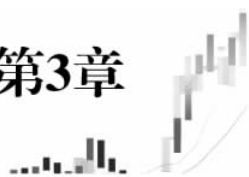


第3章



股票数据的准备

3.1 股票相关数据

长期以来,证券公司的交易系统一直处于中国各行业 IT 技术应用的领先水平。证券业已经成为当今中国计算机应用高度密集的行业之一,高度的信息化使其积累了大量的数据。股票相关的基础数据主要包括三部分:业务数据、行情数据和证券文本数据。

1. 业务数据

业务数据包括结算数据、过户数据和交易系统数据等。结算数据是由深圳和上海证券登记公司以交易席位为单位发布的证券公司当日资金、股份交收明细以及分红、送股、配股等数据。过户数据是由深圳和上海证券交易所以交易席位为单位发布的证券公司当日投资者买卖证券的过户明细数据。结算数据和过户数据由证券交易所通过地面和卫星网络系统发送到证券公司。交易系统数据是证券公司最重要和最实时的数据,它由交易系统在实时交易中产生,是进行数据挖掘、客户分析、构建 CRM 系统的主要基础数据。上述业务数据属于相对私有的数据,这些数据反映了客户的资金状况、交易状况和持仓状况等,对证券公司和交易所而言具有极高的分析价值。

2. 行情数据

行情数据是由深圳、上海证券交易所在开市期间发布的证券实时交易的成交撮合数据,是进行股市行情分析的关键数据。上海证券交易所的行情数据主要包括一张实时刷新的表,其类似结构如表 3-1 所示。深圳证券交易所的行情数据主要包括三张实时刷新的表:证券信息表、证券行情表、证券指数表。各表类似结构如表 3-2~表 3-4 所示。

表 3-1 上海证券交易所的行情数据表

字段名	字段说明	类型	长度	小数位数
S1	证券代码	CHAR	6	
S2	证券名称	CHAR	8	
S3	前收盘价格	NUM	8	3
S4	今开盘价格	NUM	8	3

续表

字段名	字段说明	类型	长度	小数位数
S5	今成交金额	NUM	12	
S6	最高价	NUM	8	3
S7	最低价	NUM	8	3
S8	最新价	NUM	8	3
S9	当前买入价	NUM	8	3
S10	当前卖出价	NUM	8	3
S11	成交数量	NUM	10	
S12	市盈率	NUM	8	3
S13	申买量一	NUM	10	
S14	申买价二	NUM	8	3
S15	申买量二	NUM	10	
S16	申买价三	NUM	8	3
S17	申买量三	NUM	10	
S18	申卖量一	NUM	10	
S19	申卖价二	NUM	8	3
S20	申卖量二	NUM	10	
S21	申卖价三	NUM	8	3
S22	申卖量三	NUM	10	
S23	申买价四	NUM	8	3
S24	申买量四	NUM	10	
S25	申买价五	NUM	8	3
S26	申买量五	NUM	10	
S27	申卖价四	NUM	8	3
S28	申卖量四	NUM	10	
S29	申卖价五	NUM	8	3
S30	申卖量五	NUM	10	

表 3-2 深圳证券交易所证券信息表

序号	字段名	字段描述	类型	长度
1	XXZQDM	证券代码	CHAR	6
2	XXZQJC	证券简称	CHAR	8
3	XXYWJC	英文简称	CHAR	20
4	XXJYDW	交易单位	NUM	4,0
5	XXHYZL	行业种类	CHAR	3
6	XXHBZL	货币种类	CHAR	2
7	XXMGMZ	每股面值	NUM	7,2
8	XXZFXL	总发行量	NUM	12,0
9	XXLTGS	流通股数	NUM	12,0
10	XXSNLR	上年每股利润	NUM	9,4
11	XXBNLR	本年每股利润	NUM	9,4
12	XXJSFL	经手费率	NUM	7,6
13	XXYHSL	印花税率	NUM	7,6

续表

序号	字段名	字段描述	类型	长度
14	XXGHFL	过户费率	NUM	7,6
15	XXSSRQ	上市日期	DATETIME	8
16	XXDJRQ	到期/交割日	DATETIME	8
17	XXMBXL	每笔限量	NUM	9,0
18	XXBLDW	买数量单位	NUM	6,0
19	XXSLDW	卖数量单位	NUM	6,0
20	XXJGDW	价格挡位	NUM	5,3
21	XXJHCS	集合竞价限价参数	NUM	7,3
22	XXLXCS	连续竞价限价参数	NUM	7,3
23	XXXJXZ	限价参数性质	NUM	1,0
24	XXZTJG	涨停价格	NUM	9,3
25	XXDTJG	跌停价格	NUM	9,3
26	XXZHBL	折合比例	NUM	5,2
27	XXJYZT	交易状态	CHAR	1
28	XXZQJB	证券级别	CHAR	1
29	XXTPBZ	停牌标志	CHAR	1
30	XXCQCX	除权除息标志	CHAR	1
31	XXCFBZ	成分股标志	CHAR	1

表 3-3 深圳证券交易所证券行情表

序号	字段名	字段描述	类型	长度
1	HQZQDM	证券代码	CHAR	6
2	HQZQJC	证券简称	CHAR	8
3	HQZRSP	昨日收盘价	NUM	9,3
4	HQJRKP	今日开盘价	NUM	9,3
5	HQZJCJ	最近成交价	NUM	9,3
6	HQCJSL	成交数量	NUM	12,0
7	HQCJJE	成交金额	NUM	17,3
8	HQCJBS	成交笔数	NUM	9,0
9	HQZGCJ	最高成交价	NUM	9,3
10	HQZDCJ	最低成交价	NUM	9,3
11	HQSYL1	市盈率 1	NUM	7,2
12	HQSYL2	市盈率 2	NUM	7,2
13	HQJSD1	价格升降 1	NUM	9,3
14	HQJSD2	价格升降 2	NUM	9,3
15	HQHYCC	合约持仓量	NUM	12,0
16	HQSJW5	卖价位五	NUM	9,3
17	HQSSL5	卖数量五	NUM	12,0
18	HQSJW4	卖价位四	NUM	9,3
19	HQSSL4	卖数量四	NUM	12,0
20	HQSJW3	卖价位三	NUM	9,3
21	HQSSL3	卖数量三	NUM	12,0

续表

序号	字段名	字段描述	类型	长度
22	HQSJW2	卖价位二	NUM	9,3
23	HQSSL2	卖数量二	NUM	12,0
24	HQSJW1	卖价位一	NUM	9,3
25	HQSSL1	卖数量一	NUM	12,0
26	HQBJW1	买价位一	NUM	9,3
27	HQBSL1	买数量一	NUM	12,0
28	HQBJW2	买价位二	NUM	9,3
29	HQBSL2	买数量二	NUM	12,0
30	HQBJW3	买价位三	NUM	9,3
31	HQBSL3	买数量三	NUM	12,0
32	HQBJW4	买价位四	NUM	9,3
33	HQBSL4	买数量四	NUM	12,0
34	HQBJW5	买价位五	NUM	9,3
35	HQBSL5	买数量五	NUM	12,0

表 3-4 深圳证券交易所证券指数表

序号	字段名	字段描述	类型	长度
1	ZSZSDM	指数代码	CHAR	6
2	ZSZSQC	指数全称	CHAR	12
3	ZSYWMC	英文名称	CHAR	20
4	ZSSSZS	昨日收市指数	NUM	11,4
5	ZKKSZS	今日开市指数	NUM	11,4
6	ZSZGZS	最高指数	NUM	11,4
7	ZSZDZS	最低指数	NUM	11,4
8	ZSZJZS	最近指数	NUM	11,4
9	ZSCJSL	成交数量	NUM	12,0
10	ZSCJJE	成交金额	NUM	17,3

3. 证券文本数据

狭义的证券文本数据是指由证券交易所通过证券卫星发送的与证券领域有关的政策和各股资讯等实时信息。广义的证券文本数据是指由各种传媒方式发布的与证券相关的信息,主要包括卫星、电视、广播、Internet、书刊杂志等传媒方式,其中 Internet 是涵盖信息量最多的传媒方式。

3.2 数据的获取源

在大数据时代,金融机构之间的竞争将在网络信息平台上全面展开,说到底就是“数据为王”。金融行业对数据有极强的依赖性,具有数据量大、时效性高的特点。谁掌握了数据,谁就拥有风险定价能力,谁就可以获得高额的风险收益,最终赢得竞争优势。

很多财经网站都有丰富、可靠的交易数据。获取数据的方式可以从网站的相关页面进

行抓取解析,也可以通过特定的接口调用来获取,还可以通过提供的资源所在位置进行直接下载等。

要获取这些数据,有免费的也有收费的渠道。收费的渠道如:①从数据提供商 Wind 购买高质量的数据,其优点是更新及时,缺点是价格比较贵;②通过购买 API 调用接口,如阿里云集成了多家第三方供应商提供的接口调用服务,价格上按调用次数进行计费,相对便宜一些,但更新上也相对慢一些。收费获取数据的渠道有很多,这里不再赘述。下面主要介绍一些免费的数据获取源。

3.2.1 从雅虎获取历史交易数据

从雅虎财经频道可以方便地获取股票的历史数据。以万科为例,访问以下网址: <https://finance.yahoo.com/quote/000002.SZ/history?p=000002.SZ>,如图 3-1 所示,单击 Download Data 可以直接下载个股的历史数据。网址中有两个参数:一个是股票代码;另一个是股票所在的市场,其中后缀.SS 代表沪市,.SZ 代表深市。下载的数据为 Excel 格式,内容如图 3-2 所示。表格中包括七个字段的数据,分别为交易日期、开盘价、最高价、最低价、收盘价、向前复权价、成交量。

