网络金融系列丛书

An Introduction to Internet Financial Information Mining

网络金融信息挖掘导论

梁 循 著

北京大学出版社
PEKING UNIVERSITY PRESS

网络金融系列丛书

An Introduction to Internet Financial Information Mining

网络金融信息挖掘导论

梁 循 著

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 提 要

网络金融信息挖掘是一个涉及互联网技术、垂直搜索引擎、信息论、计算智能、自然语言处理、金融学、计量经济学等多个学科的领域。目前，它还是一个很新的交叉领域。

本书初步介绍了作者的一些研究成果。第1章绪论对互联网金融信息挖掘轮廓做了一个勾画。绪论后的部分从结构上分为3篇。第1篇主要介绍网络金融信息的浅层挖掘，讨论了互联网金融信息半结构化文本的挖掘问题和金融信息垂直搜索引擎。第2篇研究网络金融信息流本身的特性，讨论了网络金融信息流的概率分布特性、平稳性和GARCH建模问题及残差的正态性、异方差性和自相关性，并初步探讨了网络金融信息流时间序列的形态挖掘问题。第3篇研究网络金融信息流与交易量和收益率时间序列的关联问题。

本书的读者可以是对互联网垂直金融信息搜索进行专项研究的计算机专业人士，也可以是对金融领域知识挖掘感兴趣的金融专业人士。它可供电子商务、数据挖掘、金融数据分析等领域的科技人员和高校师生参考。

图书在版编目(CIP)数据

网络金融信息挖掘导论/梁循著. —北京：北京大学出版社，2008.1
(网络金融系列丛书)
ISBN 978-7-301-13004-9

Ⅰ. 网… Ⅱ. 梁… Ⅲ. 计算机网络—应用—金融—信息处理 Ⅳ. F830.49

中国版本图书馆CIP数据核字(2007)第172904号

书　　名：网络金融信息挖掘导论
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 978-7-301-13004-9/TP·0920
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：北京宏伟双华印刷有限公司
经　销　者：新华书店
787毫米×1092毫米　16开本　14印张　346千字
2008年1月第1版　2008年1月第1次印刷
定　　价：25.00元

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

前　言

本书涉及网络金融核心的问题。本书和作者先前出版的另外四本书籍：《网络金融》(梁循和曾月卿，北京大学出版社，2005)、《数据挖掘算法与应用》(梁循，北京大学出版社，2006)、《互联网金融信息系统的设计与实现》(梁循、杨健和陈华，北京大学出版社，2006)、《电子商务理论与实践》(梁循和陈华，北京大学出版社，2007)一起，以及作者计划要出版的《网络金融信息分析》、《网络金融案例集》，从某种角度勾画出了网络金融的一个轮廓，7 本书的关系如图 0-1 所示。

