” Proceedings 11th Int’l Conf. Algorithmic Learning Theory (ALT ’99), 219-230

Denis F. 1998. “PAC Learning from Positive Statistical Queries,” Proceedings 10th Int’l Conf. Algorithmic Learning Theory (ALT ’99), 112-126.

Gulli, A. and Signorini, A. 2005. The indexable web is more than 11.5 billion pages. In Special interest Tracks and Posters of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM Press, New York, NY, 902-903.

Gyongyi Z., Garcia-Molina H., Pedersen J. 2004. Combating Web Spam with TrustRank. Technical Report, Stanford University.

Haveliwala, Taher H.. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, IEEE Transactions on Knowledge and Data Engineering,

Hawking, D., Craswell, N., Thistlewaite, P. and Harman, D. 1999. Results and challenges in Web search evaluation, in Proceedings of the Eighth International World Wide Web Conference, North Holland. 1321-1330.

Hawking D. and Craswell N. 2002. Overview of the TREC-2002 web track. In The eleventh Text Retrieval Conference (TREC-2002), volume 11. National Institute of Standards and Technology, NIST, 2003.

Hawking D. and Craswell N. 2003. Overview of the TREC-2003 Web track. In NIST Special Publication: SP 500-255, The Twelfth Text Retrieval Conference (TREC 2003).

Hawking D. and Craswell N. 2005. Very Large Scale Retrieval and Web Search, in Ellen Voorhees and Donna Harman, TREC: Experiment and Evaluation in Information Retrieval, MIT press.

Henzinger, M.R., Motwani, R., Silverstein, C. 2003. Challenges in Web Search Engines. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (2003) 1573-1579.

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. 1998. Real life information retrieval: A study of user queries on the Web. SIGIR Forum, 32(1), 5-17.

Joachims T. 2002. Evaluating Retrieval Performance Using Clickthrough Data. In Proceedings of the SIGIR Workshop on Mathematical/FormalMethods in Information Retrieval.

Kang, I. and Kim, G. 2003. Query type classification for web document retrieval. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 64-71.

Kent, A., Berry, M., Leuhrs, F.U. and Perry, J.W. 1955. Machine literature searching VIII. Operational criteria for designing information retrieval systems. American Documentation, 6(2): 93-101.

Kleinberg. J.M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604-632.

Kononenko, I., Bratko, I., & Roskar, E. 1984. Experiments in automatic learning of medical diagnostic rules (Technical report). Jozef Stefan Institute, Ljubljana, Yugoslavia.

Konopnicki, D. and Shmueli, O. 1995. W3QS: A Query System for the World-Wide Web. In Proceedings of the 21th international Conference on Very Large Data Bases (September 11 - 15, 1995). U. Dayal, P. M. Gray, and S. Nishio, Eds. Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, 54-65.

Kraaij, W., Westerveld, T., and Hiemstra, D. 2002. The Importance of Prior Probabilities for Entry Page Search. In Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 27-34.

Kushmerick N. 1999. Learning to remove Internet advertisements, in Proceedings of the Third International Conference on Autonomous Agents (Agents'99), 175-181.

Lee, U., Liu, Z., and Cho, J. 2005. Automatic identification of user goals in Web search. In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM Press, New York, NY, 391-400.

Letouzey, F, Denis, F and Gilleron, R. 2000. Learning from positive and unlabeled examples. In Proceedings of the 11th International Conference on Algorithmic Learning Theory (Lecture Notes in Computer Science, 1968). Berlin: Springer, pp.71--85.

Lin, S. and Ho, J. 2002. Discovering informative content blocks from Web documents. In Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM Press, New York, NY, 588-593.

Liu Y., Zhang M., Ma S. 2004. Effective topic distillation with key resource pre-selection.In:Myaeng SH,et al.,eds. Proceedings of the AIRS2004 .LNCS 3411, Berlin/Heidelberg: Springer-Verlag, 2005. 129-140.

Liu, Y., Wang, C., Zhang, M., and Ma, S. 2005. Web data cleansing for information retrieval using key resource page selection. In Special interest Tracks and Posters of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM Press, New York, NY, 1136-1137.

Liu Y., Wang C., Zhang M., Ma S. 2005b. Learning-based Web Data Cleansing for Information Retrieval, Journal of Computational Information Systems.2005(12).

Liu Y., Zhang M., Ru L., Ma S. 2006. Automatic Query Type Identification Based on Click Through Information, AIRS 2006, LNCS 4182, pp. 593–600.

Lyman, P. and Hal R. Varian, 2003. How Much Information, 2003 [EB/OL]. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on April 2th, 2004.

Manber U. 1994. Finding Similar Files in a Large File System, Usenix Winter 1994 Technical Conference, San Francisco (January 1994), pp. 1-10.

Manevitz, L. M. and Yousef, M. 2002. One-class svms for document classification. J. Mach. Learn. Res. 2 (Mar. 2002), 139-154.

McSherry, F. 2005. A uniform approach to accelerated PageRank computation. In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM Press, New York, NY, 575-582.

Mendelzon, A., Mihaila, G., Milo, T. 1997. Querying the World Wide Web. International Journal on Digital Libraries 1:54-67.

Mingers, J. 1989. An empirical comparison of pruning methods for decision-tree induction. Machine Learning, 4 (2):227-243.

Mitchell T.M. 1997. Machine Learning. New York: McGraw-Hill, 1997. 55-64.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. 1999. Cognitive Strategies in Web Searching. Proceedings of the 5th Conference on Human Factors & the Web.

Nuray, R. and Can, F. 2003. Automatic ranking of retrieval systems in imperfect environments. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 379-380.