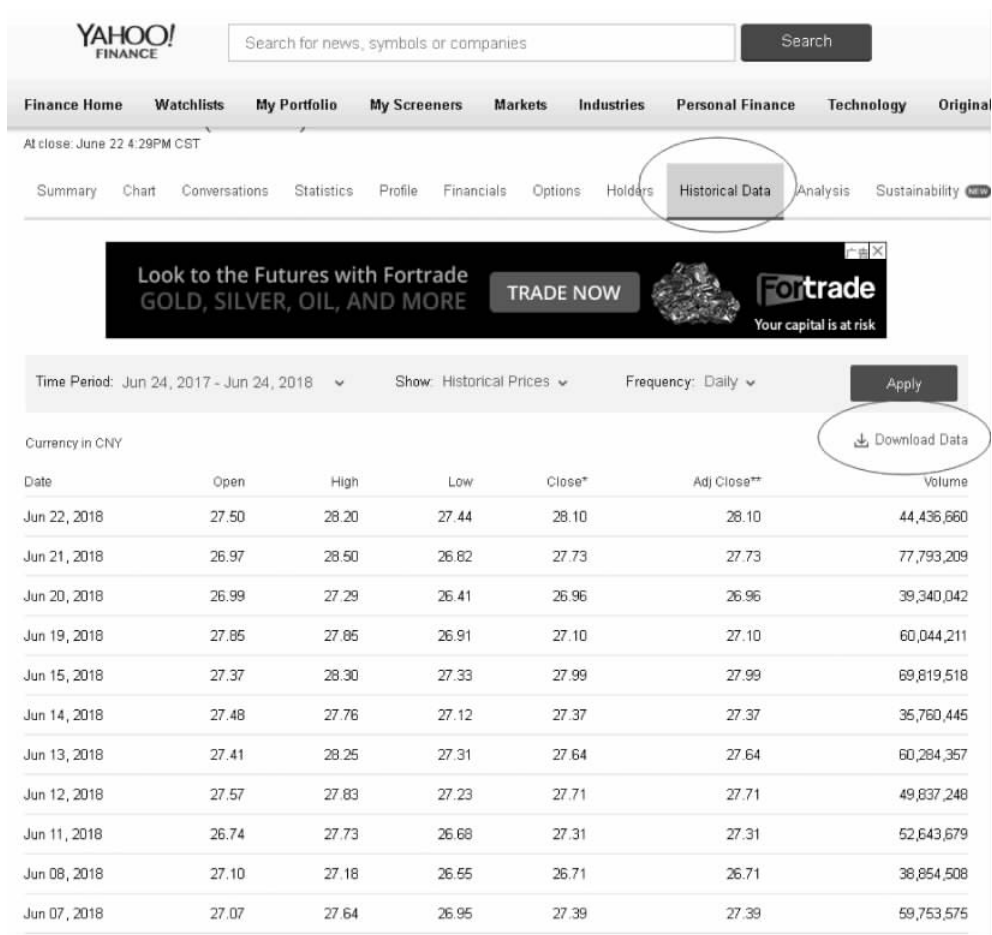
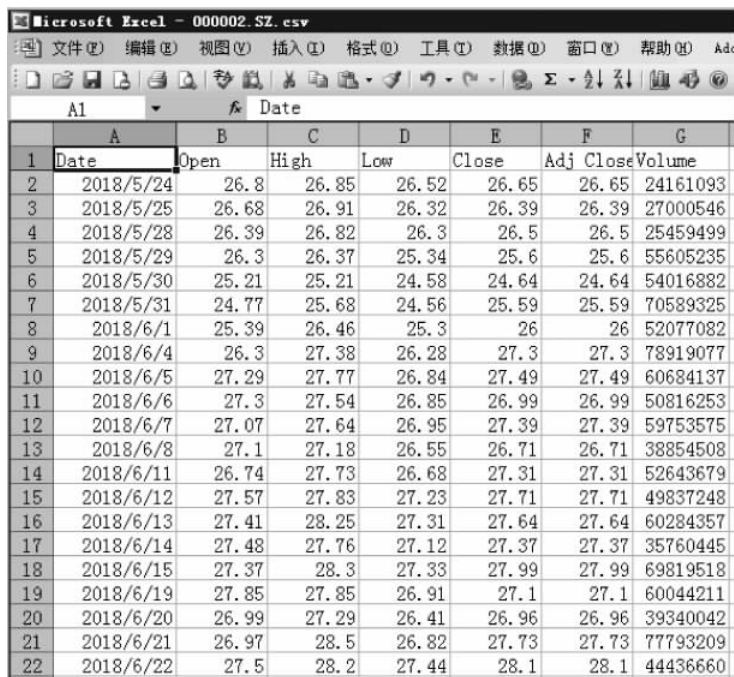


图 3-1 雅虎财经频道上个股的历史数据页面



	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Adj Close	Volume
2	2018/5/24	26.8	26.85	26.52	26.65	26.65	24161093
3	2018/5/25	26.68	26.91	26.32	26.39	26.39	27000546
4	2018/5/28	26.39	26.82	26.3	26.5	26.5	25459499
5	2018/5/29	26.3	26.37	25.34	25.6	25.6	55605235
6	2018/5/30	25.21	25.21	24.58	24.64	24.64	54016882
7	2018/5/31	24.77	25.68	24.56	25.59	25.59	70589325
8	2018/6/1	25.39	26.46	25.3	26	26	52077082
9	2018/6/4	26.3	27.38	26.28	27.3	27.3	78919077
10	2018/6/5	27.29	27.77	26.84	27.49	27.49	60684137
11	2018/6/6	27.3	27.54	26.85	26.99	26.99	50816253
12	2018/6/7	27.07	27.64	26.95	27.39	27.39	59753575
13	2018/6/8	27.1	27.18	26.55	26.71	26.71	38854508
14	2018/6/11	26.74	27.73	26.68	27.31	27.31	52643679
15	2018/6/12	27.57	27.83	27.23	27.71	27.71	49837248
16	2018/6/13	27.41	28.25	27.31	27.64	27.64	60284357
17	2018/6/14	27.48	27.76	27.12	27.37	27.37	35760445
18	2018/6/15	27.37	28.3	27.33	27.99	27.99	69819518
19	2018/6/19	27.85	27.85	26.91	27.1	27.1	60044211
20	2018/6/20	26.99	27.29	26.41	26.96	26.96	39340042
21	2018/6/21	26.97	28.5	26.82	27.73	27.73	77793209
22	2018/6/22	27.5	28.2	27.44	28.1	28.1	44436660

图 3-2 从雅虎下载的个股历史数据

如果是上述直接单击下载的方式,会发现仅有近一个月的历史数据。如果想让机器自动收集所有个股的历史数据,这时得用构造链接参数的方式。仍以万科为例,通过分析可知其实际的下载 URL 地址是: <https://query1.finance.yahoo.com/v7/finance/download/000002.SZ?period1=1527156971&period2=1529835371&interval=1d&events=history&crumb=Gp4B1z0Shfi>。

该 URL 中除了股票代码和股票所在的市场参数外,还包含几个参数: period1 是起始时间; period2 是结束时间; 其他的参数保持默认即可。需要注意,这里的时间是用 1970 年 1 月 1 日 0 点 0 分 0 秒到目标时间经过的秒数表示的,因此构造填写时需要做时间变换。通过链接构造方式,一方面可以改变时间参数获得个股的所有历史数据,另一方面可以通过变换股票代码获得所有股票的历史数据,并让机器完成自动收集。

3.2.2 从腾讯获取实时交易数据

1. 实时行情数据接口

腾讯实时行情数据的获取接口为 [http://qt.gtimg.cn/q=\(1\)\(2\)](http://qt.gtimg.cn/q=(1)(2))。接口需要提供两个参数: 第一个参数是股票所在的市场, sh 代表沪市, sz 代表深市; 第二个参数是股票代码。以万科为例,即 sz000002。调用构造的 URL,即 <http://qt.gtimg.cn/q=sz000002>,将会返回一串文本,例如:

```
v_sz000002="51~万科 A~000002~28.10~27.73~27.50~444366~210147~
234131~28.09~174~28.08~1645~28.07~1383~28.06~975~28.05~70~28.10~
435~28.11~129~28.12~223~28.13~363~28.14~353~15:00:04/28.10/12266/S/
```

34467460/9622|14:57:00/28.06/145/S/406914/9524|14:56:57/28.06/148/S/415299/9522|14:56:54/28.06/88/S/246928/9520|14:56:51/28.06/119/B/333854/9518|14:56:48/28.06/179/B/502280/9516~20180622150136~0.37~1.33~28.20~27.44~28.10/444366/1235687452~444366~123569~0.46~10.98~28.20~27.44~2.74~2729.96~3102.00~2.28~30.50~24.96~0.79~2744~27.81~86.66~11.06";

这个文本串由许多数据拼接在一起,不同含义的数据用“~”隔开,即以“~”分隔文本串中内容,下标从0开始,依次为:

- 0: “51”,未知
- 1: “万科 A”,股票名称
- 2: “000002”,股票代码
- 3: “28.10”,当前价格
- 4: “27.73”,昨收
- 5: “27.50”,今开
- 6: “444366”,成交量(手)
- 7: “210147”,外盘
- 8: “234131”,内盘
- 9: “28.09”,买一
- 10: “174”,买一量(手)
- 11: “28.08”,买二
- 12: “1645”,买二量(手)
- 13: “28.07”,买三
- 14: “1383”,买三量(手)
- 15: “28.06”,买四
- 16: “975”,买四量(手)
- 17: “28.05”,买五
- 18: “70”,买五量(手)
- 19: “28.10”,卖一
- 20: “435”,卖一量
- 21: “28.11”,卖二
- 22: “129”,卖二量
- 23: “28.12”,卖三
- 24: “223”,卖三量
- 25: “28.13”,卖四
- 26: “363”,卖四量
- 27: “28.14”,卖五
- 28: “353”,卖五量
- 29: “15:00:04/28.10/12266/S/34467460/9622|
14:57:00/28.06/145/S/406914/9524|
14:56:57/28.06/148/S/415299/9522|

14:56:54/28.06/88/S/246928/9520|
14:56:51/28.06/119/B/333854/9518|
14:56:48/28.06/179/B/502280/9516”,最近逐笔成交
30: “20180622150136”,时间
31: “0.37”,涨跌
32: “1.33”,涨跌 %
33: “28.20”,最高
34: “27.44”,最低
35: “28.10/444366/1235687452”,价格/成交量(手)/成交额
36: “444366”,成交量(手)
37: “123569”,成交额(万)
38: “0.46”,换手率
39: “10.98”,市盈率
40: “”,未知
41: “28.20”,最高
42: “27.44”,最低
43: “2.74”,振幅
44: “2729.96”,流通市值(亿)
45: “3102.00”,总市值(亿)
46: “2.28”,市净率
47: “30.50”,涨停价
48: “24.96”,跌停价
49: “0.79”,量比
50: “2744”,委差
51: “27.81”,加权平均价
52: “86.66”,PE(动)
53: “11.06”,未知

若要查询多个股票,则在 URL 最后加上一个逗号,再加上股票代码即可。例如要一次查询万科 A(sz000002)和平安银行(sz000001)的行情,构造的 URL 如下: <http://qt.gtimg.cn/q=sz000002,sz000001>。另外上证指数的代码是 sh000001,深证成指的代码是 sz399001。如果只需获取股票交易的基本数据则在之前加“s_”,即 http://qt.gtimg.cn/q=s_sz0000002。

2. 实时资金流向接口

腾讯实时资金流向的获取接口为 [http://qt.gtimg.cn/q=ff_\(1\)\(2\)](http://qt.gtimg.cn/q=ff_(1)(2))。

