

信息科学与技术丛书 移动与嵌入式开发系列

- 国内第一部全面介绍移动搜索技术的专业书籍
- 获国家自然科学基金、中国博士后科学基金、四川省教育厅自然科学基金重点项目支持

王 鹏 主编

陈高云 安俊秀 王俊岭 等编著

移动搜索引擎 原理与实践

本书涵盖的技术:

- 移动通信技术
- 搜索引擎技术
- 数据挖掘技术
- 并行计算技术



机械工业出版社
CHINA MACHINE PRESS

信息科学与技术丛书·移动与嵌入式开发系列

移动搜索引擎原理与实践

王 鹏 主编

陈高云 安俊秀 王俊岭 等编著



机械工业出版社

本书对从事移动搜索引擎研发所需技术进行了全面介绍,内容涵盖广泛,面向实战,将技术与市场紧密结合,对移动通信技术、搜索引擎技术、3G技术、WAP及短消息技术、数据挖掘技术、并行计算技术进行了详细的介绍。

本书可作为从事移动搜索技术研究、设计和开发的科技人员的重要参考书,也可作为从事移动搜索技术研究的学者及研究生的参考书籍。

图书在版编目(CIP)数据

移动搜索引擎原理与实践 / 王鹏主编. —北京:机械工业出版社, 2009.1

(信息科学与技术丛书·移动与嵌入式开发系列)

ISBN 978-7-111-26129-2

I. 移… II. 王… III. 移动通信—通信网—情报检索 IV. G354.4

中国版本图书馆CIP数据核字(2009)第010384号

机械工业出版社(北京市百万庄大街22号 邮政编码100037)

策划编辑:李萌

责任编辑:李萌

责任印制:李妍

保定市 中画美凯印刷有限公司印刷

2009年2月·第1版第1次印刷

184mm×260mm·16.5印张·315千字

0001—3000册

标准书号:ISBN 978-7-111-26129-2

定价:36.00元

凡购本书,如有缺页、倒页、脱页,由本社发行部调换

销售服务热线电话:(010) 68326294 68993821

购书热线电话(010) 88379639 88379641 88379643

编辑热线电话(010) 88379753 88379739

封面无防伪标均为盗版

出版说明

随着信息科学技术的迅速发展,人类每时每刻都会面对层出不穷的新技术、新概念。毫无疑问,在节奏越来越快的工作和生活中,人们需要通过阅读和学习大量信息丰富、具备实践指导意义的图书,来获取新知识和新技能,从而不断提高自身素质,紧跟信息化时代发展的步伐。

众所周知,在计算机硬件方面,高性价比的解决方案和新技术的应用一直备受青睐;在软件技术方面,随着计算机软件的规模和复杂性与日俱增,软件技术受到不断挑战,人们一直在为寻求更先进的软件技术而奋斗不止。目前,计算机在社会生活中日益普及,随着因特网延伸到人类世界的层层面面,掌握计算机网络技术和理论已成为大众的文化需求。由于信息科学技术在电工、电子、通信、工业控制、智能建筑、工业产品设计与制造等专业领域中已经得到充分、广泛的应用,所以这些专业领域中的研究人员和工程技术人员越来越迫切需要汲取自身领域信息化所带来的新理念和新方法。

针对人们对了解和掌握新知识、新技能的热切期待,以及由此促成的人们对语言简洁、内容充实、融合实践经验的图书迫切需要的现状,机械工业出版社适时推出了“信息科学与技术丛书”。这套丛书涉及计算机软件、硬件、网络、工程应用等内容,注重理论与实践相结合,内容实用,层次分明,语言流畅,是信息科学与技术领域专业人员不可或缺的图书。