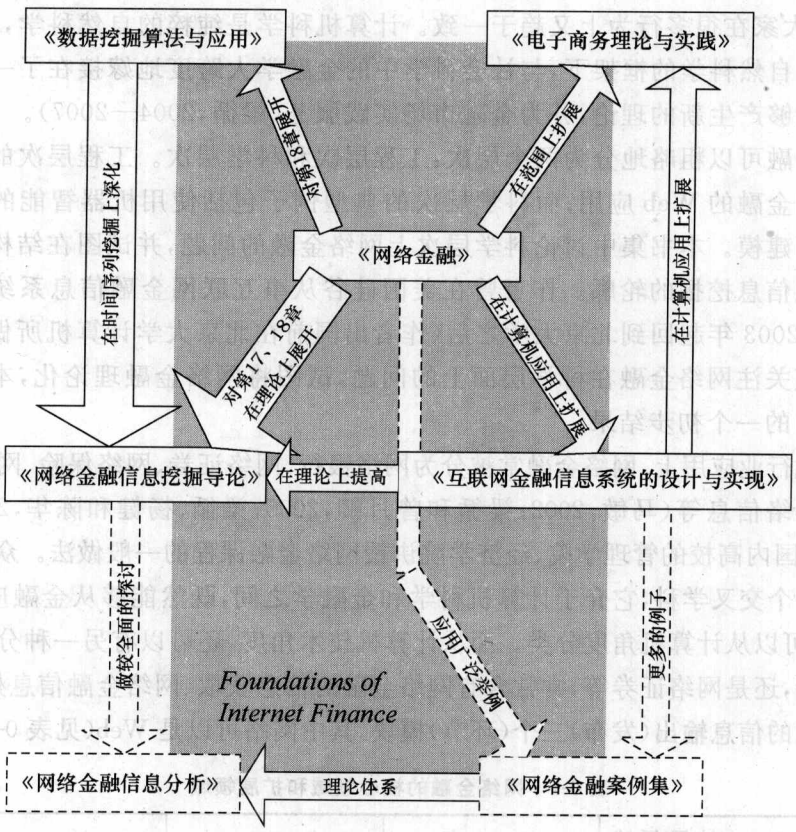

图 0-1　《网络金融》、《数据挖掘算法与应用》、《互联网金融信息系统的设计与实现》、《电子商务理论与实践》、《网络金融信息挖掘导论》、《网络金融信息分析》及《网络金融案例集》之间的关系

图 0-1 中实线框内为已经完成的书籍，虚线框内为计划续写的书籍，其中《网络金融信息分析》将把网络金融信息挖掘问题建立在一个严格的数学(概率空间)框架内，

并进行更多更为细致的实证分析，而《网络金融案例集》将把作者在北京大学计算机系讲授的学生（共5次）期末项目汇编在一起，通过案例对网络金融问题进行更直观的描述，以供同行参考。所有这些努力都是试图把网络金融的轮廓描得更清晰一些，轮廓内的“版画”刻得更细致一些。在此基础上，将7册书整合浓缩成一本 *Foundations of Internet Finance*，计划在国外出版社出版发行。

如果将网络金融看作一个金融学的一个学科分支，那么网络金融信息挖掘就是该学科分支的一个重要组成部分。本书主要介绍作者对网络金融信息挖掘的内容体系的一个初步认识。网络金融信息挖掘是一个典型的多学科交叉领域，涉及的学科有互联网技术、垂直搜索引擎、信息论、计算智能、自然语言处理、金融学、计量经济学、数学等。金融学一般被归为社会科学的范畴，但是，股市的运作规律多多少少还是有一定的“自然”科学的味道。诚然，人的行为在其中所起的作用是巨大的，不过，共同目标的驱使，使得大家在很多行为上又趋于一致。计算机科学是纯粹的自然科学，在本书中，网络金融在自然科学的框架下，与社会科学中的金融学大跨度地嫁接在了一起。希望这种嫁接能够产生新的理论，并为金融市场实践服务（梁循，2004—2007）。

网络金融可以粗略地分为两个层次：工程层次和科学层次。工程层次的典型形式为计算机加金融的 Web 应用；而科学层次的典型例子包括使用机器智能的方法对金融时间序列建模。本书集中讨论科学层次上网络金融的问题，并试图在结构上勾画一个网络金融信息挖掘的轮廓。作者曾在美国硅谷从事互联网金融信息系统的开发工作多年，自2003年起回到北京大学之后（作者出国前在北京大学计算机所做博士后研究），就一直关注网络金融在科学层面上的问题，试图将网络金融理论化，本书就是作者所做努力的一个初步结果。

从金融行业应用上，网络金融常被分为网络银行、网络证券、网络保险、网络结算、网络理财和网络信息等（马敏，2002；梁循和曾月卿，2005；梁循、杨健和陈华，2006）（见表0-1）。这是国内高校的管理学院、经济学院讲授网络金融课程的一般做法。众所周知，网络金融是一个交叉学科，它介于计算机科学和金融学之间，既然能够从金融应用角度分类，当然也可以从计算机角度分类。而从计算机技术角度，还可以有另一种分类，即不论是网络银行，还是网络证券等，常常含有网络金融的信息获取、网络金融信息处理（挖掘）和网络金融的信息输出（发布）三个（环节）模块，其中网络可以是 Web（见表0-2）。

