# 商品价格数据的两种WEB挖掘算法比较

王红艳，朱全银，严云洋，钱 进
（淮阴工学院 计算机工程学院，江苏 淮安 223003）

**摘 要**：其他网络商店的商品实时价格是Web商店店主所关注的重要数据，Web数据挖掘使得这一需求变为现实.通过正则表达式算法与分词算法的比较研究，给出了基于正则表达式的商品价格抽取算法和基于分词的网站目录树抽取算法、HTML网页商品抽取算法与商品价格抽取算法.应用系统的实践表明，正则表达式算法的挖全率与正确率较低，而分词算法的挖全率与正确率都达到99%以上，完全满足应用需求，同时可以为商品的市场预测与分析提供依据.

**关键词**：商品价格；数据挖掘；正则表达式；分词；算法比较
**中图分类号**：TP391 **文献标识码**：A **文章编号**：1000－7180(2011)10－0168－05

# Compare Two Web Mining Algorithm for Commodity Price

WANG Hong-yan, ZHU Quan-yin, YAN Yun-yang, QIAN Jin
(Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian 223003, China)

**Abstract**: Commodities price of others e-supermarkets is the most important data for the shopkeepers of shop online. This requirement becomes actuality because of the Web mining developing very fast. The algorithm based on regular expression and the extract algorithm for directory tree of Website, commodities name on the Webpage and commodities price based on participle are described in detailed respectively. All of them depend on the researched of the regular expression and the participle algorithm. The implementation shows that the lower average full rate and accuracy rate is got from regular expression algorithm. However, the participle algorithm can get more than ninety nine percent of average full rate and accuracy rate. The results show as by this way can touch the shopkeepers minds, and it can support the originality data for the commodities markets and forecast analysis.

**Key words**: commodity price; data mining; regular expression; participle; algorithm compare

## 1 引言

随着网上商店的普及应用，网上商店店主迫切需要及时了解其他同类网店同类商品的销售价格，目前采取的办法多为店主对自己所了解的网站人工查阅，无法及时了解所有或者绝大多数网店的价格，以便确立自己的销售策略，满足网店的价格优势与销售量.而Web数据挖掘技术的日益成熟，使得这一需求可以得到省时省力且自动完成.Web数据挖掘一般分为内容挖掘、结构挖掘和日志挖掘[1].目前Web数据挖掘在商业的应用比较广泛，如研究Web用户[2]、Web挖掘算法[3]、Web攻击[4]、Web日志挖掘[5]和Web用户行为习惯[6]等.就目前的Web数据挖掘应用看，研究抽取的算法多，针对应用系统，以提高抽取准确率的应用研究比较少.

## 2 商品价格的正则表达式Web挖掘算法

Web数据挖掘的正则表达式算法是利用网页的URL获取特定数据的过程.正如文献[6]中阐述的一样，这种方法的抽取速度快，针对性强，但噪声数据多，文献也提出了逆序解析DOM树的方法节省查找时间.在网络环境噪声低，关注网页集中的时

**收稿日期**：2011－04－30；**修回日期**：2011－06－22
**基金项目**：江苏省创新基金(BC2009 208)；淮安市产学研合作计划(HAC201002)

候，正则表达式具有非常高效的特性.

商品价格抽取算法：

输入：网页地址 $A$

输出：手机价格 $P$

1：初始化 $A$

2：$S1$=获取 $A$ 中源代码

3：确定开头和结尾字符串 $S,E$

4：$K$=从 $S1$ 中找到以 $S$ 开头和与 $E$ 结尾的内容

5：IF($K$ 不为空)

6：　IF($K$ 匹配正则表达式)

7：　　找到

8：　ELSE

9：　　找不到

10：ELSE

11：　　找不到

12：ENDIF

13：输出 $K$

## 3　基于分词算法的商品价格 Web 挖掘

由于正则表达式算法在实际抽取应用系统中需要通过人工方法输入目标网页的 URL 地址，因此，在实际使用中局限性比较大，而分词算法可以避免这一问题. 分词算法的主要思想是在目标网站中，首先挖取网站的目录树结构，从目录树中比较商品分类的 URL，再从目标商品分类的目录树节点抽取目标商品页面的 URL，然后从目标商品页面的 HTML 文档中，采用分词法获取目标商品名称及其对应的价格. 分词法最大的优点是可以自适应网站的目录树结构变更，且挖全率比正则表达式算法高. 以下是以不同品牌不同网站手机价格的分词算法与价格抽取算法.

