网络金融系列丛书

Foundations of Internet Financial Information Intelligent Mining

互联网金融信息智能挖掘基础

梁 循 著

北京大学出版社
PEKING UNIVERSITY PRESS

网络金融系列丛书

Foundations of Internet Financial Information Intelligent Mining

互联网金融信息智能挖掘基础

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 提 要

互联网金融信息智能挖掘是一个涉及数据挖掘、计算智能、统计学、计算语言学、模式识别、金融学等多个学科的领域。

本书综合了大量国内外的最新资料和作者的研究成果，系统而有选择地介绍了互联网金融信息智能挖掘问题。全书从结构上分为三篇。第1篇介绍了作者主持研发的一个互联网金融信息挖掘系统平台。第2篇具体介绍了一些相关技术基础，包括互联网金融信息文本分析、神经网络技术、支持向量机技术。第3篇主要介绍了互联网金融信息挖掘领域的一些问题和基本应用，包括互联网金融信息的相关分析，金融信息量、交易量和收益率时间序列的关联研究，以及基于金融信息量的交易量和收益率的控制问题。

本书的读者可以是对模式识别、计算机智能感兴趣的计算机专业人士，也可以是对互联网金融信息智能挖掘感兴趣的领域专家。它可供数据挖掘、机器智能、数据分析、金融等领域的科技人员和高校师生作研究的参考资料。

图书在版编目(CIP)数据

互联网金融信息智能挖掘基础/梁循著. —北京：北京大学出版社，2009.7
(网络金融系列丛书)
ISBN 978-7-301-15534-9

Ⅰ. 互… Ⅱ. 梁… Ⅲ. 互联网络—应用—金融—信息处理 Ⅳ. F830.49

中国版本图书馆CIP数据核字(2009)第121173号

书　　名：互联网金融信息智能挖掘基础
著作责任者：梁　循　著
责 任 编 辑：沈承凤
封 面 设 计：张　虹
标 准 书 号：ISBN 978-7-301-15534-9/TP·1040
出 版 发 行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://www.pup.cn
电 子 信 箱：zpup@pup.pku.edu.cn
电　　话：邮购部62752015　市场营销中心62750672　编辑部62752038　出版部62754962
印 刷 者：北京大学印刷厂
经 销 者：新华书店
787毫米×1092毫米　16开本　12.5印张　309千字
2009年7月第1版　2009年7月第1次印刷
定　　价：25.00元

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究
举报电话：010-62752024　电子信箱：fd@pup.pku.edu.cn

科技部863资助项目(2007AA01Z437)
国家自然科学基金资助项目(70571003、70871001)
北京大学(2007)研究生课程立项建设资助项目

前　言

本书介绍了互联网金融信息挖掘的基础理论和算法。本书与作者先前出版的另外6本书籍《网络金融》、《数据挖掘算法与应用》、《互联网金融信息系统的设计与实现》、《电子商务理论与实践 —— SCM、ERP、CRM、DW、VSE、B2C、B2B、B2M、M2M和C2C举例》、《网络金融信息挖掘导论》和《网络金融系统设计与实现案例集》之间的关系见图0-1。