表0-1 网络金融的核心领域和扩展领域

扩展领域 / 核心领域	网络结算	网络理财	网络信息	……
网络银行 网络证券 网络保险 网络彩票 ……	↕	↕	↕	……

表 0-2　网络金融从计算机角度的分类

网络金融的信息获取
网络金融信息处理(挖掘)
网络金融的信息输出(发布)

从理论上说,股价、股市交易量、金融信息量是三个重要的金融变量。一直以来,金融学者们不断地致力于研究信息与股价 P 的关系,不管是从质上还是量上。从单条金融信息质上,学者们已经有很多关于信息与股价关系的成果。从单位时间的金融信息量 W 这个课题上,在互联网产生之前,金融信息主要通过报纸、电视传播,计算机要处理这些金融信息,就需要首先进行文字识别和语音识别,所以,要按指定时段收集整理大量金融信息,再按多种要求分类,如果使用人工的办法,工作量是巨大的。人们观察到,金融信息量和股市交易量 V 有相似的波动规律,于是,一些学者借用股市交易量作为金融信息量的替代变量,通过研究股市交易量与股价之间的关系,来研究金融信息量与股价之间的关系(见图 0-2 实线部分)。然而,在互联网产生之后,金融信息量可以通过互联网获得,不需要进行文字识别和语音识别工作,使得我们直接研究的内容增大了不少(见图 0-2 虚线部分),也成为网络金融信息挖掘的主要内容之一。例如,我们可以直接研究信息量 W 自身的规律,实证过去我们使用的信息量 W 与交易量 V 之间的替代可行性、信息量 W 与股价 P 之间的关系。

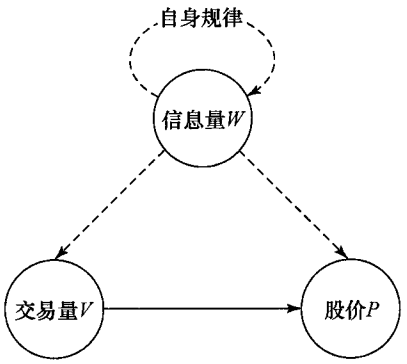

图 0-2　网络金融信息挖掘的主要研究内容之一
——金融信息量特性及其与股市交易量和股价之间的关系

本书从结构上说,第 1 章绪论对互联网金融信息挖掘轮廓做了一个勾画,并对金融信息量、股市交易量和股价三者之间的关系做了初步探讨。绪论后面的部分分为 3 篇。第 1 篇讨论网络金融信息的浅层挖掘问题,由第 2 章和第 3 章组成,主要讨论网络金融信息的搜索和前台浅层挖掘。第 2 章讨论了金融垂直信息搜索引擎,第 3 章包括金融信息去重和分类,金融信息热点排名和金融信息的情感分析。

第 2 篇和第 3 篇讨论网络金融信息的深层挖掘问题，很大程度上使用的是计量经济学的方法。第 2 篇研究网络金融信息流的特性，由第 4 章至第 6 章组成。第 4 章介绍了网络金融信息流时间序列的各种形式、概率分布、平稳性和自相关性。第 5 章讨论了网络金融信息流时间序列的 GARCH 建模，并分析了残差的正态性、异方差性和自相关性。第 6 章对股价时间序列的形态进行了挖掘和分析。

第 3 篇主要研究网络金融信息流和交易量及收益率时间序列的关联问题，由第 7 章至第 9 章组成。在探讨网络金融信息流和交易量时间序列的关联问题上，分成了两章。第 7 章基于 GARCH 模型研究了两者的关联。第 8 章使用计算智能领域中可以完成非线性映射关系的神经网络和支持向量机，对两者关系进行了关联分析。第 9 章初步研究了网络金融信息流和股价时间序列的关联。

网络金融信息量化的直接结果之一就是逆问题 $P\sim W$ 研究。我们知道，中央宏观管理机构使用信息可以调控股市波动。但是，使用多少信息以及多大信息强度，可以（理论上）将股市调整到希望的位置？目前，还没有进一步研究成果。事实上，如果没有信息强度的概念，想从量上调整股市，就无法度量。