网站目录树抽取算法：

输入：网页 URL $A$，关键字 $K$

输出：所需 URL $B$

伪代码：

1：初始化 $A$

2：$S$=$A$ 的源代码

3：$U$=匹配 $S$ 所有<a><\/a>中的信息

4：FOR($U$ 中每一个元素 $e$)

5：　IF($e$ 中的“title”信息==$K$)

6：　　$B$=$e$ 中的“href”信息

7：　ELSE

8：　　没找到

9：ENDIF

10：ENDFOR

输出 $B$

HTML 页商品分词算法：

输入：短语 $A$，词库 $B$

输出：所得词语数组 $W$

1：初始化 $A,B$

2：WHILE($A$ 长度>$n$)

3：　$C$=$A$－最后 $n$ 个字

4：　WHILE($C$ 的长度>$n$)

5：　　$C$=$C$－最后 $n$ 个字

6：　ENDWHILE

7：　$D$=$C$

8：　$i$=1

9：　WHILE($D$ 长度>0)

10：　　WHILE($D-i$>0)

11：　　FOR(B 中的每一个元素 $e$)

12：　　$G$=$D$

13：　　$F$=$D$ 的前 $n-i$ 个字

14：　　IF($F$==$e$)

15：　　找到一个词

16：　　$W$=$W+F+$“”

17：　　$D$=$D-F$

18：　　BREAKFOR

19：　　ELSE

20：　　CONTINUEFOR

21：　　ENDIF

22：　　ENDFOR

23：　　IF($G$==$D$)

24：　　$i$++

25：　　ENDIF

26：　ENDWHILE

27：　IF($G$==$D$)

28：　　$D$=$D$ 去除第一个字

29：　ENDIF

30：　ENDWHILE

商品价格抽取分词算法：

输出：$W$

输入：网页地址 $A$，品牌 $B$

输出：手机价格 $P$

1：初始化 $A,B$

2：$S$=获取 $A$ 中源代码

3：$D$=切割 $S$

4：FOR(B 中每一个元素 $e$)

5：　WHILE($i$≤$K$)

6：　　IF($e$==$D[i]$)

7：　　BREAK WHILE

8：　　ELSE

9：　　CONTINUE WHILE

10：　　ENDIF

11：　　ENDWHILE

12：　　WHILE($i$<$K$)

```
13:      IF(D[i]为空)
14:        CONTINUE WHILE
15:    ELSE
16:      IF(D[i]中包含"¥" && D[i]中包含数字)
17:        找到价格 P
18:        BREAK WHILE
19:      ELSE
20:        CONTINUE WHILE
21:      ENDIF
22:    ENDIF
23:  ENDWHILE
24:ENDFOR
25:输出 P
```

## 4 实验分析

图 1 中给出应用系统 2011 年 3 月 24 日到 4 月 30 日采用分词法对淘宝商城、京东商城与当当网挖出的手机数量，图 2 所示为 5 款相同手机 4 月 30 日三个网站的数据库记录，图 3 所示为摩托罗拉 XT800 手机 2011 年 3 月 26 日到 4 月 30 日三个网站的价格走势，图 4 所示为诺基亚 5233 手机 2011 年 3 月 26 日到 4 月 30 日三个网站的价格走势.

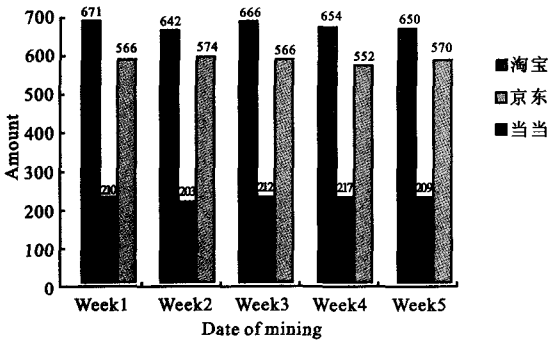


图 1 淘宝商城、京东商城与当当网销售的手机数量

表 1 与表 2 分别是正则表达式与分词算法淘宝商城单日手机价格的抽取对比，表 3 和表 4 是两种算法的挖全率与正确率对比.