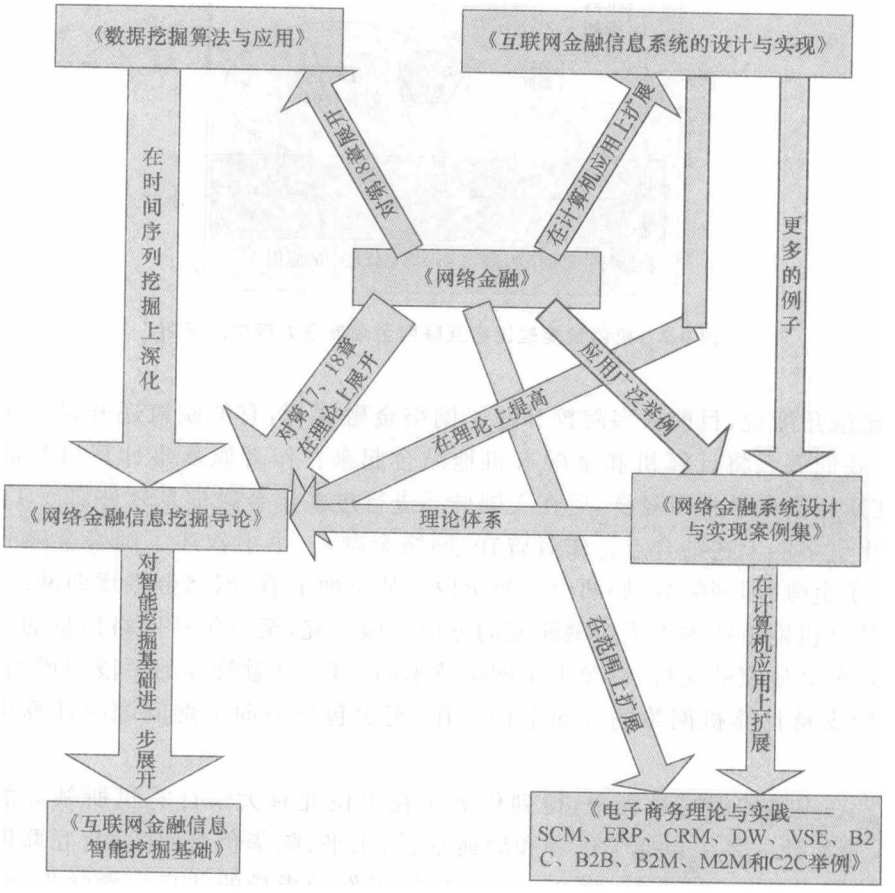

图0-1　《网络金融》、《数据挖掘算法与应用》、《互联网金融信息系统的设计与实现》、《电子商务理论与实践——SCM、ERP、CRM、DW、VSE、B2C、B2B、B2M、M2M和C2C举例》、《网络金融信息挖掘导论》、《网络金融系统设计与实现案例集》和《互联网金融信息智能挖掘基础》之间的关系

试读结束：需要全本请在线购买：　www.ertongbook.com

本书分为三篇。第1篇介绍了作者主持研发的一个互联网金融信息挖掘系统平台。第1篇为网络信息挖掘基础理论和算法，相当于图0-2中的机器智能理论与技术。本篇首先介绍一些相关的理论和技术基础知识，包括金融信息文本处理基础、神经网络技术和支持向量机技术，其中神经网络技术和支持向量机技术大多由作者发表的论文改编扩充而成。第2篇为基础应用部分，相当于图0-2中的机器智能技术与应用，将从金融领域对网络信息挖掘展开实验研究，包括：互联网金融信息的相关分析，金融信息量、交易量与收益率时间序列的关联研究，以及基于金融信息量的交易量与收益率的控制问题。

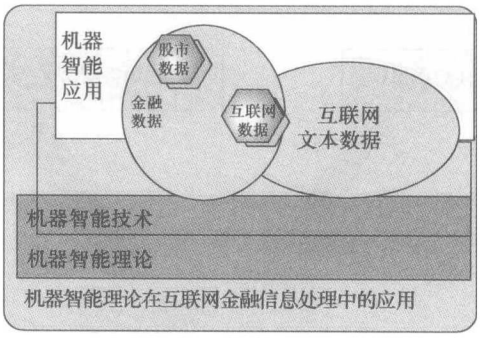

图0-2 机器智能理论在互联网金融信息处理中的应用

从金融角度说，目前很多院校开辟了网络金融课程，有的院校还开设了金融信息系，那么我们怎么将计算机和金融有机地结合起来？作者原从事计算机智能研究工作，曾在斯坦福大学学习经济，后在美国硅谷进行过多年金融信息软件研发工作，一直试图找出更多的上述结合点。作者曾在《网络金融》一书中叙述了网络金融的金融电子化、电子金融（即网络金融）两个主要阶段。从字面上看，网络金融意即网络加金融。把它从计算机网络技术支援金融研究的方向上做研究，至少包括网络信息的计算机自动分析对金融研究的支持，以及由于网络技术的出现，金融管理的手段的增加；把它从金融科学支持计算机网络的方向上做工作，至少包括面向金融问题的计算机软件平台。

作者在本书的编写过程中，得到作者所在单位北京大学计算机所领导的大力支持，第1章所述的系统包括了作者和唐典章、闫永平、陈华负责研发的《互联网金融信息智能挖掘系统》一些界面，第6、7、8章包括了作者指导的北京大学学生李欣利、李楠、赵伟、陈赢、阮进、王超、汤洋的实验结果。

本书也是作者负责的科技部863计划项目（2007AA01Z437）、国家自然科学基金资助项目（70571003、70871001）、北京大学（2007）研究生课程立项建设资助项目及留学回国启动基金（4131522）资助的成果之一。