现今,信息科学与技术的发展可谓一日千里,机械工业出版社欢迎从事信息技术方面工作的科研人员、工程技术人员积极参与我们的工作,为推进我国的信息化建设工作做出贡献。

机械工业出版社

前言

移动通信技术的高速发展正在不断地改变人们的生活和工作方式,也在改变着人们的信息搜索方式。自从 1998 年两位斯坦福大学的博士生 Larry Page 和 Sergey Brin 创立了 Google 以来,搜索引擎已逐渐成为人们获取信息的主要方式。几乎每一个上过网的用户都使用过搜索引擎,随着使用搜索引擎的人数增多,搜索引擎已经成为一种重要的媒体,许多企业通过搜索引擎让世界各地的人了解自己。但普通的网络搜索引擎已无法适应现代快速的工作节奏,随着无线通信终端的普及,移动搜索引擎将成为人们未来获取信息的主要工具。具备搜索功能的无线终端将成为未来信息社会的导航工具。Google、Baidu 等搜索引擎企业已认识到了移动搜索这一巨大的市场,我国也有多家企业涉足移动搜索行业,如 Cgogo、手索科技等企业,都在探索移动搜索的核心技术及市场模式,希望在未来市场中占有一席之地。由于移动搜索技术与传统搜索技术有一定的区别,从事移动搜索技术研究的企业现在迫切需要一本介绍移动搜索技术的专著。本书撰写的目的,就是向读者介绍从事移动搜索引擎技术研发所需要的基本知识,为从事这一领域研发的科技人员提供参考,从而促进这一新兴行业的快速发展。

本书的特点是技术与市场分析并重,内容全面,涵盖移动通信技术、搜索引擎技术、数据挖掘技术、并行计算技术等多个相关领域,可以为技术人员提供全面的技术参考,是国内第一部全面介绍移动搜索技术的专业书籍。

本书由王鹏博士主编、统稿并完成第 1~3 章的写作,陈高云副教授完成第 4~6 章的写作,余惠政完成第 7 章的写作,安俊秀副教授完成第 8~10 章的写作,谢千河和王俊岭对全书图表进行了审校。

本书作者长期从事移动搜索技术研发及市场运营,书中的许多内容来自实际开发项目。

本书在撰写过程中得到了成都信息工程学院并行计算实验室、电子科技大学、北京大学天网搜索引擎项目组的大力支持。

本书的撰写还得到：国家自然科学基金（60702075）、中国博士后科学基金一等（20070410385）、四川省教育厅自然科学重点项目（07ZA014）、成都信息工程学院发展基金（KYTZ200819，CSRF200701）等基金的资助。在此一并表示衷心感谢

北京大学王继民教授，成都信息工程学院郑皎凌老师、郭本俊老师为本书提供的重要帮助以及成都信息工程学院软件工程系李裕森、王警宇、张瑶、何子俊、吕爽、刘杨洋、刘杨、程兵、陈从兵、徐尧、付涛等同学在本书的撰写过程中也做了很多工作。

由于移动搜索引擎技术是一个新兴的技术领域，而且随着科技的发展，移动搜索引擎技术也在高速发展，加之作者水平有限，书中必然还有很多不足之处，请读者批评指正。

王 鹏

成都信息工程学院并行计算实验室
电子科技大学计算机科学与工程学院

目 录

出版说明

前言

第 1 章 移动搜索引擎概述	1
1.1 引言	1
1.2 移动搜索引擎的特点	3
1.3 移动搜索引擎的分类	4
1.4 移动搜索引擎市场现状	4
1.5 移动搜索引擎的关键技术简介	5
1.6 移动搜索与 3G	6
1.7 小结	7
第 2 章 移动通信技术概述	8
2.1 移动通信的分类	8
2.2 全球移动通信 (GSM) 网络	10
2.2.1 移动台 (MS)	11
2.2.2 基站子系统 (BSS)	12
2.2.3 网络子系统 (NSS)	12
2.2.4 操作支持子系统 (OSS)	14
2.2.5 移动搜索引擎与 GSM 网络	14
2.3 GSM 的升级 GPRS	15
2.3.1 GPRS 的特点	15
2.3.2 GPRS 的工作原理	17
2.3.3 GPRS 的网络结构和网元功能	18
2.3.4 GPRS 各个接口协议栈的功能	20
2.4 CDMA (Code Division Multiple Access)	23
2.4.1 CDMA 技术的标准化	23