对比结论：正则表达式算法的平均挖全率与正确率都比较低，分别为 78%和 83%，主要原因是网路不稳定与噪声引起的数据丢失，在实际应用系统

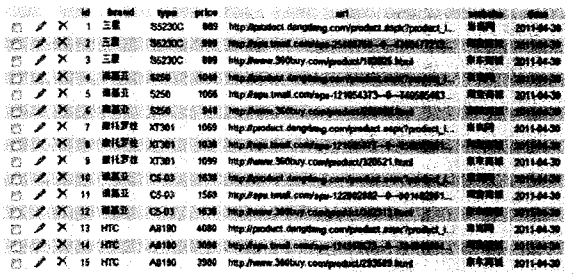

图 2 5 款相同手机三个网站的数据库记录

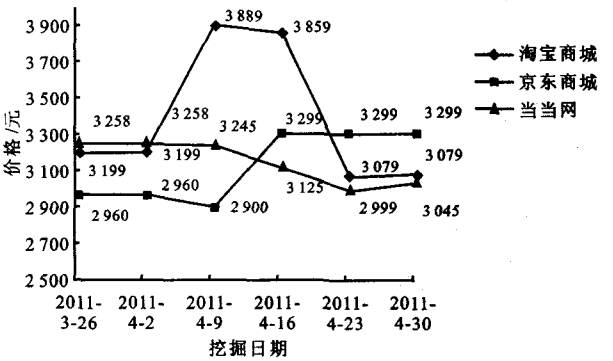


图 3 摩托罗拉 XT800 价格走势图

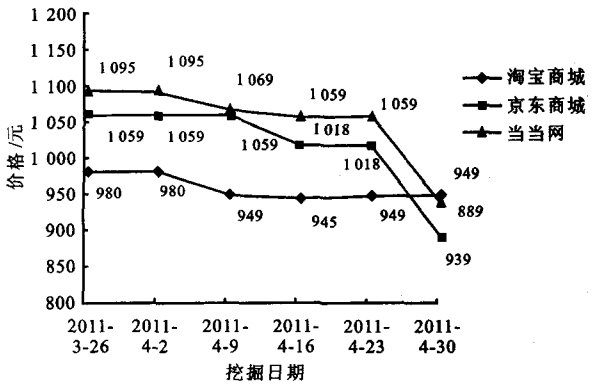


图 4 诺基亚 5233 手机价格走势图

中一般不建议采用，但是对于网络环境较好的企业，在目标网站 URL 相对稳定的商城可以使用；分词算法的平均挖全率与正确率都比较高，都达到 99%以上，完全能够满足网店店主对商品价格的数据自动抽取与分析. 如图 3 与图 4 示出的两款手机的价格走势分析. 因此具有较好的应用前景. 另外，对于价格波动相对较为敏感的商品，本系统均有较好的市场应用前景.

表 1 正则表达式法单日数据

| 序号 | 商品 | 价格 | 商城 | 时间 |
|---|---|---|---|---|
| 1 | 诺基亚 5230XM | 0 | 淘宝商城 | 2011—03—30 |
| 2 | Motorola/摩托罗拉 XT30 | 1 028 | 淘宝商城 | 2011—03—30 |
| 3 | Huawei/华为 U8500 | 1 040 | 淘宝商城 | 2011—03—30 |
| 4 | 999 元 Android2.1 安卓系统 欧盛 W180 3.5 寸 | | 淘宝商城 | 2011—03—30 |
| 5 | 诺基亚 C5—03 | 1 648 | 淘宝商城 | 2011—03—30 |
| 6 | Nokia/诺基亚 X3—02 | 0 | 淘宝商城 | 2011—03—30 |
| 7 | 天语 X90 | 530 | 淘宝商城 | 2011—03—30 |