本书中提到的沪深股市互联网金融信息数据，可以通过172.31.45.151的web-

mine、webocean 数据库免费获取(需要事先索取用户名和口令[①])。

本书和上面的 6 本书一起,组成北京大学计算机科学技术系研究生选修课教学和辅导材料。课程网页上提供另外一些相关材料,网页的地址是 http://www.icst.pku.edu.cn/course/efinance/2008/index.htm。几年来,尽管作者一直致力于从各个角度使整个"网络金融"体系完善,编写了如上的书籍,但是,作者仍常常感到"力不从心"。反观之,国内的很多"网络金融"同行们的著作和研究结果时常让我倍感鼓舞。

由于作者水平和时间的限制,书中一定存在不少缺点和错误,书中的错误由作者负责,恳请读者批评指正。作者也准备将更正及时发表在网上,作为本书的一个补充。

梁　循

2009 年 4 月于北京大学燕北园

① 由于作者精力有限,对数据的索取,只接待教师的 E-mail 或来函的要求,需要数据的同学请通过你们导师索取,请见谅。

目　录

第 1 篇　系统平台

第 1 章　互联网金融信息及其挖掘系统 …… (3)

1.1　互联网金融信息概述 …… (3)

1.2　互联网金融信息挖掘系统平台的总体结构 …… (4)

1.3　互联网信息的计算机获取 …… (5)

1.4　互联网金融信息挖掘结果展示系统 …… (8)

1.5　展望 …… (16)

第 2 篇　技术基础

第 2 章　互联网金融文本处理技术 …… (21)

2.1　概述 …… (21)

2.2　基础资源 …… (24)

2.3　词法分析 …… (32)

2.4　句法分析 …… (37)

2.5　语义分析 …… (40)

第 3 章　神经网络方法 …… (46)

3.1　学习的分类 …… (46)

3.2　群和正交群 …… (48)

3.3　前馈神经网络误差超曲面的复杂性 …… (56)

3.4　最小二乘拟和与广义逆矩阵 …… (65)

3.5　结构压缩的通用算法 …… (68)

第 4 章　支持向量机技术(Ⅰ) …… (80)

4.1　数学准备：线性空间和线性算子 …… (81)

4.2　SVC 和 SVR …… (89)

4.3　将多项式核分解为到单项式空间的映射 Φ …… (99)

第 5 章　支持向量机技术(Ⅱ) …… (108)

5.1　支持向量机的结构压缩 …… (108)

5.2　支持向量机的增量学习算法 …… (122)

5.3 支持向量机超曲面不均分两类 …… (122)
第 6 章 支持向量机技术(Ⅲ) …… (141)
6.1 H 和 U 及 K 空间的一些关系 …… (141)
6.2 通过在 U 中训练第 2 个 SVM 调整分隔超平面 Ω …… (146)

第 3 篇 基础应用

第 7 章 金融信息量和交易量及收益率时间序列的关联 …… (155)
7.1 概述 …… (155)
7.2 基于神经网络的金融信息量建模 …… (156)
7.3 基于支持向量机的金融信息量建模 …… (158)
7.4 NN 和 SVM 在挖掘新闻量和交易量关系的比较研究 …… (162)
第 8 章 基于金融信息量的股市收益率的控制问题 …… (163)
8.1 概述 …… (163)
8.2 控制系统分析 …… (164)
8.3 使用金融信息量控制收益率波动率的实验 …… (166)
8.4 展望 …… (173)

参考文献 …… (174)

第 1 篇

系统平台

试读结束：需要全本请在线购买： www.ertongbook.com

第 1 章　互联网金融信息及其挖掘系统

随着网络技术的迅猛发展，互联网在从用户规模、业务应用和技术实现等方面都不断发生着巨大的变化，并且已演化成一个虚拟社会。由于网络本身的虚拟性、隐蔽性、发散性、渗透性和随意性等特点，越来越多的人愿意通过互联网表达自己真实的想法。因此，互联网已经日渐成为信息发布和传播的主要场所之一。如何进行互联网信息收集、整理和分析，是目前的互联网问题研究热点问题。显见，面对海量的互联网信息，使用以往传统的人工方式对互联网信息的监测难以实施，智能互联网信息监控和挖掘系统平台就是在这个需求下产生的一个计算机工具。本书主要讨论互联网金融信息的相关问题。在本章中，我们首先对互联网金融信息的概念做一个综述，然后设计一个互联网金融信息智能挖掘系统平台。