接口需要提供两个参数:第一个参数是股票所在的市场,sh 代表沪市,sz 代表深市;第二个参数是股票代码。以万科为例,即 sz000002。调用构造的 URL,即 http://qt.gtimg.cn/q=ff_sz000002,将会返回一串文本,例如:

```
v_ff_sz000002="sz000002~73049.00~72789.70~259.30~0.21~50519.80~50779.10~-259.30~-0.21~123568.80~533193.8~478832.8~万科 A~20180622~
```

20180621^151818.70^116892.70~20180620^59425.90^58060.50~20180619^112247.50^101243.90~20180615^136652.70^129846.00";

以~分隔字符串中内容,下标从0开始,依次为:

0: "sz000002",股票代码

1: "73049.00",主力流入

2: "72789.70",主力流出

3: "259.30",主力净流入

4: "0.21",主力净流入÷资金流入流出总和×100

5: "50519.80",散户流入

6: "50779.10",散户流出

7: "-259.30",散户净流入

8: "-0.21",散户净流入÷资金流入流出总和×100

9: "123568.80",资金流入总和,即1+5;或资金流出总和,即2+6

10: "533193.8",未知

11: "478832.8",未知

12: "万科A",股票

13: "20180622",交易日期

14: "20180621^151818.70^116892.70",前1交易日/主力流入/主力流出

15: "20180620^59425.90^58060.50",前2交易日/主力流入/主力流出

16: "20180619^112247.50^101243.90",前3交易日/主力流入/主力流出

17: "20180615^136652.70^129846.00",前4交易日/主力流入/主力流出

3. 实时盘口分析

腾讯实时盘口分析的获取接口为 [http://qt.gtimg.cn/q=s_pk\(1\)\(2\)](http://qt.gtimg.cn/q=s_pk(1)(2))。接口需要提供两个参数:第一个参数是股票所在的市场,sh代表沪市,sz代表深市;第二个参数是股票代码。以万科为例,即sz000002。调用构造的URL,即[http://qt.gtimg.cn/q=s_pk\(1\)\(2\)](http://qt.gtimg.cn/q=s_pk(1)(2)),将会返回一串文本,例如:

v_s_pk(1)(2)sz000002="0.194~0.278~0.236~0.292";

以~分隔字符串中的内容,下标从0开始,依次为:

0: "0.194",买盘大单占比

1: "0.278",买盘小单占比

2: "0.236",卖盘大单占比

3: "0.292",卖盘小单占比

3.2.3 从新浪获取交易数据

1. 实时行情数据接口

新浪实时行情数据的获取接口为 [http://hq.sinajs.cn/list=\(1\)\(2\)](http://hq.sinajs.cn/list=(1)(2))。

接口需要提供两个参数:第一个参数是股票所在的市场,sh代表沪市,sz代表深市;第二个参数是股票代码。以万科为例,即sz000002。调用构造的URL,即<http://hq.sinajs.cn/list=sz000002>,将会返回一串文本,例如:

```
var hq_str_sz000002="万科 A, 27. 500,27. 730,28. 100,28. 200,27. 440,28. 090,  
28. 100,44436660,1235687451. 840,17400,28. 090,164500,28. 080,138300,28. 070,97500,  
28. 060,7000,28. 050,43529,28. 100,12900,28. 110,22300,28. 120,36359,28. 130,35300,  
28. 140,2018-06-22,15:05:03,00";
```

这个字符串由许多数据拼接在一起,不同含义的数据用“,”隔开,按照程序员的思路,顺序号从 0 开始。各位置的数据含义如下:

- 0: “万科 A”,股票名字
- 1: “27. 500”,今日开盘价
- 2: “27. 730”,昨日收盘价
- 3: “28. 100”,当前价格
- 4: “28. 200”,今日最高价
- 5: “27. 440”,今日最低价
- 6: “28. 090”,竞买价,即买一报价
- 7: “28. 100”,竞卖价,即卖一报价
- 8: “44436660”,成交的股票数,以股为单位
- 9: “1235687451. 840”,成交金额,以元为单位
- 10: “17400”,买一申报 17400 股,即 174 手
- 11: “28. 090”,买一报价
- 12: “164500”,买二申报股数
- 13: “28. 080”,买二报价
- 14: “138300”,买三申报股数
- 15: “28. 070”,买三报价
- 16: “97500”,买四申报股数
- 17: “28. 060”,买四报价
- 18: “7000”,买五申报股数
- 19: “28. 050”,买五报价
- 20: “43529”,卖一申报 43529 股,即 435 手
- 21: “28. 100”,卖一报价
- 22: “12900”,卖二申报股数
- 23: “28. 110”,卖二报价
- 24: “22300”,卖三申报股数
- 25: “28. 120”,卖三报价
- 26: “36359”,卖四申报股数
- 27: “28. 130”,卖四报价
- 28: “35300”,卖五申报股数
- 29: “28. 140”,卖五报价
- 30: “2018-06-22”,日期
- 31: “15:05:03”,时间

若要查询多个股票,在 URL 最后加上一个逗号,再加上股票代码即可。例如要一次查

询万科 A(sz000002)和平安银行(sz000001)的行情,构造的 URL 如下: <http://hq.sinajs.cn/list=sz000002,sz000001>。另外,上证指数的代码是 sh000001,深证成指的代码是 sz399001;如果只需获取股票交易的基本数据则在之前加“s_”,即 http://hq.sinajs.cn/list=s_sz000002。

2. K 线图接口

对于股票的 K 线图、日线图等的获取可以通过请求以下 URL 获得: [http://image.sinajs.cn/\(3\)/n/\(1\)\(2\).gif](http://image.sinajs.cn/(3)/n/(1)(2).gif)。其中(1)(2)参数同上所述,为股票所在市场和股票代码;参数(3)为查看周期,具体可取值有 min、daily、weekly、monthly 等。示例如下:

(1) 查看分时线(<http://image.sinajs.cn/newchart/min/n/sz000002.gif>),结果如图 3-3 所示。

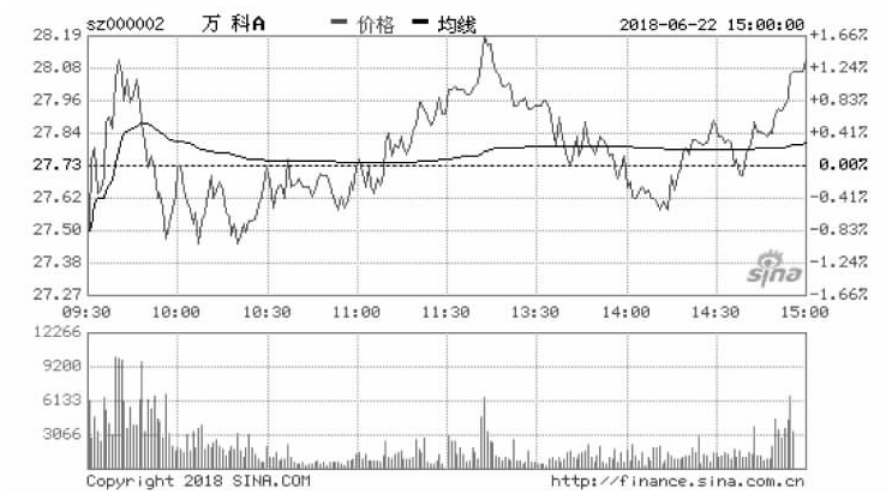


图 3-3 获取分时线

(2) 查看日 K 线(<http://image.sinajs.cn/newchart/daily/n/sz000002.gif>),结果如图 3-4 所示。

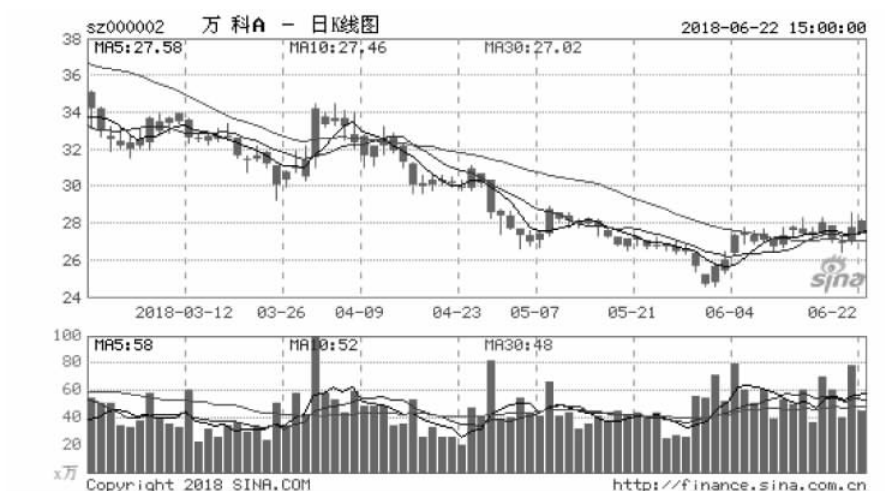


图 3-4 获取日 K 线

(3) 查看周 K 线(<http://image.sinajs.cn/newchart/weekly/n/sz000002.gif>), 结果如图 3-5 所示。



图 3-5 获取周 K 线

(4) 查看月 K 线(<http://image.sinajs.cn/newchart/monthly/n/sz000002.gif>), 结果如图 3-6 所示。

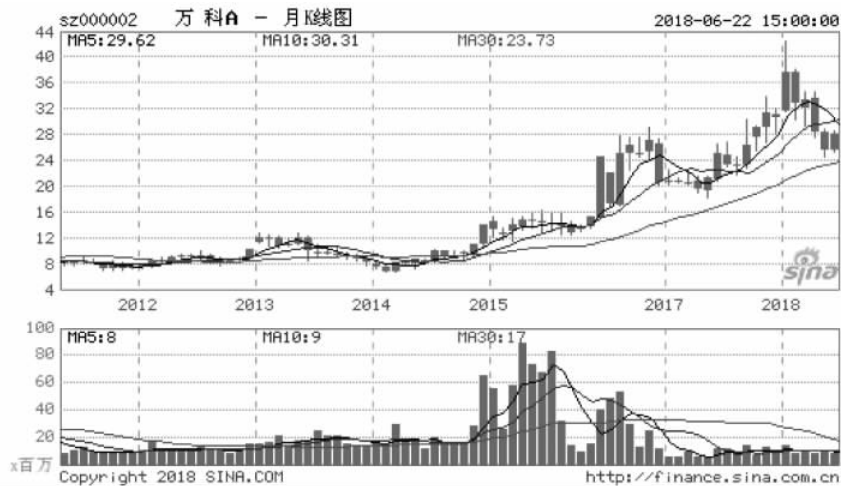


图 3-6 获取月 K 线

3. 历史交易数据的抓取

从新浪抓取股票历史数据的 URL 为:

[http://money.finance.sina.com.cn/corp/go.php/vMS_MarketHistory/stockid/\(1\).phtml?year=\(2\)&jidu=\(3\)](http://money.finance.sina.com.cn/corp/go.php/vMS_MarketHistory/stockid/(1).phtml?year=(2)&jidu=(3))

构造 URL 需要提供三个参数: 第一个参数是股票代码; 第二个参数是年份, 目前的可取范围为 1991 到 2018; 第三个参数是季度, 取值范围为 1~4。以万科为例, 2018 年第 2 季度的历史数据获取的 URL 构造如下:

http://money.finance.sina.com.cn/corp/go.php/vMS_MarketHistory/stockid/000002.phtml?year=2018&jidu=2

访问该 URL,其返回结果如图 3-7 所示。返回的表格包括七个字段,分别为日期、开盘价、最高价、收盘价、最低价、交易量(股)、交易金额(元)。接下来可以通过爬虫技术和 HTML 解析技术获得结构化的数据。通过构造不同时期的对应 URL 并通过爬虫循环抓取,即可获得所有的历史数据。

2018 ▾ 二季度 ▾ 查询

万科A(000002)年季度历史交易

日期	开盘价	最高价	收盘价	最低价	交易量(股)	交易金额(元)
2018-06-22	27.500	28.200	28.100	27.440	44436660	1235687452
2018-06-21	26.970	28.500	27.730	26.820	77793209	2167319528
2018-06-20	26.990	27.290	26.960	26.410	39340042	1059357119
2018-06-19	27.850	27.850	27.100	26.910	60044211	1639680596
2018-06-15	27.370	28.300	27.990	27.330	69819518	1951841871
2018-06-14	27.480	27.760	27.370	27.120	35760445	980418256
2018-06-13	27.410	28.250	27.640	27.310	60284357	1680583497
2018-06-12	27.570	27.830	27.710	27.230	49637248	1376048984
2018-06-11	26.740	27.730	27.310	26.680	52643679	1437705891
2018-06-08	27.100	27.180	26.710	26.550	38854508	1040501593
2018-06-07	27.070	27.640	27.390	26.950	59753575	1637918296
2018-06-06	27.300	27.540	26.990	26.850	50816253	1378427339
2018-06-05	27.290	27.770	27.490	26.840	60684137	1658473532
2018-06-04	26.300	27.380	27.300	26.280	78919077	2138820837
2018-06-01	25.390	26.460	26.000	25.300	52077082	1355014454
2018-05-31	24.770	25.680	25.590	24.580	70589325	1768260903
2018-05-30	25.210	25.210	24.640	24.580	54016882	1342440709
2018-05-29	26.300	26.370	25.600	25.340	55605235	1428341707
2018-05-28	26.390	26.820	26.500	26.300	25459499	676789913
2018-05-25	26.680	26.910	26.390	26.320	27000546	717574351
2018-05-24	26.800	26.850	26.650	26.520	24161093	643575798
2018-05-23	26.720	27.170	26.800	26.580	43317062	1164157878
2018-05-22	27.050	27.050	26.720	26.580	41546903	1111277324

图 3-7 万科 2018 年第二季度历史数据(部分)

3.2.4 从网易获取成交明细数据

从网易财经频道可以方便地获取股票的日每笔成交明细数据。以万科为例,访问以下网址: http://quotes.money.163.com/trade/cjmx_000002.html#01b05,如图 3-8 所示,单击“历史数据下载”则可直接下载个股近 5 个交易日的每笔成交明细数据。

下载的数据为 Excel 格式,内容如图 3-9 所示。表格中包括六个字段的数据,分别为成交时间、成交价、价格变动、成交量(手)、成交额(元)、性质。

为了让机器自动收集所有个股的成交明细数据,可用构造链接参数的方式进行批量下载。仍以万科为例,通过分析可知其实际的下载 URL 地址是:

<http://quotes.money.163.com/cjmx/2018/20180622/1000002.xls>



图 3-8 网易财经频道上个股的成交明细数据页面

	A1	A	B	C	D	E	F
		成交时间	成交价	价格变动	成交量(手)	成交额(元)	性质
1		成交时间	成交价	价格变动	成交量(手)	成交额(元)	性质
2		9:25:03	27.5	-0.23	4792	13177532.5	买盘
3		9:30:00	27.55	0.05	18	49586	买盘
4		9:30:03	27.52	-0.03	3289	9043984.85	中性
5		9:30:06	27.52	0	305	839516	卖盘
6		9:30:09	27.5	-0.02	354	975038.56	卖盘
7		9:30:12	27.46	-0.04	179	490928.44	卖盘
8		9:30:15	27.45	-0.01	67	183366.1	卖盘
9		9:30:18	27.45	0	206	566060.45	卖盘
10		9:30:21	27.45	0	116	319515.78	卖盘
11		9:30:24	27.5	0.05	414	1138638.15	买盘
12		9:30:30	27.5	0	161	442747	卖盘
13		9:30:33	27.5	0	68	187902.18	买盘
14		9:30:36	27.48	-0.02	78	214930.64	卖盘
15		9:30:39	27.5	0.02	29	78488.67	买盘
16		9:30:42	27.48	-0.02	52	142926	中性
17		9:30:45	27.47	-0.01	109	299875.11	卖盘
18		9:30:48	27.5	0.03	152	417818	买盘
19		9:30:51	27.47	-0.03	193	530210.65	卖盘
20		9:30:54	27.49	0.02	51	141249.27	买盘
21		9:30:57	27.49	0	68	186889	买盘

图 3-9 从网易下载的个股成交明细数据

该 URL 中有三个部分的参数：年份、交易日期、股票所在市场和股票代码。股票所在市场的参数用“0”代表沪市，“1”代表深市。交易日期参数被限定在近两周的交易日。

图 3-8 所示的个股成交明细页面的下方还给出了个股的成交明细表,通过解析该网页的表格也可以获得结构化的个股成交明细数据。

3.2.5 从巨潮资讯获取基础数据

巨潮资讯是中国证监会指定的信息披露网站,是数据提供商。巨潮资讯网提供了数据下载器的功能,可以一站式下载权威行情、财务数据等。在 PC 端访问巨潮资讯网(网址: <http://www.cninfo.com.cn>),打开的界面如图 3-10 所示。



图 3-10 巨潮资讯网的数据下载器

1. 行情数据下载

巨潮资讯网提供深沪两市挂牌的股票、基金、债券、指数、回购等交易品种上市以来所有交易日的收盘行情数据,内容包括前收、开盘、最高、最低、收盘、成交数量、成交笔、成交金额、涨跌幅等字段内容。如图 3-11 所示,个股历史收盘行情的下载步骤为:①选择一只股票;②选择起止年度;③单击“下载”按钮。



图 3-11 历史收盘行情下载器

下载结果是每年单独一个 Excel 文件,打开后其内容如图 3-12 所示。

(a)

(b)

#	A	B	C	D	E	F	G	H	I	J	K	L	M
1	证券代码	证券简称	交易日期	交易所	昨日收盘	今日开盘	成交数量	最高成交	最低成交	最近成交	总手数	涨跌幅	成交金额
2	000002	万科A	2018-1-2	深交所	31.06	31.45	68343350	32.99	31.45	32.56	63030	4.8294	2.22E+09
3	000002	万科A	2018-1-3	深交所	32.56	32.5	64687020	33.78	32.23	32.33	67257	-0.7064	2.13E+09
4	000002	万科A	2018-1-4	深交所	32.33	32.76	52908580	33.53	32.1	33.12	48689	2.4436	1.74E+09
5	000002	万科A	2018-1-5	深交所	33.12	32.98	84310196	35.88	32.8	34.76	83181	4.9517	2.92E+09
6	000002	万科A	2018-1-8	深交所	34.76	35.11	83078359	36.96	35.11	35.99	79471	3.5386	2.95E+09
7	000002	万科A	2018-1-9	深交所	35.99	35.63	47845909	36.11	34.95	35.84	56576	-0.4168	1.7E+09
8	000002	万科A	2018-1-10	深交所	35.84	35.6	45369991	36.7	35.5	35.67	47719	-0.4743	1.63E+09
9	000002	万科A	2018-1-11	深交所	35.67	35.55	40248381	35.78	34.54	35.15	45765	-1.4579	1.41E+09
10	000002	万科A	2018-1-12	深交所	35.15	35.1	53701333	36.62	35.09	35.55	54606	1.138	1.93E+09
11	000002	万科A	2018-1-15	深交所	35.55	35.84	75718609	38	35.61	37.09	72188	4.3319	2.78E+09
12	000002	万科A	2018-1-16	深交所	37.09	36.5	97846478	40.7	36.4	40.2	84078	8.385	3.76E+09
13	000002	万科A	2018-1-17	深交所	40.2	39.48	81219193	40.6	38.3	39.77	92607	-1.0697	3.2E+09
14	000002	万科A	2018-1-18	深交所	39.77	39.6	55884840	40.64	38.9	39.76	60141	-0.0251	2.23E+09
15	000002	万科A	2018-1-19	深交所	39.76	39.85	54875823	41.09	39.25	39.42	64023	-0.8551	2.21E+09
16	000002	万科A	2018-1-22	深交所	39.42	39	40652641	39.84	38.51	39.53	49336	0.279	1.6E+09
17	000002	万科A	2018-1-23	深交所	39.53	40	64865750	41.51	39.99	40.98	75731	3.6681	2.65E+09
18	000002	万科A	2018-1-24	深交所	40.98	40.7	54262959	42.24	40.46	41.13	66371	0.366	2.24E+09
19	000002	万科A	2018-1-25	深交所	41.13	41.12	64806457	41.12	38.8	40.2	79048	-2.2611	2.58E+09
20	000002	万科A	2018-1-26	深交所	40.2	40	41006237	40.66	39.12	40.04	53993	-0.398	1.64E+09

图 3-12 行情数据下载结果

(a) 数据文件列表; (b) 文件内容

2. 财务数据下载

巨潮资讯网提供上市公司最早从 1997 年以来三大财务报表(利润分配表、资产负债表、现金流量表)接近 230 个财务科目的内容。如图 3-13 所示,个股财务数据的下载步骤为:

①选择一只股票;②选择财务报表类型;③选择起止年度;④单击“下载”按钮。

A股 **港股**

个股财务数据 **历史收盘行情**

代码: 000002 类别: 利润表

起: 2016 止: 2018

更多 > 下载 免责声明

图 3-13 个股财务数据下载器

下载结果是每年单独一个 Excel 文件,打开后其内容如图 3-14 所示。

(a)

(b)