2.4.2 CDMA 的优势	24
2.5 短信方式下的移动搜索引擎协议	25
2.5.1 SMS (Short Messaging Service)	25
2.5.2 EMS (Enhanced Messaging Service)	26
2.5.3 MMS (Multimedia Message Service)	26
2.5.4 SMS、EMS、MMS 三者技术对比	28
2.6 WAP	29
2.6.1 WAP 的特点	30
2.6.2 WAP 的优势及局限性	32
2.6.3 WAP 的发展前景	33
2.7 CMWAP 和 CMNET	33
2.7.1 CMWAP 与 CMNET 的区别	34
2.7.2 CMWAP 与 CMNET 服务适用范围	35
2.8 小结	37
第 3 章 网页抓取技术	38
3.1 基于 WAP 的搜索引擎系统结构	38
3.2 WAP 爬虫基本原理	40
3.2.1 网页抓取	40
3.2.2 内容提取	41
3.2.3 爬虫的难题	43
3.2.4 爬虫更新周期	43
3.3 WAP 爬虫的关键技术	43
3.3.1 从种子站点开始逐层抓取 (逐层抓取策略)	44
3.3.2 不重复抓取策略	44
3.3.3 网页抓取优先策略	45
3.3.4 网页重访策略	53
3.3.5 Robots 协议	53
3.3.6 抓取提速策略	55
3.4 构建垂直搜索引擎	56
3.4.1 小天音乐垂直搜索引擎整体处理流程	56
3.4.2 小天音乐垂直搜索引擎服务器部分源代码	59
3.5 小结	60
第 4 章 网页信息预处理	61
4.1 结构化的数据和半结构化的数据	62
4.1.1 XML——最常见的结构化数据	62
4.1.2 半结构化的数据	63
4.2 网页信息结构化	65
4.2.1 网页信息结构化的基本流程	65

4.2.2 建立 DOM 树的原理	66
4.2.3 移动网页的信息结构化	69
4.3 中文分词	71
4.3.1 中文分词的概念	71
4.3.2 中文分词的难点	72
4.3.3 基于字典的分词方法	73
4.3.4 用统计学的方法解决新词识别问题	76
4.3.5 主流中文搜索引擎分词效果比较	78
4.3.6 典型的中文分词系统	78
4.4 小结	81
第 5 章 建立索引和实现查询	83
5.1 搜索引擎中的索引简介	83
5.1.1 基于索引的搜索引擎分类	84
5.1.2 索引器在搜索引擎中的作用	84
5.2 建立索引	85
5.2.1 倒排文件索引 (Inverted File Index) 的建立	85
5.2.2 后缀数组	90
5.2.3 签名文件索引 (Signature File Index)	95
5.3 查询服务	96
5.3.1 查询器的实现原理	97
5.3.2 查询器的分页机制	98
5.3.3 查询扩展与结果集的动态分类	98
5.4 查询结果排序	100
5.4.1 PageRank 算法	100
5.4.2 HillTop 算法	104
5.4.3 提升网页排序的方法	104
5.5 索引设计的典型案例	106
5.6 小结	107
第 6 章 WML 转换	108
6.1 标记语言简介	109
6.1.1 HTML	109
6.1.2 XML	110
6.1.3 XHTML	112
6.1.4 WML	113
6.2 WAP 体系与结构	114
6.2.1 WAP 体系结构	114
6.2.2 万维网模型	117
6.2.3 WAP 模型	118

