Commodities Price Dynamic Trend Analysis Based on Web Mining

Quanyin Zhu, Hong Zhou, Yunyang Yan, Jin Qian and Pei Zhou
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaiian, Jiangsu Province, China
hyitzqy@126.com

Abstract—Commodities price of others e-supermarkets or online shopping systems is the most important data for the shopkeepers of shop online. This requirement becomes actuality because of the Web mining developing very fast. The Web mining algorithm from extracting directory tree of different Website, the commodities name on the Webpage and commodities price based on participle are described in detailed. All of them depend on the researched of the participle algorithm. The implementation shows that the participle algorithm can get more than ninety nine percent of average full rate and accuracy rate. The error rate of price dynamic trend analysis is less than four percent. The results show as by this way can touch the shopkeepers minds, and it can support the originality data for the commodities markets and dynamic trend analysis.

Keywords- commodities price; dynamic trend analysis; mobile phone; participle algorithm; Web mining

I. INTRODUCTION

With the fast and continuous development of the electronic commerce, how to extract the price and get the dynamic trend analysis is very important for the shopkeepers selling online. More and more data grow rapidly in various and complex forms, mining of those complex data becomes an significant task in data mining realm. The knowledge discovery in database theory is applied to many field, such as electricity future market [1], product conceptual design [2] and the product conceptual design [3].

There have been many researches on mining data. Building a mining target model and use process data extracting markup language to describe the model [4], object-oriented technology is applied in defining the mining target is used to map the mining target model. Classification of data mining is adopted to reach a framework that can map data mining techniques to data stream mining challenges and requirements [5]. Cluster analysis technique have been applied in the computing professions [6], the web content data mining utilizing cluster analysis to classify data or discover new resources. Semantic technology for capturing communication utilized a semi-automatically constructor [7].

A keyword-based semantic perfecting approach is applied to internet news services and implement a client-side personalized perfecting system[8]. Web mining is always include content mining, framework mining and log mining [9]. Web mining is used widely on commercial application such as researching the Web users[10], Web attack [12] and web usage mining [13], etc. According to the researched reports, we can find that many researchers interest on the extracting algorithms but less on improving the extracting rate.

Mobile phones’ technique is making rapid progress, it has become one of the focuses in e-commerce. Shopkeepers want to know more information about the new products price and the market trend. How to extract the price is a key problem for market analysis. Our approach is based on semantic analysis method and participle algorithm. We presents a instance for commodities price dynamic trend analysis of mobile phones selling online.

II. THE PRICE MINING AND PARTICIPLE ALGORITHM

A. The Price Mining Algorithm

Through the requirement analysis, we can conclude that it is very important to get the exact price of mobile phone through a web page, and it can be used as a strong proof of mobile phone price movement for users’ analysis and making marketing strategies.

Price mining module uses two methods to achieve the information of mobile phone sales on three websites in the mobile phone. Three websites are: Jingdong Mall, Taobao Mall and Dangdang.

The method of semantic analysis is to get information from mobile phone shelf on the site, and compare the mining information to the database, if the model of mobile phone that day found had already dig in the website, then not add to the database. Detail methods and steps are as follows:

Firstly, obtain the URL of mining page that we need, then get the directory tree in current page, find the content related to the phone in the directory, enter that directory and get the correspondingly source code for the URL in the directory, at the same time, divide the string and stored in the array, traversal the array, applicant the segmentation algorithm and compare with the mobile phone brand name in the database, if found the mobile phone brand included in string, otherwise continue find, once found and then see whether the brand is the user needed, If so, get the corresponding source code of a hyperlink to the string, and then get the number behind the source code with "¥" which as the price of the mobile phone. Pseudo code is as follows:

Input: URL A, Brand B
Output: Price of phone P
I: initialize A, B
2: S1 = get the source code of A
3: D = divide S1
4: FOR(e: element in D)
5:   WHILE(i <= K)
6:     IF(e == D[i])
7:       BREAK WHILE
8:   ELSE
9:     CONTINUE WHILE
10:  ENDWHILE
11:  WHILE(i < K)
12:   IF(D[i])
13:     CONTINUE WHILE
14:   ELSE
15:     IF(D[i] concludes “¥” && D[i] is number)
16:       find P
17:       BREAK WHILE
18:     ELSE
19:       CONTINUE WHILE
20:    ENDWHILE
21:  ENDWHILE
22:  ENDWHILE
23:  ENDFOR
24:  print P

Advantage of this method is that no matter how much the site templates change, you can find accurate information. The detail knowledge of semantic analysis will be introduced in the Part B.

B. The Participle Algorithm

Maximal matching algorithm include positive maximal matching algorithm, reverse maximal matching algorithm, two-way matching algorithm, etc. The main principle of it is word segmentation, and then compares with the word in the database, if it is a word, recorded it, or through increased or decreased a word, continue to compare, until remain a word is terminated, if the word cannot be segmented any more, stop the process. In Mobile Phone Sales Decision Support System, we use positive maximal matching algorithm. The characteristic of this algorithm is faster speed and easy to achieve.

The achieve path of positive maximal matching algorithm is suppose the maximum word length in the dictionary is m, we divide a string as some short string based on punctuation, then get the first m words, look up this word in the dictionary to judge it is a word, if it is true, delete this word from the string, if it is false, delete the last word of this string and check it is a single word or not, if it is true, output this word and delete this word from this string, if it is false, continue look up this word in the dictionary to judge it is a word, repeat this process until output a word. After it return the beginning of it. Put it this way, a long string can divided into a combination of some short words.