#	A	B	C	D	E	F	G	H	I	J	K
1	机构ID	机构名称	公告日期	开始日期	截止日期	报告年度	合并类型	报表来源	一、营业总收入	其中:营业收入	二、营业总成本
2	000002	万科A	2017-4-28	2017-1-1	2017-3-31	2017-3-31	合并本期	定期报告	18589228819	18589228819	13048871983
3	000002	万科A	2017-8-25	2017-1-1	2017-6-30	2017-6-30	合并本期	定期报告	69810477567	69810477567	47015024669
4	000002	万科A	2017-10-27	2017-1-1	2017-9-30	2017-9-30	合并本期	定期报告	1.17101E+11	1.17101E+11	79943369965
5	000002	万科A	2018-3-27	2017-1-1	2017-12-31	2017-12-31	合并本期	定期报告	2.42897E+11	2.42897E+11	1.6008E+11

图 3-14 财务数据下载结果

(a) 数据文件列表; (b) 文件内容

3. 基础数据下载

在巨潮资讯网访问一只个股,可以看到该个股详细的基础数据信息。如图 3-15 所示,

基础数据信息包括了 14 个方面,分别为最新资料、公司概况、发行筹资、分红配股、高管人员、股本结构、十大股东、财务指标、公告摘要、公告全文、定期报告、投资者关系信息、章程制度、持续督导意见。

巨潮资讯网的信息页面相对来说已经较接近结构化信息,主要是以表格和列表为主。通过网络爬虫技术及内容解析技术,可以全面获取个股所有相关的数据信息。

最新资料	公司概况	发行筹资	分红配股	高管人员	股本结构	十大股东	财务指标	公告摘要	公告全文	定期报告	投资者关系信息	章程制度	持续督导意见	
股票代码: 000002 股票简称: 万科A 代码/简称/拼音 选择股票														
十大股东						十大股东	流通股东							
截止时间	股东名称					持股数量(股)	持股比例(%)	股份性质						
20180331	1.深圳市地铁集团有限公司					3,242,810,791	29.380	流通A股						
	2.HKSCC NOMINEES LIMITED					1,314,913,539	11.910	境外可流通股						
	3.深圳市钜盛华股份有限公司					926,070,472	8.390	流通A股						
	4.国信证券-工商银行-国信金鼎分级1号集合资产管理计划					456,993,190	4.140	流通A股						
	5.前海人寿保险股份有限公司-海利年年					349,776,441	3.170	流通A股						
	6.招商财富-招商银行-博赢1号专项资产管理计划					329,352,920	2.980	流通A股						
	7.安邦财产保险股份有限公司-传统产品					258,167,403	2.340	流通A股						
	8.安邦人寿保险股份有限公司-保守型投资组合					243,677,851	2.210	流通A股						
	9.西部利得基金-建设银行-西部利得金裕1号资产管理计划					225,494,379	2.040	流通A股						
	10.前海人寿保险股份有限公司-聚富产品					218,081,383	1.980	流通A股						
最新公告														
万科A: 关于2017年度股东大会的提...														
万科A: 关于向合格投资者公开发行...														
万科A: 关于变更证券事务代表的公...														
万科A: 2018年度第一期超短期融资...														
万科A: 关于按照《香港上市规则》...														
万科A: 2018年五月份销售及近期新...														
万科A: 根据70亿美元中期票据计划...														
万科A: 关于“15万科01”、“17万...														
万科A: 关于万科物业为物业香港发...														
万科A: 2017年公司债券(第二期)...														
公司备忘														
备忘事件 备忘时间														
股东大会召开日 20180629														
网络投票结束日 20180629														
网络投票开始日 20180628														
股东资格登记日 20180621														

图 3-15 个股的相关基础数据信息

3.2.6 多源获取 UGC 数据

UGC 数据,即用户产生的内容。在股票领域,为分类方便,本书暂且把新闻资讯、股吧论坛、机构研报、用户自选股、用户单击等都算是 UGC 数据。这些 UGC 数据的来源主要是一些门户网站、社区网站等,可通过网络爬虫技术进行获取。由于本书后续有专门章节讨论相应的 UGC 数据并进行深度挖掘分析,这里先不做介绍,详见相应的章节。

3.3 数据获取技术

为了从数据源获取股票相关数据,接下来需要采用网络爬虫技术对数据进行爬取。在爬取网址或调用接口的过程中,其返回的数据主要体现为三种格式: HTML、XML 和 JSON,这时需要应用相应的解析方法将返回数据解析成结构化信息。

3.3.1 网络爬虫技术

1. 爬虫的基本原理

网络爬虫也叫网络蜘蛛,它是一个按照一定的规则自动提取网页的程序。根据爬虫爬取的范围,可以将其分为两类: 通用爬虫和主题网络爬虫。

通用爬虫是从一个或多个初始网页的 URL(种子 URL)开始,获得初始网页上的 URL 列表;在爬取网页的同时,不断从当前网页提取新的 URL 放入待爬行队列中,直到满足程

序的停止条件。一个通用的网络爬虫的框架如图 3-16 所示,其基本工作流程如下:

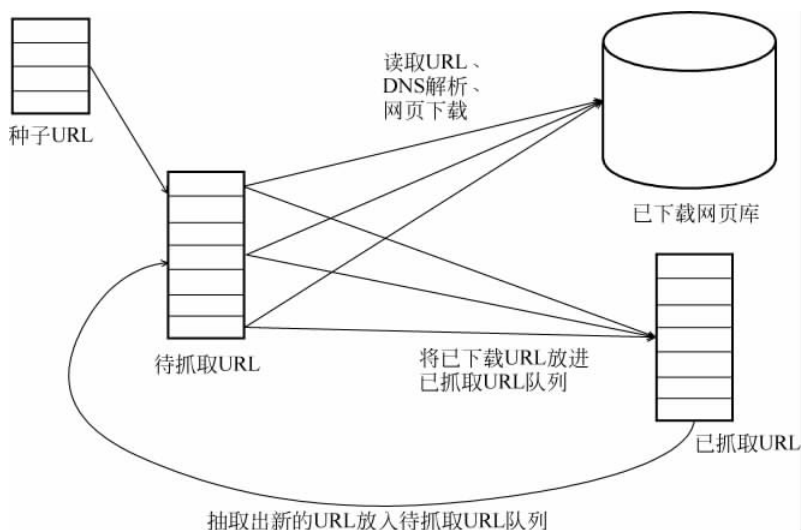


图 3-16 通用的网络爬虫框架

- (1) 选取一部分精心挑选的种子 URL。
- (2) 将这些 URL 放入待爬取 URL 队列。
- (3) 从待爬取 URL 队列中取出 URL,解析 DNS,得到主机的 IP,并将 URL 对应的网页下载下来,存储进已下载网页库中。此外,将这些 URL 放进已爬取 URL 队列。
- (4) 分析已爬取 URL 队列中的 URL,并且将满足条件的 URL 放入待爬取 URL 队列,从而进入下一个循环。

主题网络爬虫根据一定的网页分析算法过滤与主题无关的链接,保留与主题相关的链接并将其放入待爬取的 URL 队列中,通过一定的搜索策略从队列中选择下一步要爬取的 URL,重复以上步骤直到满足程序的停止条件。所有被网络爬虫爬取的网页将会被系统存储,进行一定的分析、过滤,并建立索引。对于主题网络爬虫来说,这一过程所得到的分析结果还可能对后续的爬取过程进行反馈和指导。现有主题网络爬虫对爬取目标的描述可分为基于目标网页特征、基于目标数据模式和基于领域概念三种。本书对目标数据的爬取属于主题网络爬虫的方式。

2. 爬取策略

在爬虫系统中,待爬取 URL 队列是很重要的一部分。待爬取 URL 队列中的 URL 以什么样的顺序排列也是一个很重要的问题,因为这涉及先爬取哪个页面,后爬取哪个页面。而决定这些 URL 排列顺序的方法,叫做爬取策略。下面重点介绍常见的爬取策略。

1) 深度优先搜索策略

深度优先搜索的目的是要达到被搜索结构的叶节点(即那些不包含任何超链的 HTML 文件)。例如在一个 HTML 文件中,当一个超链被选择后,将对被链接的 HTML 文件执行深度优先搜索,也就是说,在搜索其余的超链结果之前必须先完整地搜索单独的一条链。深度优先搜索沿着 HTML 文件上的超链走到不能再深入为止,然后返回到某一个 HTML 文件,再继续选择该 HTML 文件中的其他超链。当不再有其他超链可选择时,说明搜索已经结束。

2) 广度优先搜索策略

广度优先搜索策略是指在爬取过程中,在完成当前层次的搜索后,才进行下一层次的搜索。为覆盖尽可能多的网页,一般使用广度优先搜索方法。也有很多研究将广度优先搜索策略应用于聚焦爬虫中,其基本思想是认为与初始 URL 在一定链接距离内的网页具有主题相关性的概率很大。另外一种方法是将广度优先搜索与网页过滤技术结合使用,先用广度优先策略爬取网页,再将其中无关的网页过滤掉。这些方法的缺点在于,随着爬取网页的增多,大量的无关网页将被下载并过滤,这将导致算法的效率变低。

3) 最佳优先搜索策略

最佳优先搜索策略按照一定的网页分析算法,预测候选 URL 与目标网页的相似度或与主题的相关性,并选取评价最好的一个或几个 URL 进行爬取。它只访问经过网页分析算法预测为“有用”的网页。存在的一个问题是,在爬虫爬取路径上的很多相关网页可能被忽略。由于最佳优先策略是一种局部最优搜索算法,因此需要将最佳优先结合具体的应用进行改进,以跳出局部最优点。

3. 更新策略

互联网是实时变化的,具有很强的动态性。网页更新策略主要是决定何时更新之前已经下载过的页面。常见的更新策略有以下三种。

1) 历史参考策略

顾名思义即根据页面以往的历史更新数据,预测该页面未来何时会发生变化。

2) 用户体验策略

用户往往只关注前几页结果,因此爬取系统可以优先更新那些显示在查询结果前几页中的网页,然后再更新后面的网页。这种更新策略也需要用到历史信息。用户体验策略保留网页的多个历史版本,并且根据过去每次内容变化对搜索质量的影响,得出一个平均值,用这个值作为决定何时重新爬取的依据。

3) 聚类抽样策略

前面两种更新策略都有一个前提:需要网页的历史信息。这样就存在两个问题:①系统需要为每个系统保存多个版本的历史信息,无疑增加了很多负担;②如果新的网页完全没有历史信息,就无法确定更新策略。

聚类抽样策略认为,网页具有很多属性,类似属性的网页可以认为其更新频率也是类似的。要计算某一个类别网页的更新频率,只需要对这一类网页抽样,以它们的更新周期作为整个类别的更新周期。基本思路如图 3-17 所示。