6.3 HTML 到 WML 的转换	120
6.3.1 标记语言转换概述	120
6.3.2 HTML 与 WML 的主要语法标记	122
6.3.3 HTML 与 WML 的文档结构对比	124
6.3.4 HTML 与 WML 的转换	125
6.3.5 正则表达式和 Java 开源项目	127
6.4 衡量搜索引擎性能	137
6.4.1 衡量搜索引擎的性能指标	137
6.4.2 搜索引擎的评价方式	139
6.5 小结	141
第 7 章 基于短信的移动搜索引擎	142
7.1 概述	142
7.2 互联网和短信并存的效应	143
7.3 短信平台	145
7.4 短信接入	146
7.4.1 协议规范	146
7.4.2 协议流程	147
7.4.3 接口更改	157
7.4.4 接口开发	159
7.5 短信搜索的功能设计	167
7.5.1 功能组织结构	167
7.5.2 短信监控功能设计	167
7.5.3 短信监控的数据库设计	168
7.5.4 分发平台功能设计	171
7.5.5 处理模块功能设计	172
7.6 小结	173
第 8 章 并行技术在移动搜索中的应用	174
8.1 并行计算	174
8.1.1 并行计算的概念	174
8.1.2 并行计算的发展	176
8.2 MPI 简述	177
8.2.1 MPI 的产生	178
8.2.2 MPI 的语言绑定	178
8.2.3 目前主要的 MPI 实现	179
8.2.4 MPI 程序结构	180
8.3 并行计算技术在移动搜索引擎中的应用	183
8.4 并行计算技术在索引/排序部分的应用	184
8.5 并行计算技术在移动搜索部分的应用	186

8.5.1 高性能并行 Crawler 的设计和实现	187
8.5.2 并行 Searcher 部分的设计和实现	191
8.6 基于并行技术的移动搜索引擎的发展趋势	192
8.7 小结	192
第 9 章 3G 时代的移动搜索引擎展望	193
9.1 3G 移动搜索引擎概述	193
9.1.1 3G 时代	194
9.1.2 3G 技术标准	195
9.1.3 3G 移动搜索引擎趋势	197
9.2 Wi-Fi 技术和 WiMax 技术	199
9.2.1 Wi-Fi 技术	199
9.2.2 WiMax 技术	202
9.3 WAP-PUSH 技术	205
9.3.1 WAP-PUSH 的含义	206
9.3.2 WAP-PUSH 的框架	206
9.3.3 WAP-PUSH 的协议和推送方式	207
9.3.4 WAP-PUSH 的实现	208
9.4 音频搜索技术	211
9.4.1 文本搜索	211
9.4.2 语音分析	211
9.5 图片搜索技术	212
9.5.1 图像匹配技术	212
9.5.2 人脸识别技术	213
9.6 视频搜索技术	215
9.6.1 文本搜索	215
9.6.2 语音分析	215
9.6.3 视觉特征分析	216
9.7 3G 技术条件下的其他搜索技术	218
9.7.1 即时信息搜索	218
9.7.2 手机购物搜索	219
9.7.3 商务搜索	219
9.8 小结	220
第 10 章 数据挖掘技术在移动搜索引擎中的应用	221
10.1 数据挖掘简介	221
10.2 聚类在移动搜索引擎中的应用	223
10.2.1 聚类简介	223
10.2.2 数据结构和数据类型	224
10.2.3 聚类分析方法	228

10.2.4 聚类在移动搜索引擎中的应用	230
10.3 关联规则在移动搜索引擎中的应用	232
10.3.1 关联规则相关概念	232
10.3.2 关联规则度量	234
10.3.3 关联规则分类	234
10.3.4 Apriori 算法	234
10.3.5 频繁项集产生强关联规则	239
10.3.6 关联规则在移动搜索引擎中的应用	239
10.4 Web 数据挖掘	241
10.4.1 Web 结构挖掘	242
10.4.2 Web 内容挖掘	244
10.4.3 Web 日志挖掘	245
10.5 小结	247
参考文献	249