表 2　分词法单日数据

| 序号 | 商品 | 型号 | 价格 | URL | 商城 | 时间 |
|---|---|---|---|---|---|---|
| 1 | 诺基亚 | 5230XM | 1 048 | http://spu.tmall.com/spu-68865206---0--.htm? | 淘宝商城 | 2011-03-30 |
| 2 | 摩托罗拉 | XT301 | 1 028 | http://spu.tmall.com/spu-121085877---0--.htm? | 淘宝商城 | 2011-03-30 |
| 3 | 华为 | U8500 | 1 040 | http://spu.tmall.com/spu-121085850---0--.htm? | 淘宝商城 | 2011-03-30 |
| 4 | 诺基亚 | 5250 | 905 | http://spu.tmall.com/spu-121054373---0--.htm? | 淘宝商城 | 2011-03-30 |
| 5 | 欧盛 | W180 | 999 | http://item.tmall.com/item.htm? id=9128294003&is_b=1 | 淘宝商城 | 2011-03-30 |
| 6 | 诺基亚 | C5-03 | 1 648 | http://spu.tmall.com/spu-122802802---0--.htm? | 淘宝商城 | 2011-03-30 |
| 7 | 诺基亚 | X3-02 | 1 055 | http://spu.tmall.com/spu-120764571---0--.htm? | 淘宝商城 | 2011-03-30 |
| 8 | 天语 | X90 | 530 | http://spu.tmall.com/spu-58969121---0--.htm? | 淘宝商城 | 2011-03-30 |

表 3　正则表达式算法挖全率与正确率　　%

| 抽取时间 | HTC | | Nokia | | MOTO | | Sony Ericsson | | Apple | |
|---|---|---|---|---|---|---|---|---|---|---|
| | full rate | accuracy | full rate | accuracy | full rate | accuracy | full rate | accuracy | full rate | accuracy |
| 2011/3/1——2011/3/3 | 81 | 85 | 68 | 83 | 67 | 98 | 98 | 82 | 67 | 100 |
| 2011/3/4——2011/3/6 | 29 | 100 | 36 | 81 | 18 | 87 | 69 | 67 | 83 | 58 |
| 2011/3/7——2011/3/9 | 100 | 52 | 94 | 81 | 100 | 79 | 100 | 78 | 100 | 58 |
| 2011/3/10——2011/3/12 | 81 | 63 | 67 | 79 | 72 | 81 | 72 | 93 | 83 | 75 |
| 2011/3/13——2011/3/15 | 100 | 71 | 83 | 92 | 92 | 77 | 91 | 89 | 67 | 81 |
| 2011/3/16——2011/3/18 | 90 | 69 | 69 | 97 | 80 | 86 | 85 | 91 | 50 | 72 |
| 2011/3/19——2011/3/21 | 81 | 65 | 60 | 97 | 85 | 80 | 93 | 89 | 58 | 89 |
| 2011/3/22——2011/3/24 | 90 | 58 | 53 | 96 | 82 | 87 | 99 | 91 | 50 | 100 |
| 2011/3/25——2011/3/27 | 100 | 72 | 69 | 97 | 94 | 92 | 87 | 91 | 58 | 89 |
| 2011/3/28——2011/3/30 | 100 | 67 | 75 | 99 | 100 | 90 | 85 | 100 | 75 | 100 |
| 2011/3/31——2011/4/2 | 81 | 77 | 67 | 99 | 89 | 86 | 91 | 95 | 67 | 100 |
| 2011/4/3——2011/4/5 | 100 | 81 | 67 | 96 | 76 | 88 | 86 | 95 | 67 | 78 |
| Total full rate: | 78 | | | | | | | | | |
| Total accuracy: | 83 | | | | | | | | | |

表 4　分词算法挖全率与正确率　　%

| 抽取时间 | HTC | | Nokia | | MOTO | | Sony Ericsson | | Apple | |
|---|---|---|---|---|---|---|---|---|---|---|
| | full rate | accuracy | full rate | accuracy | full rate | accuracy | full rate | accuracy | full rate | accuracy |
| 2011/3/25——2011/3/27 | 100 | 100 | 100 | 99 | 100 | 99 | 100 | 100 | 100 | 100 |
| 2011/3/28——2011/3/30 | 100 | 100 | 99 | 99 | 100 | 99 | 99 | 99 | 100 | 100 |
| 2011/3/31——2011/4/2 | 100 | 100 | 100 | 99 | 99 | 99 | 100 | 99 | 100 | 100 |
| 2011/4/3——2011/4/5 | 100 | 100 | 99 | 99 | 100 | 99 | 100 | 100 | 100 | 100 |
| 2011/4/7——2011/4/10 | 100 | 100 | 100 | 99 | 100 | 99 | 100 | 99 | 100 | 100 |
| Total full rate: | 99 | | | | | | | | | |
| Total accuracy: | 99 | | | | | | | | | |

## 5 结束语

Web数据外挖掘可以为商业带来市场的行情，了解市场是商家有的放矢的基础.文中通过网店店主的应用需求，利用正则表达式法和分词法分别开发了基于Web的手机价格数据挖取，实验结果表明，利用正则表达式法对网络环境的依赖性比较强，但速度快，而分词法对网络带宽的依赖性比较低，且对网页框架及内容的变化适应性强，但对计算机的运算性能要求高.随着技术的日益进步，除商业应用外，Web挖掘应用将可以为人们提供更多的个性化服务，如为人们提供Web新闻[8]、分析E-mail邮件[9]和找工作[10]等.相信随着Web应用的丰富，基于Web的挖掘一定将存在于我们生活的方方面面.