4. 分布式策略

一般来说,爬取系统需要面对的是整个互联网上数以亿计的网页,单个爬取程序不可能完成这样的任务,往往需要多个爬取程序一起来处理。爬取系统往往是一个分布式的三层结构,如图 3-18 所示,最下一层是分布在不同地理位置的数据中心,在每个数据中心里面有若干台爬取服务器,而每台爬取服务器上可能部署了若干套爬虫程序。这就构成了一个基本的分布式爬取系统。

对于一个数据中心内的不同爬取服务器,协同工作的方式有以下几种。

1) 主从式

主从式(Master-Slave)的基本结构如图 3-19 所示。对于主从式模式而言,有一台专门

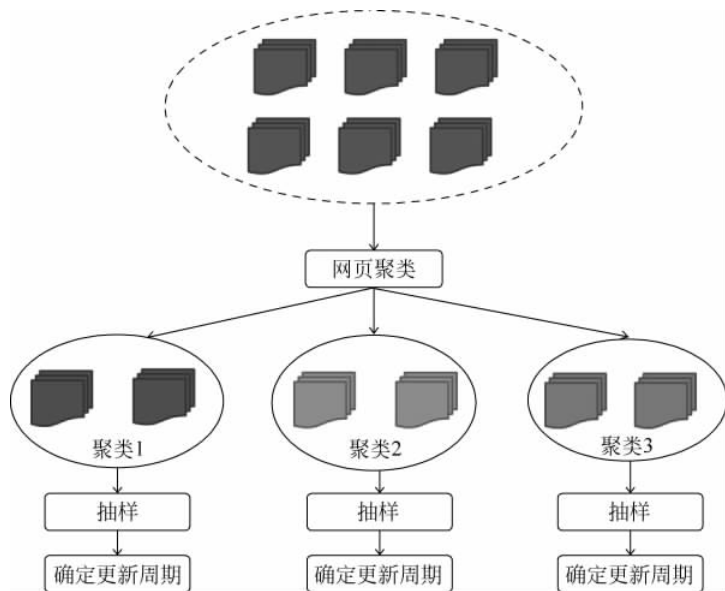


图 3-17 聚类抽样更新策略

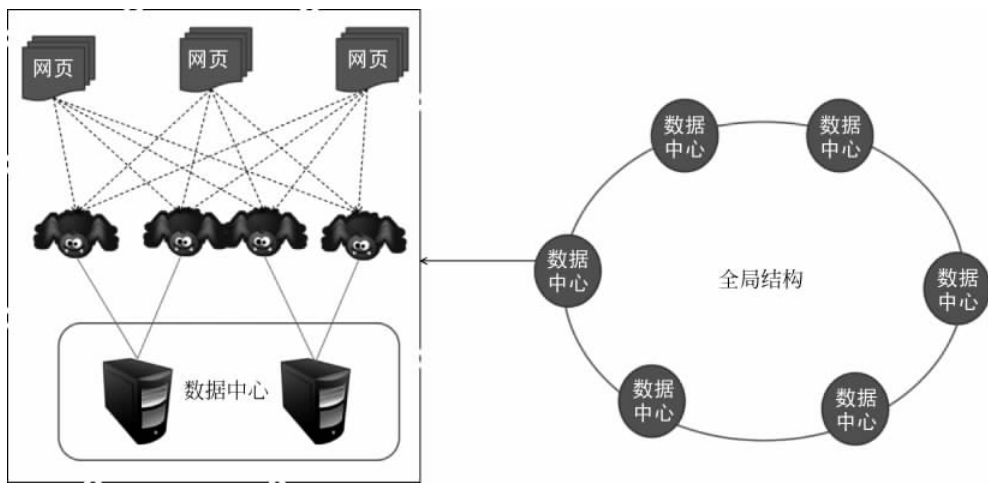


图 3-18 分布式爬取系统的架构

的 Master 服务器用于维护待爬取 URL 队列,它负责每次将 URL 分发到不同的 Slave 服务器,而 Slave 服务器则负责实际的网页下载工作。Master 服务器除了维护待爬取 URL 队列以及分发 URL 之外,还要负责协调各个 Slave 服务器的负载情况,以免某些 Slave 服务器过于清闲或者过于劳累。这种模式下,Master 往往容易成为系统瓶颈。

2) 对等式

对等式(Peer to Peer)的基本结构如图 3-20 所示。在这种模式下,所有的爬取服务器在分工上没有不同,每一台爬取服务器都可以从待爬取 URL 队列中获取 URL。首先对该 URL 的主域名进行哈希(Hash)值计算得到 H ,然后计算 $H \bmod m$ (其中 m 是服务器的数量),最后计算得到的数就是处理该 URL 的主机编号。举例来说,假设对于 URL `www.`

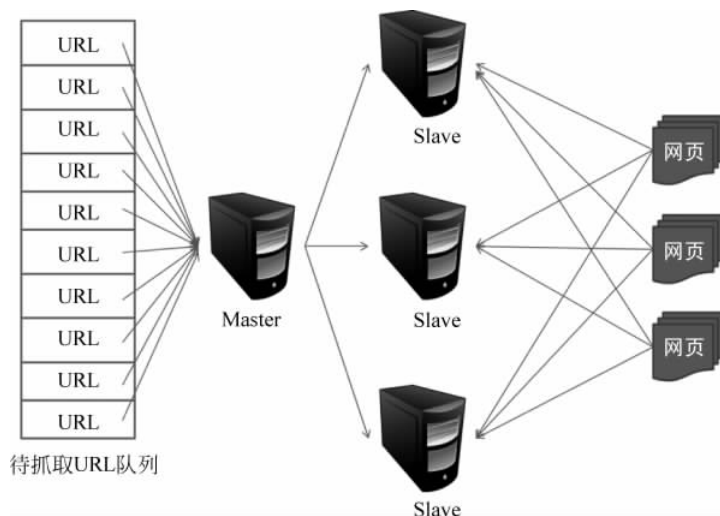


图 3-19 主从式结构

stocktobe.com, 计算其哈希值 $H=8$, $m=3$, 则 $H \bmod m=2$, 因此由编号为 2 的服务器进行该链接的爬取。这种模式有一个问题, 当有一台服务器死机或添加新的服务器, 那么所有 URL 的哈希求余的结果就都要变化。也就是说, 这种方式的扩展性不佳。

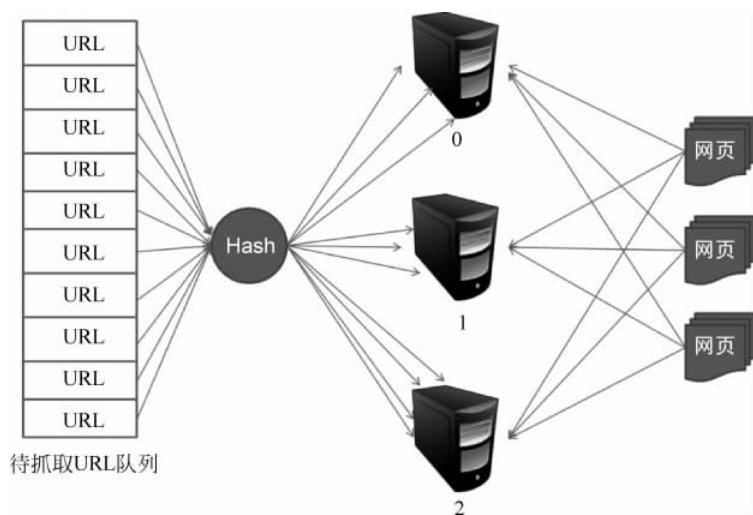


图 3-20 对等式结构

3) 一致性哈希结构

针对对等式结构的情况, 提出了一种改进方案, 这种改进的方案是用一致性哈希法确定服务器的分工。其基本结构如图 3-21 所示。一致性哈希法将 URL 的主域名进行哈希运算, 映射为一个范围在 $0 \sim 2^{32}$ 之间的某个数。而将这个范围平均地分配给 m 台服务器, 根据 URL 主域名哈希运算的值所处的范围判断由哪台服务器进行爬取。如果某一台服务器出现问题, 那么本该由该服务器负责的网页则按照顺时针顺延, 由下一台服务器进行爬取。这样, 即使某台服务器出现问题, 也不会影响其他服务器的工作。

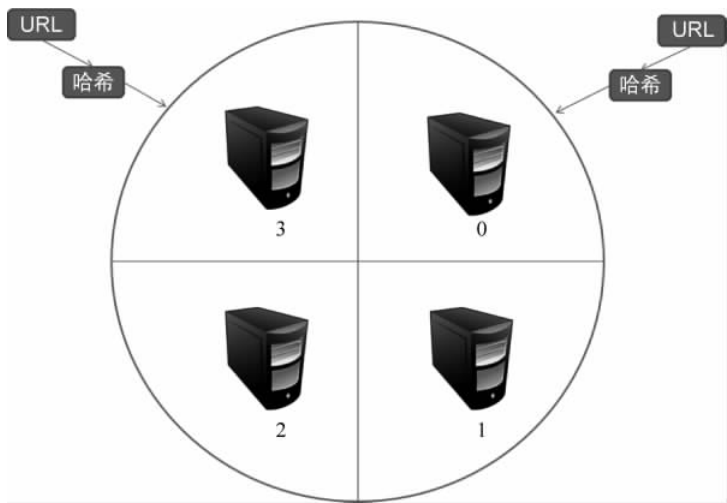


图 3-21 一致性哈希结构

5. 爬虫开源框架

目前市场上的开源爬虫框架很多,不同语言下有不同类型的爬虫框架。常用语言的爬虫框架列表如表 3-5 所示。

表 3-5 常用语言的爬虫框架

Java	Python	PHP	C #	C/C++
Apache Nutch2	Scrapy	PHPSpider	DotnetSpider	Open-source-search-engine
WebMagic	Crawley	BeanBun	NWebCrawler	Cobweb
Heritrix	Portia	PHPCrawl	SmartSpider	Upton
WebCollector	PySpider	PHP Selenium	Abot	Wombat
Crawler4j	Grab		xNet	Spider
SpiderMan	Cola		AngleSharp	Larbin
SeimiCrawler	Python Selenium		HtmlAgilityPack	
Jsoup			CSQuery	
Java Selenium				
HtmlUnit				

下面对一些流行爬虫框架做简要介绍。

1) Nutch

Nutch 是 Apache 的子项目之一,是一个基于 Lucene,类似 Google 的完整网络搜索引擎解决方案,其基于 Hadoop 的分布式处理模型保证了系统的性能,类似 Eclipse 的插件机制保证了系统的可定制化。Nutch 包括全文搜索和 Web 爬虫,支持分布式,可以通过配置网站地址、规则,以及采集的深度对网站进行采集。