如果把 $P\sim W$ 当成反馈控制，可以形成一个 $W\sim P$ 调控系统，该系统是一个闭环的金融宏观调控控制系统（见图 0-3）。我们可以在此基础上，进一步探索一系列时域和频域的线性和非线性控制问题等。

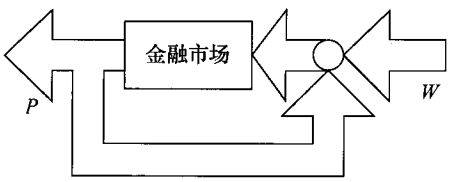

图 0-3 金融宏观调控系统

我们以日本和中国房地产金融市场为例说明 $W\sim P$ 控制关系。众所周知，日本的房市在 20 世纪 90 年代的大幅下跌，以至于至今日本经济仍然在它的阴影覆盖下，根据邱敏和曾向荣（2007）的报道，起因就是“1990 年 9 月，日本国营广播电视台 NHK 连续 5 个晚上在黄金时段播放了有关土地问题的特别节目，指出地价是可以下跌的，……这一节目像颗重磅炸弹，其巨大的舆论冲击力揭开了日本泡沫经济破灭的序幕。以这一节目的播出为转折点，日本的地价自二战结束以来第一次开始急速下跌。”然而，在电视台连续几天预测房价下跌之前，当时全世界都认为日本的房价有泡沫了，也讨论了很长时间这个泡沫了。可见，房价 P 对 W 有时非常迟钝，而超过一定阈值时又可能非常敏感。当然，本质的原因还是有泡沫，W 只是导火索而已。我们还可以以 2007 年秋冬之际中国房价为例。目前，政府和各新闻媒体不断提醒购房者房价可能下降的风险，大家也在普遍担心是否会出现房价大跌的情形，然而，到作者出版此书时，

北京地区的房价不降反升。事实上，究竟房价是升是降，这个动作是一个“将来时”。所以，怎样使用 W 控制 P 是一个值得研究而且应用性很强的问题。

应当说明的是，本书只是列举一些作者目前想到的网络金融的新课题，远远不是互联网金融信息的挖掘与预测的总结之作，相反地，这仅仅是一个开始。在这个领域，本书的研究是初步的，每个问题都可以进一步展开，在内涵上不仅需要更深一步的研究，在外延上也还有更多的问题需要探究。此外，很多研究还不得不等待自然语言处理技术的进展。还有，我们服务器的数据量积累也不是很大，对很多问题也只能做个初步实验。因此，本书只能称为导论，很多问题尚未囊括进来。后面究竟有多少问题，讲实话，作者也是“雾里看花”，而且时间序列本身就是一个非常复杂的问题。分析它的方法很多，本书更像是提出一些 ideas，旨在抛砖引玉，以便与更多的专家共同探讨这类问题时，有这样一本小册子为大家服务。

这本小册子所探讨的网络金融信息挖掘是一个很大的领域，例如，本书中对待交易量的各种 GARCH 类的方法，以及对待收益率的各种方法，可以互为参考，从而引出更多的实证分析工作，这将在下一本书《网络金融信息分析》中尝试完成。此外，本书只是对简单的一类网络金融信息流时间序列，即单位时间的信息数量进行了分析，对更复杂类型的网络金融信息流时间序列，例如基于自然语言理解的、基于信息熵的网络金融信息流时间序列，也将留待《网络金融信息分析》去完成。此外，《网络金融信息分析》将从计算机智能角度开始展开，做更多分析建模。所以，从这个意义上来说，这本小册子称为导论是名副其实。

