

前 言

进入 21 世纪,互联网技术的发展及其所催生的“新经济”在世界经济格局中所占的分量呈现大幅上升的趋势。互联网的发展和应用趋势也为越来越多的人所关注。近几年来搜索引擎与信息获取技术随着万维网(WWW)这种方便易用的媒体的普及而得到了迅速的发展。由于互联网技术开放性的特点,使得网上的信息越来越丰富,这一方面为用户获取信息提供了很大的便利,另一方面由于信息量的飞速增长,使得用户查找所需信息时犹如大海捞针。新的信息获取技术——搜索引擎技术应运而生,并得到了飞速发展。搜索引擎已经成为在互联网上仅次于电子邮件的第二大网络应用。

信息获取技术从出现到现在已经有 20 多年的历史。在 WWW 出现之前,信息获取技术的发展比较缓慢,主要应用在图书馆、科技信息研究等专业部门,涉及的用户相对较少。在 WWW 出现之后,基于 WWW 的信息获取技术——搜索引擎技术出现并得到迅速发展。由于互联网

的开放性,使得搜索引擎可以面向普通用户,用户的需求推动其进一步发展。我国在信息获取领域起步较晚,在中文信息获取需求的驱动下,出现一些中文搜索引擎网站。但是从国内搜索引擎的应用效果和对搜索引擎技术的掌握上与发达国家比较,则仍然存在较大的差距,特别是在智能搜索引擎的开发、建设和应用水平方面差距更大。这种差距主要表现在两个方面:第一是搜索引擎的性能和搜索结果的质量与国外的搜索引擎相比存在很大的差距。这就是为什么国人在选择搜索引擎的时候普遍选择国外著名搜索引擎的缘故。第二是对搜索引擎与信息获取技术的掌握和应用上有待于提高和加强,应用人才急需培养,经验有待积累和总结。前者的改善应依靠于国内网络性能的整体改善和先进信息获取技术的应用;而后的改善则更多地依赖于人们对搜索引擎和信息获取技术的掌握。

我的博士生张卫丰同志从本科学习阶段开始就跟随我进行搜索引擎的研究,后又跟我攻读硕士学位(因成绩优异被批准改为硕博连读),经过近五年的刻苦努力,在国内外发表了大量学术论文,取得了一系列研究成果。本书即是在对搜索引擎的有关问题进行深入研究的基础上,充分吸收现有研究成果,一起编写而成的。

本书涵盖了搜索引擎与信息获取技术的主要内容,力争系统地全面地介绍信息获取的关键技术,并通过实例来说明,使得复杂的概念容易理解。本书主要包含三部分:第一部分(第1章到第9章)介绍信息获取的基本原理与技术;第二部分(第10章到第11章)主要介绍我们在搜索引擎与信息获取领域所做的工作和取得的最新科研成果,它主要是实现用户个性化搜索的相关技术;第三部分(第12章到第13章)面向普通用户的应用需求,分别从Web站点维护者的角度和普通用户使用搜索引擎的角度讨论了如何提高自己的网页在搜索引擎的排名和如何提高查询的搜索

精度。

我们希望本书的出版能够对搜索引擎的设计者、Web 站点的管理员以及广大用户有所裨益,也希望它能成为有关领域学生的学习参考书。

十分感谢清华大学出版社的同志为本书的出版所做的工作。

作 者

2002 年 2 月于南京

目 录

第 1 章 概述	1
1.1 引言	1
1.2 信息获取与数据获取	2
1.3 信息获取技术的发展	3
1.4 信息获取基本概念	4
1.5 信息获取系统的过去、 现在和将来	5
1.6 信息获取的过程	6
1.7 本书的结构	7
1.8 本章小结.....	10
第 2 章 信息获取模型	11
2.1 引言.....	11
2.2 布尔模型.....	11
2.3 向量模型.....	12
2.4 概率论模型.....	14
2.5 神经网络模型.....	16
2.6 基于命题逻辑的模型及其应用	18
2.6.1 基本概念不相交及其 与向量模型的关系.....	19