网址: <http://nutch.apache.org>。

2) WebMagic

WebMagic 是一个开源的 Java 垂直爬虫框架,目标是简化爬虫的开发流程,让开发者

专注于逻辑功能的开发。WebMagic 采用完全模块化的设计,功能覆盖整个爬虫的生命周期(链接提取、页面下载、内容抽取、持久化),支持多线程爬取,分布式爬取,并支持自动重试、自定义 UA/cookie 等功能。

网址: <http://webmagic.io>。

3) Heritrix

Heritrix 是 SourceForge 上的开源产品,是一个由 Java 开发的、开源的网络爬虫,用户可以使用它来从网上爬取想要的资源。其最出色之处在于它良好的可扩展性,方便用户实现自己的爬取逻辑。Heritrix 是个“archival crawler”,可用来获取完整的、精确的站点内容的深度复制。

网址: <https://sourceforge.net/projects/archive-crawler/>。

4) Jsoup

Jsoup 是一款 Java 的 HTML 解析器,可直接解析某个 URL 地址和 HTML 文本内容。它提供了一套非常省力的 API,可通过 DOM、CSS 以及类似于 jQuery 的操作方法来取出和操作数据。

网址: <https://jsoup.org>。

5) Scrapy

Scrapy 是 Python 开发的一个快速、高层次的屏幕爬取和 Web 爬取框架,用于爬取 Web 站点并从页面中提取结构化的数据。Scrapy 用途广泛,可以用于数据挖掘、监测和自动化测试。Scrapy 是一个框架,任何人都可以根据需求方便地修改,它也提供了多种类型爬虫的基类,如 BaseSpider、Sitemap 爬虫等。

网址: <https://scrapy.org>。

6) PySpider

PySpider 是一个由国人编写的网络爬虫系统,并带有强大的 WebUI。它采用 Python 语言编写,为分布式架构,支持多种数据库后端,WebUI 支持脚本编辑器、任务监视器、项目管理器以及结果查看器。

网址: <http://www.pyspider.cn>。

7) PHPSpider

PHPSpider 是用 PHP 开发的一款优秀的爬虫开发框架。基于该框架,不用了解爬虫的底层技术实现,只用简单几行 PHP 代码,就可以创建自己的爬虫。利用框架封装的多进程 Worker 类库,代码更简洁,执行效率更高。

网址: <https://github.com/owner888/phpspider>。

8) PHPCrawl

PHPCrawl 是一个 PHP 开源的 Web 检索爬虫类库。PHPCrawl 可以选择性地指定爬虫的行为,例如喜欢的网址、内容类型、过滤器、cookie 的处理等方式。

网址: <http://phpcrawl.cuab.de>。

9) DotnetSpider

DotnetSpider 是由国人编写的、开源的一个跨平台、高性能、轻量级的爬虫软件,采用 C# 开发。目前是 .NET 开源爬虫最为优秀的爬虫之一。

网址: <http://www.dotnetspider.com>。

10) xNet

xNet 是一个由俄罗斯人写的开源工具,该工具代码不一样的地方在于将所有 HTTP 协议的底层都重新实现了一遍。这样做的好处是在做爬虫开发调试时,可以进入到更深入的底层组件,如底层的字节流。

网址: <https://github.com/X-rus/xNet>。

3.3.2 HTML 解析

HTML 是一种标记语言,是 Hypertext Markup Language 的缩写,即超文本标记语言。HTML 是标准通用标记语言(SGML)下的一个应用,也是一种规范或标准,它通过标记符号来标记要显示的网页中的各个部分。网页文件本身是一种文本文件,通过在文本文件中添加标记符,可以告诉浏览器如何显示其中的内容,例如文字如何处理,画面如何安排,图片如何显示,等等。

对 HTML 数据的解析有多种方式,如正则表达式、字符串截取等。这里介绍一种开源的工具——Htmlparser。顾名思义,Htmlparser 是一种解析提取 HTML 的工具。Htmlparser 具有小巧、快速的优点,缺点是相关文档比较少。使用 Htmlparser 可以很好地对 HTML 标签进行操作,如:

- 文本信息抽取,例如对 HTML 进行有效信息搜索;
- 链接提取,用于自动给页面的链接文本加上链接的标签;
- 资源提取,例如对一些图片、声音资源的处理。

使用 Htmlparser 将 HTML 字符串生成 DOM 树就可以对 DOM 树进行操作,这样也就不再担心网站改版。Htmlparser 生成 DOM 树的过程如下:

- (1) 新建一个 XmlDocument 对象,并添加根节点 root;
- (2) 根据获取的 HTML 字符串生成一个 Parser 对象,同时也生成 NodeFilter 对象 Filter;
- (3) 根据 Parser 对象的 Parser 方法获取需要过滤的节点集合 NodeList;
- (4) 循环 NodeList,使用递归方式向 XML 添加 DOM 节点。

Htmlparser 的下载地址是 <http://htmlparser.sourceforge.net>。

3.3.3 XML 解析

XML,英文全称 Extensible Markup Language,即可扩展标记语言。XML 是标准通用标记语言(SGML)的子集,是一种用于标记电子文件使其具有结构性的标记语言。XML 可以用来标记数据、定义数据类型,是一种允许用户对自己的标记语言进行定义的源语言。XML 非常适合万维网传输,其提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。

1. XML 数据格式

XML 文档中的元素形成了一棵文档树。如以下示例所示,这棵树从根部开始,扩展到树的最底端,所有的元素都可以有子元素。用父、子、同胞等术语描述元素之间的关系:父元素拥有子元素,相同层级上的子元素称为同胞。所有的元素都可以有文本内容和属性。

```
<书架>
  <书>
    <书名>Java</书名>
    <作者>张三</作者>
    <售价>39.00 元</售价>
  </书>
  <书>
    <书名>JavaScript</书名>
    <作者>李四</作者>
    <售价>28.00 元</售价>
  </书>
</书架>
```

2. XML 数据解析

SUN 提供了两种不同的解析 XML 数据的方式：DOM 解析、SAX 解析。

DOM 的解析原理是在内存中构造 XML 文件的 DOM 树。因此使用 DOM 解析 XML 数据时，解析器会将整个 XML 文件直接装载进内存形成 DOM 树，如果该 XML 文件过大，可能会导致内存的溢出或者操作比较慢。

SAX 解析使用事件机制和方法回调的技术实现。SAX 解析是为了解决 DOM 解析的缺点，因此 SUN 在设计时借鉴了 GUI 的事件处理机制。SAX 在解析 XML 文件时效率要高于 DOM 解析，但 SAX 解析器只能做元素的获取，不能进行其他的操作。

开源组织免费开发了一个 DOM4J 框架技术，可以使得开发者快速并高效地解析 XML 数据。DOM4J 该解析技术结合了 DOM 和 SAX 的优点，因此在实际的项目开发中会使用该解析技术进行解析。DOM4J 的下载地址是：<https://sourceforge.net/projects/dom4j/>。

使用 DOM4J 进行开发时，每次都需要从根元素开始查找元素的位置，因此，希望有一个类似操作系统路径的方式可以直接描述 XML 中各个节点的关系，以方便 DOM4J 进行元素的查找。此时 DOM4J+XPath 的技术应运而生。

XPath 是一种用来确定 XML 文档中某部分位置的语言。XPath 基于 XML 的树状结构，提供在数据结构树中找寻节点的能力。XPath 使用路径表达式选取 XML 文档中的节点或者节点集。路径表达式是从一个 XML 节点（当前的上下文节点）到另一个节点或一组节点的书面步骤顺序。这些步骤以“/”字符分开，每一步有三个构成成分：①轴描述（用最直接的方式接近目标节点）；②节点测试（用于筛选节点位置和名称）；③节点描述（用于筛选节点的属性和子节点特征）。操作 XPath 的几种常用方法如下：

- 获取多个元素：selectNodes(String xpath)
- 获取单个元素：selectSingleNode(String xpath)
- 获取属性：valueOf(String xpath)

3.3.4 JSON 解析

JSON(JavaScript Object Notation)是一种轻量级的数据交换格式，具有良好的可读性和便于快速编写的特性，可在不同平台之间进行数据交换。JSON 采用兼容性很高的、完全独立于语言的文本格式，是一种理想的数据交换语言。

1. JSON 数据格式

和 XML 一样，JSON 也是基于纯文本的数据格式。JSON 的数据格式非常简单，可以

用 JSON 传输一个简单的 String、Number、Boolean,也可以传输一个数组,或者一个复杂的 Object 对象。JSON 表示一个数组对象,使用[]包含所有元素,每个元素用逗号分隔,元素可以是任意的 Value,例如以下数组包含了一个 String、Number、Boolean 和一个 null:

```
["abc",12345,false,null]
```

Object 对象在 JSON 中是用{}包含一系列无序的 Key-Value 键值对表示的,这里 Key 只能用 String 表示。例如,一个 Address 对象包含如下 Key-Value:

```
city:Beijing
street:Chaoyang Road
postcode:100025
```

用 JSON 表示为:{"city":"Beijing","street":"Chaoyang Road","postcode":100025},其中 Value 也可以是另一个 Object 或者数组,因此复杂的 Object 可以嵌套表示,例如一个 Person 对象包含 name 和 address 对象,可以表示如下:

```
{"name":"Michael","address":
  {"city":"Beijing","street":"Chaoyang Road","postcode":100025}
}
```

2. JSON 数据解析

在服务器端生成 JSON 格式的数据发送到客户端后,在客户端需要对接收到的纯文本数据进行解析。不同编程语言有不同的解析方式,这里简单介绍几种解析方式。

1) 使用 JavaScript 处理 JSON 数据

在 Web 页面中经常使用 JavaScript 处理 JSON 数据。这里通过一个简单的 JavaScript 方法看客户端如何将 JSON 数据表示给用户。

```
function handleJson() {
    var j = {"name":"Michael","address":
        {"city":"Beijing","street":"Chaoyang Road","postcode":100025}
    };
    document.write(j.name);
    document.write(j.address.city);
}
```

从代码可见,只需将服务器返回的 JSON 数据赋值给一个 JavaScript 变量,就可以立刻使用该变量并更新页面中的信息。相比 XML 需要从 DOM 中读取各种节点而言,JSON 的使用非常容易。JSON 已经是 JavaScript 标准的一部分。目前,主流的浏览器对 JSON 支持都非常完善,对那些应用 Ajax 的 Web 2.0 网站来说,JSON 是目前最灵活的轻量级方案。