在热心的同事和朋友多次善意的催促下，本书得以出版，然而，作者对于本书内容的发表还是觉得有些仓促，有些地方还需要再进一步斟酌。不过话又讲回来，和任何一个新的学科分支一样，“网络金融信息挖掘”的架构和内容也必然是一个不断完善发展的过程，而这个过程靠作者一个人是不可能实现的，必然需要各位同行专家的共同努力。因此，从这个意义上说，本书的内容虽然还不甚完善，但如能引起同行研究者的兴趣，并在此基础上大家共同致力于网络金融信息挖掘领域的研究，将是本书最大的成功。

本书的研究出版受到了国家自然科学基金资助项目(70571003)、留学回国启动基金(4131522)及北京大学(2007)研究生课程立项建设的资助。在编写过程中，还得到作者所在单位北京大学计算机所领导的大力支持以及一些同事和同学的热心帮助。在此一并致谢。

此外，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 webmine 数据库免费获取(需要事先索取用户名和口令[①])。

本书和前面提到的已经出版的四本书一起，共同组成北京大学计算机系研究生选

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。

修课教学和辅导材料。课程网页上还提供了另外一些相关材料，网页的地址是http://www.icst.pku.edu.cn/course/efinance/2007/index.html，在网页上还提供了另外一些相关材料。

由于作者水平和时间的限制，书中定会存在不少不足和错误，恳请读者批评指正。作者会将更正及时发表在网上，作为本书的一个补充。

梁 循

2007年9月于北京大学燕北园

目　录

第1章 绪 论

众所周知，金融市场是一个很复杂的系统。股市的运作规律，至今为止，一直是人们努力探求的内容。本章力图从一个新的角度——互联网信息量、股市交易量和股价出发，概述它们的特点以及相互之间的关联，这些研究对于投资者评价股票价值和进行投资都具有一定的参考作用。

1.1 金融信息对市场影响的研究

很多年来，基于股市信息的股价预测是一个重要的金融方向，它已经吸引很多金融学家和股市分析家。金融理论认为股市信息对市场的影响是至关重要的，它起着一个对市场调节的杠杆作用。当金融信息进入市场时，股市的平均收益将做出相应的变化调整。从某种意义上说，股市信息对股市收益是有预测作用的。所以，我们有理由重视股市信息对股市的影响。

一般在市场上涌现的消息由于性质上会有差异，可带来市场价格的波动。金融市场的波动在某种程度上看来是可以被预知的，这就需要从导致市场波动的原因出发来进行研究。目前看来，原因是多方面的，既有宏观经济因素、政策因素的影响，又有国际金融因素的影响，同时还会受到投资者心理预期等因素的制约。对于金融市场的重要组成部分的股票市场而言，除了这些影响因素，还要受到包括行业前景和公司业绩等因素的影响。这些因素的变化常常会被各种媒介的信息报道所反映。毫无疑问，如果我们捕捉并研究这些报道，就可以观察到他们对市场产生的影响。

信息理论告诉我们，信息量是消息中意外的数量(常恫，1993)，因此信息过程是独立的，当前的价格已经吸收了从过去和现在的信息集推断的所有知识，而新信息的到达是一独立事件。各种信息，影响着投资者的心理，传播到股市，刺激股民进行交易。一般地，当利多消息大量出现时可能会造成投资者进行较多的买进活动，即好的信息可以刺激购买，可能引起金融资产价格的上涨，使股价增长，造成市场向上运动；相反地，当利空消息大量涌现时常常会造成投资者进行较多的卖出活动，即坏的信息可以刺激卖出，从而引起金融资产价格的下降，造成市场向下运动。这类信息包括：利率、汇率，工业增长、物价，以及上市公司信息，例如分红、盈利公告、年报、消费价格指数(consumer price index，CPI)、谣言，还有社会上的突发事件。假定价格具有充分弹性，交易者对市场具有完全的信息。当新信息到达时，市场参数发生变动，价格能瞬间做出调整，平均收益必然会对新信息做出反应而产生变化。也就是说，新信息进入市