Ottenstein K. J. 1977. An Algorithmic Approach to the Detection and Prevention of Plagiarism, SIGCSE Bulletin 8(4), pp. 30-41.

Quinlan, J.R. 1986. Induction of decision trees. Machine Learning, 1(1): 81-106.

Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M.M. & Gatford, M. 1994. Okapi at TREC-3 (pp. 109-126). NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC-3), Washington: Department of Commerce and National Institute of Standards and Technology.

Rose, D. E. and Levinson, D. 2004. Understanding user goals in web search. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 13-19.

Saracevic, T. 1995. Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 138-146.

Shivakumar N, Garcia-Molina H. 1995. SCAM: A copy detection mechanism for digital documents. In: Proceedings of the 2nd International Conference in Theory and Practice of Digital Libraries (DL'95).

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6-12.

Singhal, A. 2005. Challenges in running a commercial search engine. In Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM Press, New York, NY, 432-432.

Soboroff, I., Nicholas, C., and Cahan, P. 2001. Ranking retrieval systems without relevance judgments. In Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 66-73.

Song, R., Liu, H., Wen, J., and Ma, W. 2004. Learning block importance models for web pages. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 203-211.

Spoerri A. 1995. InfoCrystal: A Visual Tool For Information Retrieval [EB/OL]. Ph.D. Research and Thesis at Massachusetts Institute of Technology. On line at http://www.scils.rutgers.edu/~aspoerri/InfoCrystal/InfoCrystal.htm

Sullivan D. 2005. Search Engine Sizes [EB/OL]. Retrieved from search engine watch web site http://searchenginewatch.com/reports/article.php/2156481 .

Sullivan D. 2006. Searches Per Day [EB/OL]. Retrieved from search engine watch web site http://searchenginewatch.com/reports/article.php/2156461 .

Tang, M.-C. Sun, Y. 2003. Evaluation of Web-Based Search Engines Using User Effort Measures. Library and Information Science Research Electronic Journal 13(2).

Tauscher, L., & Greenberg, S. 1997. How people revisit web pages: Empirical findings and implications for the design of history systems. International Journal of Human-Computer Studies, 47, 97-137.

Tax, D. M. and Duin, R. P. 2002. Uniform object generation for optimizing one-class classifiers. J. Machine Learning Research. 2 (Mar. 2002), 155-173.

Upstill, T., Craswell, N., and Hawking, D. 2003. Query-independent evidence in home page finding. ACM Transaction on Information Systems 21, 3 (Jul. 2003), 286-313.

Voorhees. E. M. 2001. The philosophy of information retrieval evaluation. In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum, (CLEF 2001), pages 355-370.

Weisstein E. W. 2005. K-Means Clustering Algorithm [EB/OL]. From MathWorld Web Site, Online at http://mathworld.wolfram.com/K-MeansClusteringAlgo-rithm.html

Yi, L., Liu, B., and Li, X. 2003. Eliminating noisy information in Web pages for data mining. In Proceedings of the Ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Washington, D.C., August 24 - 27, 2003). KDD '03. ACM Press, New York, NY, 296-305.

Yu, H., Han, J., and Chang, K. C. 2004. PEBL: Web Page Classification without Negative Examples. IEEE Transactions on Knowledge and Data Engineering 16, 1 (Jan. 2004), 70-81.

Zhang M., Song R., Lin C., Ma S., Jiang Z., Jin Y., Liu Y., Zhao L. 2002. THU TREC 2002: Novelty Track Experiments. TREC 2002

Zukerman, I., Albrecht, D. W., and Nicholson, A. E. 1999. Predicting users' requests on the WWW. In Proceedings of the Seventh international Conference on User Modeling (Banff, Canada). J. Kay, Ed. Springer-Verlag New York, Secaucus, NJ, 275-284.

刘悦，杨志峰，程学旗，王斌. 2002. 利用链接分析技术提高搜索引擎查找质量的研究。微电子学与计算机，2002 年第 5 期，18-21 页。

孟涛，闫宏飞，王继民. 2005. 一个增量搜集中国 Web 的系统模型及其实现。《清华大学学报（自然科学版）》，2005.9，vol.45， No.S1，1882-1886

王斌. 2006. 中国科学院“现代信息检索”研究生课程讲义。

余慧佳，刘奕群，张敏，茹立云，马少平. 2007. 基于大规模日志分析的网络搜索引擎用户行为研究. 中文信息学报 Vol. 21(1): pp. 109-114, 2007.

张俊林，刘洋，孙乐，刘群. 2006. 2005 年度 863 信息检索评测方法研究和实施 ，中文信息学报增刊，Vol. 20(s): 1-6.

张敏，马少平，宋睿华. 2005. DF 还是 IDF?主特征模型在 Web 信息检索中的使用，软件学报，Vol. 16(5): 1012-1020.

赵乐，岑荣伟，王灿辉，齐伟，金奕江，张敏，马少平. 2006. 清华 THUIR 2005 年 863 信息检索评测，中文信息学报增刊，Vol. 20(s): 91-95。

中国互联网络信息中心(CNNIC). 2004. 第 14 次中国互联网络发展状况统计报告，2004 年 7 月.

中国互联网络信息中心(CNNIC). 2004b. 2003 年中国互联网络信息资源数量调查报告，2004 年 4 月.

中国互联网络信息中心(CNNIC). 2005. 第 16 次中国互联网络发展状况统计报告，2005 年 7 月.

中国互联网络信息中心(CNNIC). 2005b. 2004 年中国互联网络信息资源数量调查报告，2005 年 4 月.

中国互联网络信息中心(CNNIC). 2006. 2005 年中国互联网络信息资源数量调查报告，2006 年 3 月.