2) 使用 Java 处理 JSON 数据

在 Java 编程中,可以使用 JSONObject 和 JSONArray 对返回的 JSON 字符串进行解析。JSONObject 与 JSONArray 的区别在于 JSONObject 是对象形式,而 JSONArray 是数组形式。下面以代码举例说明。

```
import net.sf.json.JSONArray;
import net.sf.json.JSONObject;
public static void main(String[] args) {
```

```
String jsonString = "{\"UserName\":\"HONG\",\"age\":\"38\",\"workIn\":\"IBM\",\"Array\":" +
    "[\" HONG \",\"38\",\"IBM\"]}\"";
//将 JSON 字符串转为 Java 对象
JSONObject obj = JSONObject.fromObject(jsonString);
//获取 Object 中的 UserName
if (obj.has("UserName")) {
    System.out.println("UserName:" + obj.getString("UserName"));
}
//获取 JSONArray
if (obj.has("Array")) {
    JSONArray transitListArray = obj.getJSONArray("Array");
    for (int i = 0; i < transitListArray.size(); i++) {
        System.out.print("Array:" + transitListArray.getString(i) + " ");
    }
}
}
```

3) 使用 .NET 处理 JSON 数据

.NET 可以使用 System.Runtime.Serialization.dll 与 System.Web.Extensions.dll 这两个 DLL 把对象序列化和反序列化成 JSON 数据。也可以使用第三方的 Newtonsoft.Json.dll 来操作 JSON 数据。使用它会更方便,其功能也强一些,下载地址是: <http://json.codeplex.com>。

3.4 数据预处理

现实世界中的数据基本是不完整、不一致的脏数据,无法直接进行数据挖掘,或挖掘结果差强人意。为了提高数据挖掘的质量,产生了数据预处理技术。数据预处理有多种方法:数据清理、数据集成、数据变换、数据归约等,如图 3-22 所示,这些数据预处理技术在数据挖掘之前使用,将大大提高数据挖掘模式的质量,降低实际挖掘所需要的时间。

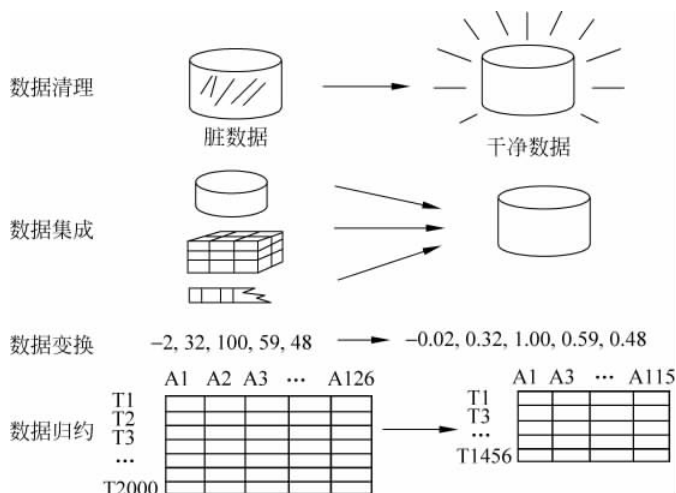


图 3-22 数据预处理的形式

3.4.1 数据清理

数据清理是通过填写缺失的值、光滑噪声数据、识别或删除离群点并解决不一致性等来“清理”数据。主要是达到格式标准化、异常数据清除、错误纠正、重复数据的清除等目标。

1. 缺失值

假设要分析一份销售和顾客的数据,但注意到许多元组的一些属性,如顾客的收入没有记录值。这时怎样才能为该属性填上缺失值? 表 3-6 给出了一些方法。表中方法 3~6 数据倾斜时填入的值可能不正确。然而,方法 6 是最常用的方法,与其他方法相比,它使用现存数据的最多信息来推测遗漏值。在估计收入的遗漏值时,通过考虑其他属性的值,有更大的机会保持收入和其他属性之间的联系。

表 3-6 缺失值的处理

编号	方 法	适 用	缺 点
1	忽略元组	元组有多个属性缺少值	不能使用该元组剩余属性值,这些数据可能有用
2	人工填写缺失值	缺少数据少	费时,数据集大且缺失值多时不适用
3	使用全局常量填充缺失值		简单不可靠
4	使用属性的均值或中位数填充缺失值	正常数据适用均值,倾斜数据使用中位数	数据不可靠
5	使用与给定元组属同类样本的均值或中位数填充缺失值	给定类的数据若分布倾斜,则选择中位数	数据不可靠
6	使用最可能的值填充缺失值	可以使用回归、贝叶斯形式、决策树归纳确定	最流行但数据不可靠

2. 噪声数据

噪声是测量变量的随机错误或偏差。给定一个数值属性,例如价格,怎样才能平滑数据去掉噪声呢? 下面是一些数据平滑技术。

(1) 分箱。分箱方法通过考察“邻居”(即周围的值)来平滑存储数据的值。存储的值首先被分布到一些箱中,对于按平均值平滑,箱中每一个值被箱中的平均值替换;类似地,可以使用按中值平滑或按边界平滑。对于按边界平滑,箱中的每一个值被最近的边界值替换。

(2) 回归。通过让数据适合一个函数(如回归函数)来平滑数据。线性回归涉及找出适合两个变量的“最佳”直线,使得由一个变量能够预测另一个变量。多元线性回归是线性回归的扩展,它涉及多于两个变量的情况,数据要适合一个多维面。使用回归找出适合数据的数学方程式能够帮助消除噪声。

(3) 离群点分析。局外者可以被聚类检测。聚类将类似的值组织成群或“聚类”。直观地,落在聚类集合之外的值被视为局外者,即离群点。

3. 不一致数据

对于有些事务,所记录的数据可能存在不一致。有些数据不一致可以使用其他材料人工地加以更正。例如,数据输入时的错误可以使用纸上的记录加以更正。知识工程工具也可以用来检测违反限制的数据。例如,知道属性间的函数依赖,可以查找违反函数依赖的

值。数据集成中也可能产生不一致：一个给定的属性在不同的数据库中可能具有不同的名字，这时可能存在冗余。

3.4.2 数据集成

数据集成是把不同来源、格式、特点和性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。在企业数据集成领域，已经有了很多成熟的框架可以利用。目前通常采用联邦式、基于中间件模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享问题，为企业提供决策支持。

数据集成的过程是将多个数据源中的数据结合起来并统一存储。在数据集成时，需要考虑许多问题。

1. 实体识别问题

例如，数据分析者或计算机如何才能确信一个数据库中的 `customer_id` 和另一个数据库中的 `cust_number` 指的是同一实体？通常，数据库和数据仓库有元数据——关于数据的数据。这种元数据可以帮助避免模式集成中的错误。

2. 冗余问题

一个属性如果能由另一个或另一组属性“导出”，则这个属性可能是冗余的。有些冗余可以被相关分析检测到。例如，给定两个属性，根据可用的数据，相关分析可以度量一个属性能在多大程度上蕴含另一个。除了检测属性间的冗余外，“重复”也应当在元组级进行检测。重复是指对于同一数据，存在两个或多个相同的元组。

3. 数据值冲突问题

对于现实世界的同一实体，来自不同数据源的属性值可能不同，这可能是因为表示、比例或编码不同。例如，重量属性可能在一个系统中以公制单位存放，而在另一个系统中以英制单位存放；不同旅馆的价格不仅可能涉及不同的货币，而且可能涉及不同的服务和税。

仔细将多个数据源中的数据集成起来，能够减少或避免结果数据集中数据的冗余和不一致性。这有助于提高其后挖掘的精度和速度。

3.4.3 数据变换

数据变换将数据转换成适合于挖掘的形式，数据变换可能涉及以下内容：

- 平滑：去掉数据中的噪声。相关技术包括分箱、聚类和回归等。
- 聚集：对数据进行汇总。例如，可以聚集日销售数据，计算月和年销售额。
- 数据泛化：使用概念分层，用高层次概念替换低层次“原始”数据。例如，分类的属性如 `street`，可以泛化为较高层的概念，如 `city` 或 `country`。类似地，数值属性如 `age`，可以映射到较高层概念如 `young`、`middle-age` 和 `senior`。
- 规范化：将属性数据按比例缩放，使之落入一个小的特定区间，如 $-1.0 \sim 1.0$ ，或 $0.0 \sim 1.0$ 。
- 属性构造：可以构造新的属性并添加到属性集中，以帮助挖掘过程。

这里主要介绍几种属性规范化方法。对于基于距离的方法，诸如神经网络、最近邻分类或聚类算法，规范化方法特别有用，可以帮助防止具有较大初始值域的属性与具有较小初始值域的属性相比时的权重不均衡问题。

1. 最小-最大规范化

最小-最大规范化对原始数据进行线性变换。假定 $\min_A$ 和 $\max_A$ 分别为属性 A 的最小和最大值。最小-最大规范化通过计算:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

将 A 的值 v 映射到区间 $[\text{new_min}_A, \text{new_max}_A]$ 中的 v' 。最小-最大规范化保持原始数据值之间的关系。但如果今后的输入落在 A 的原数据区之外,该方法将面临“越界”错误。

2. z-score 规范化

在 z-score 规范化(或零-均值规范化)中,属性 A 的值基于 A 的平均值和标准差规范化。 A 的值 v 被规范化为 v' ,由下式计算:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

式中, $\bar{A}$ 和 σ_A 分别为属性 A 的平均值和标准差。当属性 A 的最大和最小值未知,或局外者左右了最大-最小规范化时,该方法是有用的。

3. 小数定标规范化

小数定标规范化通过移动属性 A 的小数点位置进行规范化。小数点的移动位数依赖于 A 的最大绝对值。 A 的值 v 被规范化为 v' ,由下式计算:

$$v' = \frac{v}{10^j}$$

式中, j 是使得 $\max(|v'|) < 1$ 的最小整数。

3.4.4 数据归约

数据挖掘时往往数据量非常大,直接在海量数据上进行复杂的数据分析和挖掘将需要很长时间,因此这种分析是不现实或不可行的。数据归约技术可以用来得到数据集的归约表示,它小得多,但仍然接近于保持原数据的完整性。这样在归约后的数据集上挖掘将更有效,并产生相同(或几乎相同)的分析结果。

数据归约的策略如下:

- (1) 数据立方体聚集: 聚集操作作用于数据立方中的数据。
- (2) 维度归约: 可以检测并删除不相关、弱相关或冗余的属性。维度归约的方法主要包括主成分分析、小波变换等。
- (3) 数据压缩: 使用编码机制压缩数据集。分为有损和无损。
- (4) 数值规约: 用替代的、较小的数据表示替换或估计数据,如参数模型(只需要存放模型参数,而不是实际数据)或非参数方法,如聚类、选样和使用直方图等。
- (5) 离散化和概念分层产生: 属性的原始值用区间值或较高层的概念替换。概念分层允许挖掘多个抽象层上的数据,是数据挖掘的一种强有力的工具。

综上所述,现实世界的数据一般是脏的、不完整的和不一致的。数据预处理技术可以改进数据的质量,从而有助于提高其后的挖掘过程精度和性能。由于高质量的决策必然依赖于高质量的数据,因此数据预处理是知识发现过程的重要步骤。