场后，立即被市场吸收，市场瞬间达到新的均衡。因此，新信息的进入并不影响市场均衡，信息理论和均衡模型是相容的，均衡可以认为是连续的。

市场可作为信息交换的场所，市场和价格是信息的一种测度，是提供给经济主体最好做什么的基本判断标准。一般信息以时间离散的方式到达，金融资产收益序列往往是一个围绕均值跳动的过程。由于市场存在摩擦，信息序贯地到达，金融资产的定价机制即是对序贯信息传播过程的反应。对股市信息和股价的关系问题，大部分研究集中于股市对某条股市信息的反应。显然，要整体地研究信息对股市的影响，只考虑一条信息是不全面的。在本书中，我们记单位时间这些信息量为随机变量 W，这是一个计算机-金融变量。如果单位时间由 0、1、2 开始，这些按单位时间序贯到达的信息量就形成了网络金融信息量时间序列 $W_t, t=0,1,2,\cdots$。本书中，有时在上下文明显的地方，为了避免繁琐，也简单地称 W_t 为 W。

股票市场的例子也很有代表性，根据市场上信息量的强弱，我们可以对价格波动与成交量的联系进行判断。在信息量极少或重要性不大的情形下，市场相对是平静的，成交量代表的信息流不发生作用；相反地，当影响市场的各种金融事件发生频繁，股票价格的波动就主要受到市场上的信息流的影响，此时成交量对股价波动的影响会很大。

对于中国国内的股票市场而言，经历了十多年的发展已经开始正规和成熟起来。与欧美的股票市场相比，还有一些不一样的特点，在中国股票市场上短线投资者众多，真正长线的投资者比较少，而短线投资者对于即时信息的依赖程度就比较高。此外，中国国内的股民炒作能力很强，导致某些时候会出现一些暴涨的新闻信息量。中国股票市场的一些政策法规出台也会对它产生巨大的影响。也就是说，信息在中国股市的作用很大，引起国内股市变动的可能性要比欧美股市中更显著。

在对金融信息影响市场的研究中，以对股票市场的研究最为典型。在互联网产生之前，研究限于电视和报纸上的股市信息对股价的关系，一般是通过预先给定股市信息内容的方式来研究股价变化，由于此类研究的局限性，不易对所有的报纸和电视的股市信息进行逐日乃至逐小时的收集，导致研究的不充分，仅仅分析了一些具有特殊代表性的股市信息。而互联网出现为此类研究提供了巨大的便利，网上股市信息的收集成为了轻而易举的事情，因此可以从整体上来分析股市信息总强度 W 的特性，进而研究它与股价波动的联系。

一般的看法是，交易量是影响股价变化的重要因素，大量实验研究的结果也支持了两者之间变化的显著关系。试图用交易量解释股价行为的模型有三类：一是信息理论模型，该模型认为信息导致了交易量和股价的变动；二是交易理论模型，该模型认为投资者习惯在市场活跃时候进行交易，即交易量和股价有集群性；三是理念分散模型，该模型认为投资者对金融信息的看法越分散，股价波动越大，交易量也越大。目前，信息理论模型的研究比较多，实证研究也更多地支持了这一理论的模型（朱永安和

石礼英,2003)。

在互联网时代到来之前,股市信息在传播途径主要是报纸和电视,以及人们的口头传播。显然,在这个条件下,直接度量某单位时间的金融信息流量是不可能的。互联网时代到来后,在互联网上发布的股市信息迅速增长。如今,股市信息渗透互联网的每一个角落。在互联网在帮助下,计算机服务器可以非常容易地处理成千上万条互联网股市信息,仅仅因为互联网容纳了海量股市信息并且互联网信息可以非常容易地被我们的计算机服务器获取。今天,互联网已经变成了一个传播股市信息的非常重要的媒体。

比较中国和欧美股市,中国股市还是一个十几岁的少年。有的学者列出了很多中国股市不同于一个成熟的市场的许多方面,不少学者认为,中国市场的这种特征使得她对金融信息更敏感。在金融时间序列分析领域,已经有很多学者进行了研究,并且取得了很多成果。