2.6.2	基本概念相交及其与布尔模型的关系·····	21
2.7	本章小结·····	23
第3章	标记语言与文本操作 ·····	24
3.1	引言·····	24
3.2	标记语言·····	24
3.2.1	HTML语言·····	25
3.2.2	XML语言·····	26
3.3	文本预处理·····	29
3.3.1	文本的词法分析·····	30
3.3.2	中文分词技术·····	31
3.3.3	无用词汇的删除·····	32
3.3.4	词干提取技术·····	32
3.3.5	索引词条的选择·····	37
3.3.6	词典·····	37
3.4	文档聚类·····	38
3.5	文本压缩·····	39
3.5.1	基本概念·····	39
3.5.2	统计方法·····	40
3.5.3	字典方法·····	41
3.5.4	倒排文件压缩·····	42
3.5.5	文本压缩方法比较·····	44
3.6	本章小结·····	45
第4章	索引和搜索 ·····	46
4.1	引言·····	46
4.2	倒排文件·····	47
4.2.1	倒排文件的搜索·····	48
4.2.2	倒排文件的构造·····	49
4.3	后缀树与后缀数组·····	50

1.4	布尔查询	52
1.5	顺序查询	53
1.6	结构化查询	54
1.7	对压缩文本的搜索	55
1.8	模式匹配	56
1.8.1	容错匹配	56
1.8.2	正规表达式和扩展模式	56
1.8.3	利用索引进行模式匹配	57
4.9	本章小结	58
第5章	信息获取系统评价	59
5.1	引言	59
5.2	相关性	60
5.3	召回率和精度	61
5.3.1	召回率与精度的计算	61
5.3.2	汇聚技术	62
5.4	复合度量	64
5.5	本章小结	65
第6章	查询处理	66
6.1	引言	66
6.2	基于用户反馈信息的查询扩展	66
6.2.1	向量模型的查询扩展和词条权重重新计算	67
6.2.2	概率论模型中的词条权重重新计算	69
6.3	自动局部分析	71
6.3.1	通过局部聚集进行查询扩展	71
6.3.1.1	关联聚集	72
6.3.1.2	距离聚集	73
6.3.1.3	标量聚集	74
6.3.1.4	搜索表达式的改变	75

6.3.2	通过局部上下文分析进行查询扩展	76
6.4	自动全局分析	78
6.4.1	基于相似词典的查询扩展	78
6.4.2	基于统计词典的查询扩展	80
6.5	本章小结	82
第7章	目录式检索服务与聚类分析	83
7.1	引言	83
7.2	目录检索服务的构成	84
7.2.1	网页采集过程	84
7.2.2	网页分类方法	85
7.3	聚类过程	86
7.3.1	文档关联度的衡量	86
7.3.1.1	相似度	86
7.3.1.2	相异度	87
7.3.2	文档聚类	88
7.3.2.1	基于相似度的分类过程	88
7.3.2.2	基于相异度的分类过程	92
7.4	基于聚类的信息获取	94
7.5	本章小结	94
第8章	基于因特网的搜索引擎	95
8.1	引言	95
8.2	基于因特网的搜索引擎的构成	97
8.3	搜索引擎的主要指标及其分析	98
8.3.1	搜索引擎的精度	99
8.3.2	搜索引擎受欢迎的程度	100
8.3.3	搜索引擎相关性考虑	101
8.4	搜索引擎的数据结构	102
8.4.1	Bigfile 文件系统	103

8.4.2	信息库	103
8.4.3	文本索引	104
8.4.4	词典	104
8.4.5	采样表	104
8.4.6	前向索引	105
8.4.7	后向索引	106
8.5	网页的获取	107
8.6	建立索引的方法和过程	108
8.6.1	搜索引擎建立索引的方法	108
8.6.2	索引的过程	111
8.7	搜索过程	112
8.8	搜索结果排序方法	112
8.9	搜索引擎的发展趋势	116
8.10	本章小结	118
第9章	元搜索引擎	120
9.1	引言	120
9.2	基本构成	120
9.3	元搜索引擎分类	122
9.4	与独立搜索引擎的比较	124
9.5	主要指标及其分析	126
9.6	元搜索引擎面临的问题、对策和发展趋势	129
9.6.1	查询预处理	131
9.6.2	搜索结果集成	132
9.7	元搜索引擎调度策略研究	134
9.7.1	GSE 基本思想	134
9.7.2	遗传算法在元搜索引擎调度中的应用	135
9.7.2.1	编码方法	136
9.7.2.2	适应函数和选择	137