1.1　互联网金融信息概述

1.1.1　互联网金融信息及其主要特点

每天，互联网上各种不同版面都有成千上万的数据信息发布。网络成为一种新兴传播载体之后，已经变成民众表达思想和情绪的重要窗口（刘毅，2007）。互联网金融信息包含了较多民众对各种金融问题所表达的信念、态度、意见和情绪等。随着互联网在全球范围内的飞速发展，网络媒体已被公认为是继报纸、广播、电视之后的“第四媒体”，网络成为反映互联网金融信息的主要载体之一。网络环境下的金融信息的主要来源有：新闻评论、BBS(bulletin board system)、聊天室、博客、聚合新闻。互联网金融信息表达快捷、信息多元、方式互动、具备传统媒体无法比拟的优势，其特点于网络传播的方式密切相关，它的开放性和虚拟性，决定了网络信息具有以下特点：

1. 自由和随意性

这点是从互联网金融信息的传播来看的。由传播技术的发展史可以知道，一般来说，每次出现一种新的媒体，人们传播新闻和发表言论的自由度都会被扩大。人们可以通过 E-mail 传递信息，可以通过即时通信工具沟通和交流感情，还可以在 BBS 和博客 BLOG 上自由和随意地发表言论和表达看法。人们还可以在网络上建立自己的网站，来发表自己的见解，或者出版自己的著作或报纸，成本低廉，程序简便。

2. 即时互动和突发性

从网络媒体区别于传统媒体的主要传播特性来看，互联网金融信息传播是即时和互动的交流。通过 BBS、新闻点评和博客网站，网民可以立即发表意见，民意表达更加畅通。与传统媒体单向的信息传播通道不同，网络能够双向交互。就互联网金融信息而言，其交互性主要体现在网民之间、网民与网络媒体之间的互动。此外，在网络环境下，互联网金融信息的传播和表达具有较高的时效性，一些大型门户网站更加突出了反映重大事件的原创性言论的即时性。

互动性使得互联网金融信息的形成往往非常迅速，具有突发性的特点。一个热点事件的存在加上一种情绪化的意见，就可能成为点燃一片舆论的导火索，并导致金融市场的波动。

3. 情绪化

互联网金融信息的质量在于理性程度，但是，问题是股民发布的互联网金融信息常常带有很强的倾向化。不过，这正为使用计算机对互联网金融信息进行智能分析提供了有利因素。

1.1.2 互联网金融信息传播的主要途径

互联网金融信息通过网络载体进行传播，目前主要通过以下几种方式：① 电子邮件 E-mail；② 即时通信工具；③ 电子公告板 BBS 及网上社区；④ BLOG。这四类网络言论的传播渠道各有特色，它们为广大网络用户提供了就一些公共事务发表意见和表达互联网金融信息的场所，形成了一种开放的、交互的、并能迅速影响广泛的公共空间。

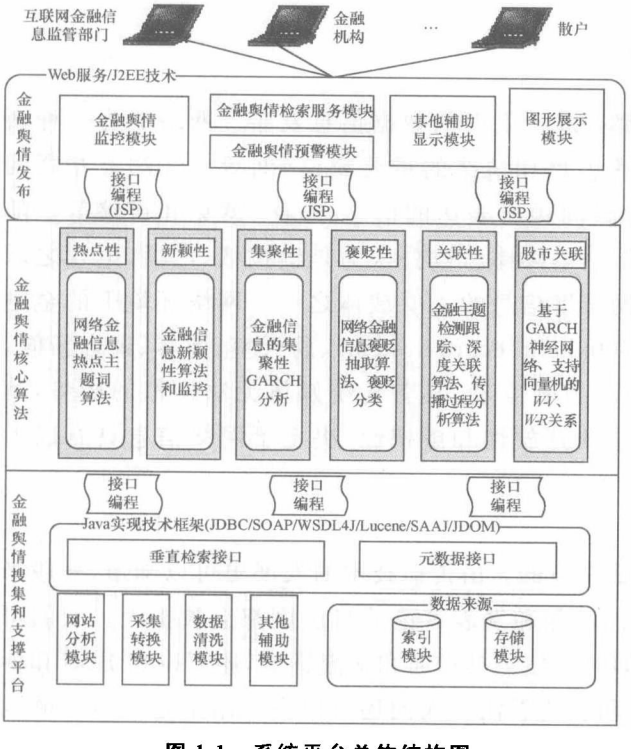

图 1-1 系统平台总体结构图

1.2 互联网金融信息挖掘系统平台的总体结构

我们的互联网金融信息挖掘系统平台的总体结构大致可分为三层，用户界面层，核心算法层和数据库持久化层，如图 1-1 所示。

1.3　互联网信息的计算机获取

互联网信息的计算机获取程序共分为两个部分：基于各种网站的信息抓取系统和基于 BBS 的信息抓取系统。对前者读者可参见作者以前出版的《数据挖掘算法与应用》、《互联网金融信息系统的设计与实现》和《网络金融信息挖掘导论》中的相应部分，下面集中讨论对 BBS 信息获取问题。