第 1 章

移动搜索引擎概述

1.1 引言

移动搜索是指用户在移动通信网络中,通过移动终端,利用 SMS、WAP、IVR 等多种特定搜索方式获取所需信息的搜索行为。移动搜索服务的核心是将搜索引擎与移动设备有机结合,生成符合移动产品和用户特点的搜索结果,从而脱离对固定设备和固定通信网的依赖,实现随时随地的信息获取。手机移动搜索作为搜索技术与移动通信技术的一种结合体,融合了两种技术的特点。移动搜索是网络搜索引擎在移动终端的延伸,是未来搜索引擎发展的一个重要趋势。尽管面临着诸多问题,但移动搜索的崛起已成为大势所趋。伴随着 3G 商用的来临,移动搜索必将逐步迈进高速发展的快车道。

1. 从有线到无线

1994 年大家熟知的 Lycos 成为第一个现代意义的搜索引擎,1998 年 Google 正式推出,成为全球最受欢迎的搜索引擎。我国的搜索引擎从北大天网开始逐步发展,到 2000 年出现“百度”搜索引擎,有线互联网搜索引擎已形成从中文到英文信息的全覆盖。

在有线互联网快速发展的同时,无线通信技术也以更快的速度在发展,手机已成为人们必不可少的沟通工具,搜索引擎的战火已从互联网烧到了无线通信网。2004 年我国开始出现移动搜索引擎,随后 Google、“百度”相继推出了自己的移动搜索引擎。由于人们生活节奏的加快,传统的信息搜索方式已无法满足快节奏的生活和工作方式,无线搜索引擎必将会有广阔的市场空间。

2008 年中国无线互联网搜索引擎用户数达到 11700 万并继续保持高速增长, 到 2010 年无线搜索用户规模将增长到 2 亿户。巨大的市场推动了移动搜索引擎的发展, 可以预见未来人们的搜索行为将主要来自移动搜索, 移动搜索会成为网络搜索强大的竞争对手。

2. 从自然语言到受限语言模型

以 Google 和 Baidu 为代表的传统关键词搜索, 满足了人们从海量信息中搜索所需信息的需要, 搜索已成为了人们生活与工作习惯, 搜索技术特别是智能搜索技术将成为最有应用潜力的关键技术。

随着手机的普及, 手机用户的数量十分巨大, 我国移动用户数已达到 6 亿户。手机作为一种个人移动信息设备具有普通计算机网络无法取代的优点, 基于手机应用的电信增值业务已逐渐形成一个巨大的产业, 但由于基于传统的关键词搜索的手机搜索技术不够人性化, 手机屏幕和内容有限, 无法满足人们在手机上随时随地搜索的需求, 阻碍了整个手机增值业务产业的发展。

移动搜索引擎的查询要以用户满意为检验技术的最终标准, 人们提出了适合手机应用的自然语言理解技术, 解决了系统对手机自然语言的理解问题, 避免了采用传统自然语言理解技术大而全, 但理解率低的问题。利用人类自然语言通过短消息或 WAP 信息提交实现对信息的检索, 真正实现了“口语化提问, 精确解答”, 它的应用和普及将改变人们传统的信息获取方式, 使手机不但作为通信工具, 而且成为个人的信息导航工具。手机的轻便、灵活是普通计算机网络所无法取代的, 它改变了人们传统信息获取习惯, 适应了当今社会快速化、移动化的生活节奏, 对人们的日常生活、工作产生极大的影响。

在自然语言理解模型的基础上, 通过对客户历史数据的挖掘, 实现了对不同用户的个性化服务, 系统根据用户的使用习惯自动实现关键词的进一步限制, 防止检索到大量的垃圾信息, 提高了用户的使用感受, 使信息更加精准, 实现了“以用户为导向的智能搜索技术”。智能搜索技术大大降低了用户采用手机进行信息检索的难度, 符合人们日常的语言习惯, 返回用户所需的精确结果。