1.2 网络金融信息挖掘的研究领域

互联网上的信息量众多,是当前最大和最普及的半文本文档仓库。互联网已经深刻地影响到我们生活的许多方面,改变着人们的生活方式。想要更好地利用互联网上的海量信息,需要将它们系统地组织起来,以免被浩如烟海的网页所淹没,因此,针对网页数据的挖掘,或自动从互联网中发现有价值信息已经成为数据挖掘中的一个重要课题。伴随着互联网的不断成长,搜索引擎技术和互联网信息挖掘技术产生和发展起来了。搜索引擎的诞生使得用户可以使用自定义的关键词在网页数据库中寻找感兴趣的内容,互联网的信息挖掘技术则更为复杂,它提供对网页数据的深层处理功能,可挖掘出网页包含的知识以及各网页之间的联系,帮助人们进行决策。

作为数据挖掘领域的一个新的分支,针对互联网的数据挖掘主要是指利用数据挖掘技术从互联网数据中抽取知识,它是传统的数据挖掘技术在互联网上的应用,所处理的是半结构化的网页。按照互联网挖掘的数据来源进行分类,目前的互联网挖掘可分为互联网信息结构挖掘、互联网信息元数据挖掘、互联网信息内容挖掘、互联网使用信息挖掘以及互联网信息摘要和总结。

互联网作为投资者获取金融信息的一个渠道而言,不同于其他媒介。电视和报刊传播的金融信息以音像和纸张为载体,不易使用计算机来进行处理,并且各类媒体之间的重要内容往往会互相重复,而由于互联网信息更易于获取,能最大程度地与计算机技术相结合,所以更便于进行相关的研究。对于从事金融研究的工作者而言,由于在互联网上有着众多的金融信息,获取十分方便,使得我们可以从更大量上对金融信息和金融市场的关联进行实证,包括进行相关性的统计分析,而不仅仅停留在对若干典型金融信息的描述性分析上。