获取 BBS 的信息比获取普通信息网站的难度要大。主要难度体现在以下两点，我们下面举例说明。

1. 各个 BBS 板块的公共信息数量的变化

在 BBS 上，各个 BBS 板块的信息数量在不断变化（见图 1-2 和图 1-3）。

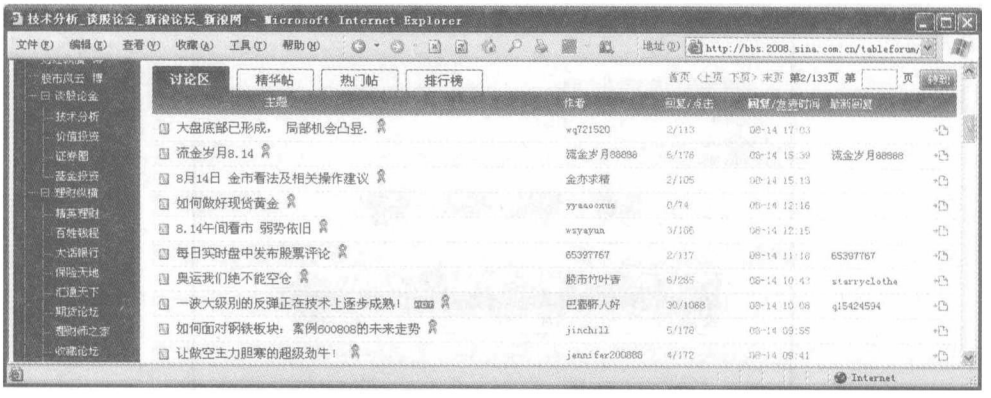

图 1-2　新浪 BBS 财经论坛中的“谈股论经”讨论区下的“技术分析”子讨论区中的公共信息

数据来源：http://bbs.2008.sina.com.cn/tableforum/App/index.php?tree=0&bbsid=62&subid=1&p=2

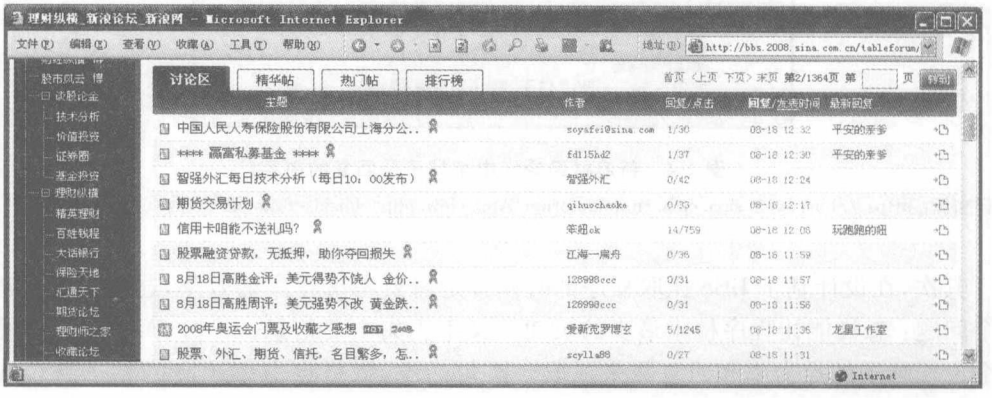

图 1-3　新浪 BBS 财经论坛中的“理财纵横”讨论区下的“精英理财”子讨论区中的公共信息

数据来源：http://bbs.2008.sina.com.cn/tableforum/App/index.php?tree=0&bbsid=63&subid=0&p=2

不难看出，论坛中的公共信息会出现在所有的讨论页面中，但是这些公共信息的数量和内容是在不确定性的变化的，所以，我们在抓取的过程中就要剔除那些公共信息，只抓取非公共信息的部分。

2. BBS 中帖子的实时抓取

随着互联网的飞速发展，网民的数量也急剧增大，而 BBS 正好为大家提供了一个开放的交流平台，所以我们也应当将大家所发表的意见逐一抓取下来。但是网民发表意见的时间和几率具有不规律性，从而也给我们实时抓取带来了以下一些难点。