基于自然语言理解的下一代手机智能搜索技术的研发和应用, 将作为 3G 关键应用服务, 推动 3G 的应用和普及。

3. 移动搜索风起云涌

自从移动搜索技术的概念被提出, 大量企业进入到这一新兴的领域, 如 Google、百度、Cgogo、UUCUN、手索科技等企业, 成了这一行业的领跑者, 整个移动搜索

行业呈现出一片风起云涌的态势。行业的繁荣既表明了竞争的激烈,但也预示着美好的前景,各企业均坦然地面对竞争,努力寻找搜索技术的赢利模式。

由于移动搜索受众广泛,因此在与一定的商业模式结合后,必然具有灵活的赢利模式,如手机用户对天气、列车、航班、餐饮娱乐等日常信息的检索,并可采用该技术实现对无线手机商场、企业黄页等的搜索。智能搜索技术的实现不但在技术上具有创新性,而且该技术有着强大的产业前景,如手机检索结果的竞价排名,分词关键词竞价,企业信息服务付费,无线手机商场,定向客户营销等。通过各手机搜索企业的共同努力,越来越多的企业赢利模式将被挖掘出来,为整个移动搜索产业提供了强大的发展动力。

1.2 移动搜索引擎的特点

传统的互联网搜索引擎已成为亿万用户日常不可缺少的工具,具有很强的用户基础,移动搜索引擎在这种情况下能迅速引起人们的重视在于它具有的独特特点。移动搜索引擎是基于先进的移动通信技术在移动信息终端上实现的搜索引擎系统。移动搜索将移动通信和搜索引擎有机地结合起来,使得手机用户能够在无法使用计算机和互联网络时利用手机等移动终端进行移动搜索,移动搜索的特点主要有以下几点:

(1) 搜索的便利性

与互联网搜索相比,移动搜索的自由度更大,实现了随时、随地、随身搜索。现代人们的生活节奏越来越快,很多时候都处在运动中,不具有互联网上网的条件,只有移动搜索技术才能使用户不受时间和地点的限制,实现随身搜索,从而获得自己需要的信息。

(2) 搜索的精准性

通常用户使用移动搜索的目的性很强,移动搜索系统需要提供给用户的是相对精确的信息,从而适应手机终端屏幕小、网络接入速度慢等特点,移动搜索技术将更注重使用的简约化和查询实效性,将具备更强的自然语言分析应答能力,并提供更为精准的垂直搜索结果。例如,以短信问答为交互方式的查询,简单直观。这一特点也为研究和设计移动搜索平台提出了更高的要求,大量的智能化技术将在系统中被采用。

(3) 能提供个性化服务

移动设备通常是个人的随身用品,个性化特征非常突出,移动搜索引擎系统一般会通过数据挖掘技术对用户的搜索记录、搜索习惯等个人偏好进行智能化分析,为用户提供最符合个人需求的搜索功能。通过与定位技术的结合,移动搜索服务商可以提供更有针对性的产品。如定向信息投放,只向用户提供其感兴趣的信息,避免了垃圾

信息的骚扰。

(4) 用户终端数量巨大

移动搜索的最大吸引力就是巨大的用户数量,移动终端的数量大大超过了互联网用户终端,据预测,全球的手机用户数已达 20 亿,到 2010 年,全球手机用户数将突破 30 亿。全球电脑用户只有手机用户的十分之一,而其中还有相当一部分不能上网。巨大的用户终端带来了巨大的市场空间,提供了丰富的产业应用模式。Google 仅仅依靠互联网搜索就创造了惊人的奇迹,移动搜索必然会创造更大的奇迹。

1.3 移动搜索引擎的分类

目前市场上的移动搜索方式大致分为三种:基于浏览器的移动搜索、基于短信的移动搜索、短信和微浏览器相结合的移动搜索。

基于浏览器的移动搜索:采用浏览器进行搜索是人们在互联网搜索时代非常熟悉的搜索方式,现代的普通手机都内置了类似于网页浏览器的微浏览器,手机用户可以通过微浏览器来连接互联网。现代主流的移动搜索服务就是基于浏览器的搜索,但有部分手机用户未使用这一功能,这影响了它的普及。