9.7.2.3	初始化种群	139
9.7.2.4	重组	139
9.7.2.5	变异	140
9.7.3	GSE 中的智能调度器	141
9.7.4	实验——自适应过程运行周期的确定	142
9.8	文档选择	143
9.8.1	用户决定法	145
9.8.2	权重分配法	145
9.8.3	基于学习的方法	146
9.8.4	确保取回法	147
9.9	结果归并	150
9.9.1	基本定义	150
9.9.2	元搜索引擎结果集成方法	152
9.9.2.1	几种常用元搜索引擎结果集成 方法及其存在问题	152
9.9.2.2	摘要排序法	153
9.9.2.3	位置排序法	154
9.9.2.4	摘要/位置排序法	155
9.9.3	搜索结果集成技术比较	155
9.9.4	实验分析	157
9.9.5	元搜索引擎搜索结果集成技术展望	158
9.10	元搜索引擎可扩展性	159
9.10.1	XML 与 XSL 语言	160
9.10.2	可扩展元搜索引擎的基本结构	161
9.10.3	元查询映射	163
9.10.4	结果归并	166
9.10.5	搜索引擎接入元搜索引擎的过程	171
9.11	本章小结	172

第 10 章 基于客户端的个性化应用研究	173
10.1 利用代理个性化搜索结果	173
10.1.1 用户兴趣模型	174
10.1.1.1 个性化信息抽取与兴趣生成树	174
10.1.1.2 词干抽取与信息预处理	176
10.1.1.3 用户个人兴趣模型	177
10.1.1.4 共同兴趣模型	178
10.1.2 个性化搜索代理系统 PSA	180
10.1.2.1 用户个人兴趣代理	180
10.1.2.2 共同兴趣代理	181
10.1.2.3 利用兴趣剖像过滤搜索结果	182
10.1.3 工作流程	182
10.1.4 性能分析	183
10.2 数据挖掘技术在 Web 预取中的应用研究	184
10.2.1 简化 WWW 数据模型	185
10.2.2 兴趣关联知识库与用户行为预测	187
10.2.3 数据挖掘技术	190
10.2.4 基于代理的 Web 预取技术	193
10.2.5 实例研究	195
10.3 本章小结	196
第 11 章 基于服务器端的个性化应用研究	198
11.1 引言	198
11.2 带反馈自适应搜索引擎系统	199
11.3 数据采集与反馈信息库的生成	200
11.3.1 数据采集	200
11.3.2 反馈信息库的生成及其算法	202
11.4 反馈响应过程	205
11.5 自适应搜索引擎系统原型设计与实验	207

11.5.1	一个实验性带反馈自适应搜索引擎 ASE	207
11.5.2	实验	208
11.6	本章小结	211
第 12 章	搜索引擎策略——站点角度	212
12.1	引言	212
12.2	提高网站在搜索引擎中的排名位置的方法	213
12.2.1	了解不同的搜索引擎	213
12.2.2	关键词的选择	214
12.2.3	标题	217
12.2.4	Meta 值的使用	217
12.2.5	提升自己网站排名的技巧	220
12.2.5.1	隐藏的表单 input	220
12.2.5.2	不可见关键词堆砌	221
12.3	如何提交自己的网站	221
12.3.1	提交工具	221
12.3.2	如何跟踪	222
12.4	阻止网络检索器索引网页	222
12.4.1	阻止网络检索器的方法	223
12.4.2	文件 Robots.txt 的格式	224
12.4.3	Robots.txt 使用实例分析	225
12.5	本章小结	226
第 13 章	搜索引擎策略——用户角度	227
13.1	引言	227
13.2	数学命令在搜索中应用	229
13.2.1	查询条件具体化	229
13.2.2	使用加号 +	229
13.2.3	使用减号 -	230
13.2.4	使用引号 " "	230

13.2.5 组合符号.....	231
13.3 增强的搜索命令.....	232
13.3.1 搜索标题.....	232
13.3.2 搜索网站.....	233
13.3.3 百搭命令(*).....	233
13.4 搜索引擎的辅助功能.....	234
13.4.1 相关搜索.....	234
13.4.2 搜索结果重组.....	237
13.4.3 相近搜索.....	239
13.4.4 延伸搜索条件.....	241
13.5 搜索引擎功能特点分析图表.....	242
13.6 本章小结.....	245
附录 1 搜索引擎导航	246
附录 2 术语	250
参考文献	252

图 目 录

图 2-1	包含两个文档与三个关键词的 简单神经网络	17
图 2-2	布尔信息获取文档表示示例	23
图 4-1	构造倒排索引实例	50
图 4-2	后缀树实例	51
图 4-3	一有层次关系的关键词集合	52
图 5-1	精度-召回率曲线的拟合	64
图 6-1	由全连接算法生成的 3 个类别的 层次结构	81
图 7-1	根据相似度和阈值生成的对象图	89
图 7-2	根据子图来划分的可能类	90
图 7-3	层次树状图	92
图 7-4	单连接方法根据相异度生成图 进行分类的过程	93
图 8-1	搜索引擎受欢迎程度比较	100
图 8-2	信息库数据结构	103
图 8-3	采样的结构(两个字节)	105
图 8-4	前向索引数据结构	106
图 8-5	后向索引数据结构	107