**参考文献：**


[1] Chen Qi, Hou Ming. XML－based data mining design and implementation [C]//International Conference on Computer Design and Applications. Qinhuangdao, China:IEEE, 2010: 610－613.

[2] Antony S, Wu Ping, Agrawal D. et al. Aggregate skyline: analysis for online users [C]//Ninth Annual International Symposium on Applications and the Internet. Bellevue, Washington, USA:IEEE, 2009: 50－56.

[3] Alla H, Al－Ghreimil N. A novel efficient classification algorithm for search engines [C]//Computational Intelligence for Modelling Control & Automation. Vienna, Austria:IEEE, 2008: 773 － 778.

[4] Atanasova T, Kasheva M, Sulova S, et al. Analysis of the possible application of Data Mining, Text Mining and Web Mining in Business Intelligent Systems [C]// Proceedings of the 33rd International Convention. Opatija, Croatia:IEEE, 2010: 1294 － 1297.

[5] 何波，涂飞，程勇军. Web日志挖掘数据预处理研究.[J].微电子学与计算机，2011，28(4)：111－114.

[6] Salin S, Senkul P. Using semantic information for web usage mining based recommendation [C]//International Symposium on Computer and Information Sciences. Northern Cyprus:IEEE, 2009: 236 － 241.

[7] 张瑞雪，宋明秋，公衍磊. 逆序解析DOM树及网页正文信息提取[J].计算机科学，2011，38(4)：213－215.

[8] Xu Cheng Zhong, Ibrahim T I. A keyword－based semantic prefetching approach in Internet news services [J]. Knowledge and Data Engineering, 2004, 16(5): 601 － 611.

[9] Grobelnik M, Mladenic D, Fortuna B. Semantic technology for capturing communication inside an organization [J]. Internet Computing, 2009, 13(4): 59 － 67.

[10] Litecky C, Aken A, Ahmad A, et al. Mining for computing jobs [J]. Software, 2010, 27(1): 78 - 85.


**作者简介：**


**王红艳** 女，(1979－)，硕士，讲师.研究方向为计算机应用；

**朱全银** 男，(1966－)，教授.研究方向为智能信息处理、接口与通信；

**严云洋** 男，(1967－)，博士，教授.研究方向为模式识别、数字图像处理.


(上接第167页)

态方式，并从信息论的角度出发，给出分辨系数取值的具体算法，解决了以往分辨系数难以量化的问题.并将该算法应用于河南省铁路客运量实例，给出了基于动态灰关联分辨系数的河南省铁路客运量影响因素诊断模型.仿真结果表明，该算法是合理的，而且能有效提供关联度分辨力，使关联分析更符合实际，是一种合理的选择输入变量的有效方法.

**参考文献：**


[1] 申卯兴，薛西锋，张小水.灰色关联分析中分辨系数的选取[J].空军工程大学学报，2003，4(1)：68－70.

[2] 刘 琦.影响铁路客流的因素及相关度分析[J].上海铁道大学学报，1999，20(2)：19－20.

[3] 汪健雄，刘春煌，单杏花. 基于双层次正交神经网络模型的铁路客运量预测[J]. 中国铁道科学，2010，31(3)：126－132.

[4] 夏国恩，金炜东，张葛祥. 改进SVR及其在铁路客运量预测中的应用[J]. 西南交通大学学报，2007，42(4)：494－498.

[5] 吕锋.灰色系统关联度之分辨系数的研究[J].系统工程理论与实践，1997，17(6)：49－54.

[6] 范凯，吴皓莹.灰色系统关联度中一种新的分辨系数确定方法[J].武汉理工大学学报，2002，24(7)：87－89.


**作者简介：**


**王文莉** 女，(1978－)，硕士，讲师.研究方向为数据挖掘、计算机网络；

**杨俊红** 女，(1970－)，硕士，副教授.研究方向为数据挖掘、数字视频处理.