网络金融信息挖掘问题可以粗略地分为两大块：网络金融信息的搜索等浅层挖掘，以及网络金融信息流与金融市场的关联等深层挖掘(见图 1-1)。

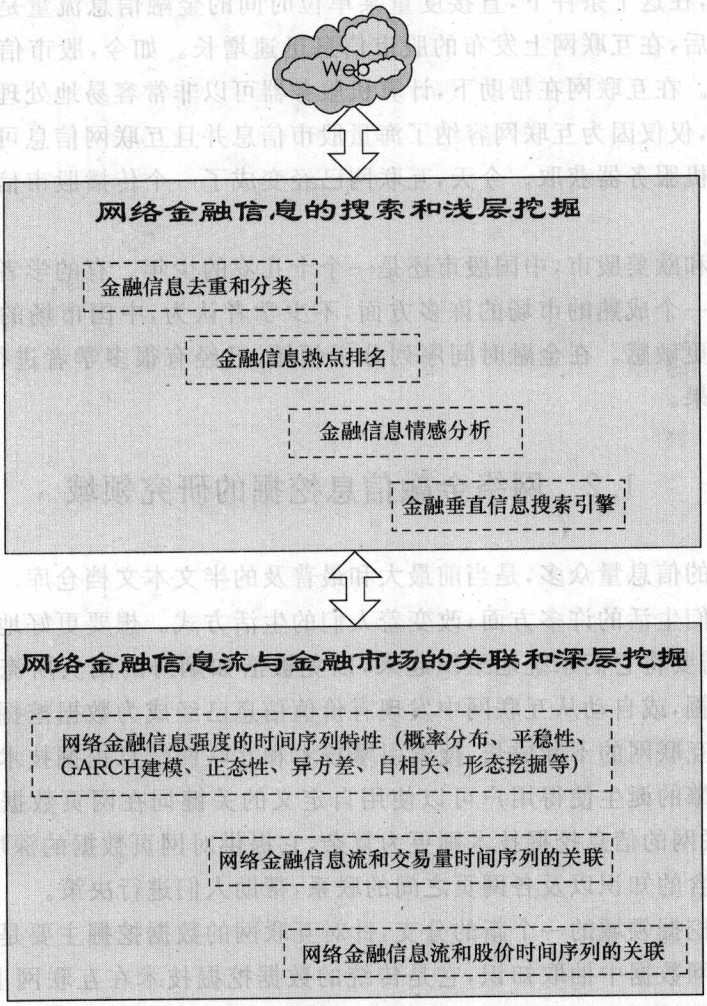

图 1-1 网络金融信息挖掘内容

20 世纪末，HTML 语言的产生和应用迅速推动了 Web 的发展，使得网络这一原本限制在专业人员的范围内的事务，进入公司学校，进入家家户户。可以说，网络的迅猛发展，HTML 功不可没。但是，由于设计 HTML 时只定义了内容的呈现形式，例如字体的大小、颜色，只考虑了人的阅读，并未定义内容的含义，未考虑机器的阅读，使得后来的搜索引擎和挖掘徒增了不少负担(梁循，2004—2007)，而且速度和质量都受到不小影响。如果将来 Web 上使用 XML 取代 HTML，则机读的问题可以得到解决，因为 XML 不但可以给出了网页的呈现形式，而且给出了其内容的含义。在这种条件下，我们相信，互联网搜索引擎和挖掘的速度和质量都可以得到很大提供，并且在实用性

上得到用户的广泛认可。

由于信息量比较大的公司与信息量相对较少的公司的表现常常不同，所以，在本书中不少地方是将两者分开讨论的。为了叙述方便，我们把信息量比较大的公司简单地称为大公司，而信息量比较小的公司简单地称为小公司。

1.3 信息量 W 与股价 P 及交易量 V 的关联

在金融市场上，股市投资者依赖股市信息来进行股票交易，好的市场信息可刺激购买、提高股价，同时坏的信息可能造成交易时的股价降低，股市信息在很大程度上与投资者在股市上的各种活动相关联。各种信息每天在股市上传播，影响着投资者的心理并激励他们做决定。从来源上看，股市信息大体可以分为股市新闻和股市消息。股市新闻指公司、政府或新闻机构正式公布的资料，例如公司的年度报告，政府公布的消费价格指数，全球的紧急事件。股市消息指个人财务分析师和个人投资者在网上张贴的非正式信息。本书的研究主要针对股市新闻。

“混合分布假说”(mixture distribution hypothesis，MDH)认为，股价的变化率服从相互独立、对称的、相对尖峰的近似正态分布，而价格和交易量的联合分布是被一种潜在的混合变量共同驱动，这个混合变量常常被假设为信息流。金融学者们一直在致力于研究信息与股价 P(或收益率 R)的关系，既涉及质也涉及量。从单条金融信息质上，学者们已经有很多关于信息与股价关系的成果。从单位时间的金融信息量 W 这个课题上，在互联网产生之前，金融信息主要通过报纸、电视传播，计算机处理这些金融信息，这就需要首先进行文字识别和语音识别，所以，要按指定时段收集整理大量金融信息、再按多种要求分类，如果使用人工的办法，工作量是巨大的。人们观察到，金融信息量和股市交易量 V 有相似的波动规律，于是，一些学者借用股市交易量作为金融信息量的替代变量，通过研究股市交易量与股价之间的关系，来研究金融信息量与股价之间的关系(见图 1-2 实线部分)。

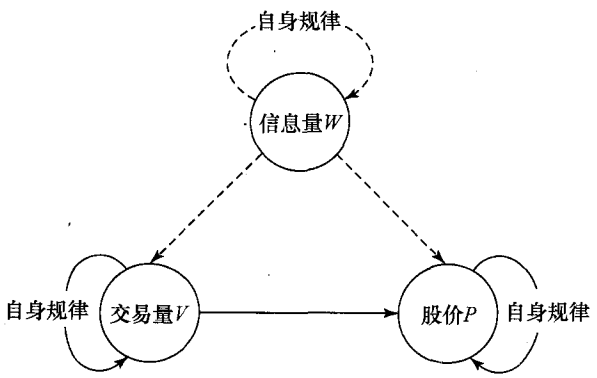

图 1-2 金融信息挖掘研究实体之间的关系