基于短信的移动搜索:移动搜索系统通过短信接受用户的查询请求,并将查询结果通过短信的方式返回给用户。这种应用方式具有很强大的用户基础,大量用户对短信的收发都非常熟悉,适合所有手机用户使用。但短信的信息表现能力很差,能为用户提供的信息相当有限,限制了这一搜索模式的进一步发展。

基于短信和微浏览器相结合的移动搜索:用户可以使用移动搜索服务商提供的客户端,当用户通过客户端向移动搜索系统提交查询请求时,客户端会根据用户的检索行为,自动选择使用通过短信或者微浏览器的方式返回查询结果。

在早期,移动搜索服务商更喜欢选择基于短信的移动搜索,主要是因为移动网络的短信具有统一的开放标准和规范。但基于短信的移动搜索只能向手机用户发送最长为 70 个汉字或 160 个纯英文字母,这大大地限制了基于短信的移动搜索的结果呈现。随着带宽的增加,基于浏览器的移动搜索必然会成为主流的搜索方法,短信搜索可能只在一些特殊的领域发挥作用。

1.4 移动搜索引擎市场现状

经过几年的发展以及 Google、百度等搜索巨头的推波助澜,移动搜索已在用户心中有了初步的概念,市场的井喷在技术发展的推动下一触即发。大量的资金,大量

的公司进入这一领域,准备分享这一块市场的大蛋糕。北京、上海、广州、深圳、杭州和成都等主要城市都有从事移动搜索服务的公司,如北京易查、上海明复等都是这一行业的领跑者。国内每年都会举行多次移动搜索技术与市场的产业研讨会,吸引了大量的业内人士参加。移动搜索产业是一个技术和资金密集的产业,许多企业选择了和移动、联通等运营商合作,利用其强大的资金和市场平台进行移动搜索技术的推广,实现了“和大象共舞”的合作战略。

不过我们也应该看到,移动搜索市场的发展还有许多未定的因数:

1) 用户还没有全面养成使用移动搜索的习惯,如何培养用户采用移动终端进行搜索并形成习惯是这一个产业需要面对的问题。这一培养过程可能涉及大量资金的投入。

2) Google、百度具有成熟的搜索技术,强大的市场基础,并且都已进入移动搜索领域,如何在他们强大的身影下获得生存的空间对中小搜索企业来说是至关重要的,产品的差异化和垂直化是生存的基本规律。

3) 移动搜索企业不同于互联网企业,它受到移动运营商强大的制约,运营商具有网络、资金、市场的优势,如何处理与运营商的关系对移动搜索企业特别是发展初期的企业是一个巨大的考验。

1.5 移动搜索引擎的关键技术简介

移动搜索技术是搜索引擎技术向无线网络的拓展,从关键技术上来看,搜索引擎中的关键技术同样也是移动搜索引擎的关键技术。搜索引擎系统是一个复杂的系统,涉及的技术点非常多,计算机、网络、通信、安全等学科的技术都在搜索引擎中得到了应用,而移动搜索引擎由于使用方法的的不同在某些技术上又有特殊的要求,总的来看移动搜索引擎的关键技术主要有以下几个:

(1) 海量数据的采集与管理

搜索引擎是一种信息密集型的系统,信息的数量是系统的灵魂,搜索引擎系统需要从网络大量收集信息,并向用户提供服务。海量信息的采集和管理一直是计算机科学研究的一个重要方向,当数据量十分大时,数据的采集和管理会成为制约系统性能的一个重要因素。构建性能良好的移动搜索系统,必需要有性能良好的海量数据采集和管理方法。这里主要涉及以下几个方面:① 高速网络信息采集;② 高效的数据索引技术;③ 海量数据的高速存储与访问。

(2) 海量数据的分析处理

对于移动搜索系统,数据分析主要包括两个方面:① 信息数据分析;② 用户