首先，在对论坛实行实时性抓取时，其中的回复内容的范围跨度不好控制，例如，由于帖子的回帖时间具有不确定性，有可能是白天，也有可能是在凌晨，如何让程序能准确无误、实时地抓取到这一条回帖，是一个挑战（见图 1-4）。

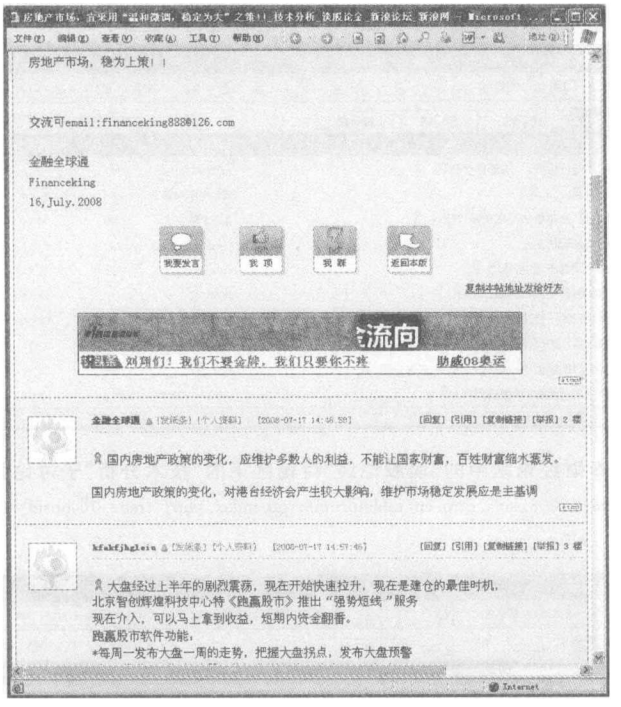

图 1-4 新浪财经论坛中某帖子及回复的内容

数据来源：http://bbs.2008.sina.com.cn/tableforum/App/view.php? bbsid=62&subid=1&fid=450&tbid=375

其次，在设计基于 BBS 抓取程序的时候肯定需要考虑程序的效率，这就不得不思考一个问题：怎样能让程序检索最少的页面但又不丢失主要信息。一种方法是，可以根据每个论坛帖子的点击率和回复数来抓取。换句话说，就是点击率越高的帖子排位越靠前，说明其关注度就越高，回复数就越多。目前，大多数网站按时间排序，不过也有一些网站可以按回复数和点击数分别排序（见图 1-5、图 1-6、图 1-7）。如果某网站可以按

回复数或点击数排序，我们就可以根据帖子在相应讨论区中位置排名来决定是否进入其对应的讨论页面抓取回复信息。

图 1-5　新浪财经论坛中“技术分析”讨论区的帖子排序情况

数据来源：http://bbs.sina.com.cn/192/hits_g10/1.shtml

图 1-6　网易财经论坛中“股海范舟”讨论区的帖子排序情况

数据来源：http://bbs.stock.163.com/list/fanzhou.html

图 1-7 搜狐财经论坛中帖子排序情况

数据来源：http://club.stock.sohu.com/main.php?c=20&b=simplestock&a=566702

3. 如何避免重复抓取

如何实现信息不重复抓取也是一个很有复杂度的问题。以下是其中几个解决途径：

(1) 在对信息每一次抓取之前，对信息的回复时间进行判断。这种方法虽然能实现抓取功能，但是由于太过耗费内存资源，所以造成程序效率也很低下。

(2) 使用 Java 中提供的多线程技术。首先，在程序中构造一个大的线程池，调整每个进程的休眠时间，当一个线程重新启动后就正好可以抓取休眠后所出现的信息。这种方法的缺点就是当遇到某帖子被大量网民关注或者是回帖时，就容易造成回复信息的丢失。

(3) 编写强大的网页模板解析器，能让程序自动识别网页内容，优化程序底层的结构和算法，使程序能具有搜索引擎的实时搜索功能，只有这样才能从根本上解决以上所述问题。目前国内的搜索引擎在信息的实时性方面最为著名的是酷讯网。它能将网络信息实时性缩减至 1 分钟。

1.4 互联网金融信息挖掘结果展示系统

1.4.1 概述

本软件系统是北京大学计算机所网络金融实验室研发的一套互联网金融信息采集与监控的软件。系统平台指整合互联网搜索技术及信息智能处理技术和知识

试读结束：需要全本请在线购买： www.ertongbook.com