图 9-1	元搜索引擎原理图	121
图 9-2	GSE 中的智能调度器	141
图 9-3	几种排序方法偏移度比较	157
图 9-4	可扩展元搜索引擎 SMetaSearch 框架	162
图 9-5	元查询映射	164
图 9-6	扩展元查询实例	164
图 9-7	扩展元本地查询 DTD 和用其格式化后的 扩展本地查询 XLQ	165
图 9-8	扩展本地查询向本地查询映射转换程序	166
图 9-9	搜索结果及其模板	168
图 9-10	搜索结果 DTD 及其用其格式化后的搜索结果	169
图 9-11	将搜索结果转换为 HTML 形式的 XSL 程序	170
图 10-1	兴趣生成树	176
图 10-2	个性化搜索代理系统 PSA	180
图 10-3	简化 WWW 数据模型	186
图 10-4	利用知识库预测用户链接次序	188
图 10-5	基于代理的 Web 预取系统	194
图 11-1	带反馈自适应搜索引擎系统原理图	200
图 11-2	反馈信号响应过程	205
图 11-3	自适应搜索引擎 ASE	207
图 11-4	相异度曲线及趋势	210
图 11-5	相异度总体曲线与总体趋势	211
图 13-1	AltaVista 的相关搜索功能	234
图 13-2	AOL Search 中的相关搜索链接	235
图 13-3	Excite 列出的相关关键词	235
图 13-4	Excite 提供的相关词搜索	236
图 13-5	Go 提供的相关搜索	236
图 13-6	HotBot 的相关结果搜索	237

图 13-7	Yahoo 的相关搜索词	237
图 13-8	AltaVista 的搜索结果重组功能	238
图 13-9	Excite 将网页按网站组合的功能	238
图 13-10	Go 中的关闭重组功能	239
图 13-11	HotBot 中查看网站中其他网页的功能	239
图 13-12	Northern Light 的搜索结果重组功能	240
图 13-13	AOL Search 的相近搜索功能	240
图 13-14	Excite 的相近搜索功能	240
图 13-15	Google 的相近搜索功能	241
图 13-16	HotBot 的延伸搜索功能	241
图 13-17	MSN Search 的延伸搜索功能	242
图 13-18	Snap 的延伸搜索功能	242

表格目录

表 3-1	文本压缩方法比较	44
表 5-1	相关文档的集合定义	61
表 5-2	(a) 查询 1 的召回率与精度	63
	(b) 查询 2 的召回率与精度	63
	(c) 对查询 1 与查询 2 汇聚后 得到的召回率与精度	63
表 8-1	搜索引擎结果中的质量	100
表 8-2	搜索引擎索引方法比较	110
表 9-1	元搜索引擎比较	128
表 9-2	GSE 中遗传算法执行频率与 智能代理性能的变化	143
表 10-1	取不同参数 n 对用户满意 程度的影响	184
表 10-2	实验结果分析	196
表 11-1	反馈信息库数据结构	208
表 13-1	如何选择搜索引擎	228
表 13-2	搜索引擎的数学命令	242
表 13-3	增强的搜索命令	243
表 13-4	辅助搜索功能	244
表 13-5	结果显示功能	245

概 述

1.1 引 言

随着计算机技术和互联网技术的飞速发展,信息获取已经从手工获取,到计算机信息获取,以及到现在的通过网络进行信息获取。网络的最大优点就是将大量的信息相互共享,而且只要通过一台接入互联网的计算机就可以方便地获取信息。利用互联网,用户一方面可以快速、方便地接触到各种信息,但是另一方面通过普通浏览的方式很难在信息的海洋中找到真正需要的信息。网络时代的信息量每 8 个月就翻一倍,如今的网页以 10 亿来计算,要在浩如烟海的网络世界寻找需要的信息,作为现代信息获取技术的主要应用——搜索引擎是必不可少的。搜索引擎正在不断地改变人们获取信息的方式。利用搜索引擎可以快速找到需要的信息。信息获取技术现在广泛应用于搜索引擎、数字图书馆等。

搜索引擎是仅次于门户的互联网第二大核心技术,伴随互联网的普及和网上信息的爆炸式增长,它越来越引起人们的重视。

搜索引擎技术的市场不仅限于门户网站,专业网站同样需要快速有效的搜索。此外,各个企业、机构自己的网站也是一个极其广阔的市场领域。

目前,国内不少企业花了很多钱构建了一个内容丰富的网站,