The step of positive maximal matching algorithm is like following:

- Initialize and get short string D1;
- If D1 is not null, get the first m words into W;
- Look up the dictionary, if W words in the dictionary, put the W into D2, D1=D1-W, if W words in not in the dictionary, delete the last word of W;
- Judge the W is a single word or not, if it is a single word, put W into D2, D1=D1-W, if not return step 3 go on to look up the dictionary, until W in the dictionary;
- Judge the D1 is null or not, if is null, stop to analysis, if not return step 2, continue to loop, until D1 is divided into words.

Relative to the algorithm based on understanding and statistics for, the characteristic of positive maximal matching algorithm is smaller complex, easy to achieve, only need to establish a table. But it will cause a string have different means. So usually we use this algorithm with other algorithm. Though the algorithm based on understanding and statistics for will have a very accuracy result, almost 100%, but it need to establish a very big regular database, so it will spend much time to look up. In general, users are advised to use the positive maximal matching algorithm combine with reverse maximal matching algorithm to participle, it will enhance the accuracy. Pseudo code is as follows:

Input: String A, Lexicon B
Output: Words W
1: initialize A, B
2: WHILE(length(A)>n)
3:   C = A-last n words
4:   WHILE(length(C)>n)
5:     C = C-last n words
6:   ENDWHILE
7:   D = C
8:   i = 1
9:   WHILE(length(D)>0)
10:  WHILE(D-i>0)
11:    FOR(e: elements in D)
12:      G = D
13:      F = first n-I words of D
14:      IF(F==e)
15:        find a words
16:        W = W+F+" "
17:        D = D-F
18:      ELSE
19:        D = delete first word of D
20:    ENDWHILE
21:  ENDWHILE
22:  ENDWHILE
23:  print W

III. THE DYNAMIC ANALYSIS ALGORITHM

We can assume the equation of the average values is:
\[ y = ax^2 + bx + c \]  
(1)

\( y \): the next average value of seasonal factor.
\( x \): the value of weeks.

In order to simplify the computing, let’s \( x=1; \ x=2; \ x=3 \) into the equation.

We can get the Ternary Once Equations:

\[
\begin{align*}
y_1 &= a + b + c \quad \text{that is} \quad a = (y_3 - 2y_2 + y_1)/2 \\
y_2 &= 4a + 2b + c \quad \quad \quad \quad b = (-3y_1 + 4y_2 - 3y_3)/2 \\
y_3 &= 9a + 3b + c \quad \quad \quad \quad c = 3y_1 - 3y_2 + y_3 
\end{align*}
\]  
(2)

Because of \( y_1, y_2, y_3 \) are known values. So we can get the value of \( a, b \) and \( c \). we can express the equations (1) as:

\[
y = \frac{(y_1 - 2y_2 + y_3)}{2}x^2 + \frac{(-3y_1 + 4y_2 - 3y_3)}{2}x + (3y_1 - 3y_2 + y_3)
\]  
(3)

Then if \( x=4 \), we can get the \( y=0.07692 \).

So, we can use \( y=0.07692 \) to work out the next period seasonal factor:0.88532.

Then each mobile price can use the monthly average prices multiply 0.88532.

Of course, in the actual program, this factor is changed but not static.

In order to convenient for compare the results, we set \( N=3 \) and \( N=5 \) to simulation only one month data to determine the price analysis model.

\[
\hat{y}_{t+1} = \frac{y_t + y_{t+1} + y_{t+2}}{3} \quad (N=3)
\]  
(4)

\[
\hat{y}_{t+1} = \frac{y_t + y_{t+1} + y_{t+2} + y_{t+3} + y_{t+4}}{5} \quad (N=5)
\]  
(5)

Using the equation (4) and (5), we can get the different step size \( N \) effect the analysis results which are show as Figure 1 and 2 respectively.

### Table I. Full Rate and Accuracy of Participle Algorithm

<table>
<thead>
<tr>
<th>Time</th>
<th>HTC Full Rate</th>
<th>HTC Accuracy</th>
<th>Nokia Full Rate</th>
<th>Nokia Accuracy</th>
<th>MOTO Full Rate</th>
<th>MOTO Accuracy</th>
<th>Sony Ericsson Full Rate</th>
<th>Sony Ericsson Accuracy</th>
<th>Apple Full Rate</th>
<th>Apple Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011/3/25——2011/3/31</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/4/1——2011/4/7</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/4/8——2011/4/14</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/4/22——2011/4/20</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/4/28——2011/5/4</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/5/5——2011/5/11</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/5/12——2011/5/18</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/5/19——2011/5/25</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2011/5/26——2011/6/1</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table II. Four Months Forecasting Price and Errors

<table>
<thead>
<tr>
<th>Mobile Phone</th>
<th>Mobile Phone 1</th>
<th>Mobile Phone 2</th>
<th>Mobile Phone 3</th>
<th>Mobile Phone 4</th>
<th>Mobile Phone 5</th>
<th>Mobile Phone 6</th>
<th>Mobile Phone 7</th>
<th>Mobile Phone 8</th>
<th>Mobile Phone 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>actual</td>
<td>forecast</td>
<td>error</td>
<td>actual</td>
<td>forecast</td>
<td>error</td>
<td>actual</td>
<td>forecast</td>
<td>error</td>
</tr>
<tr>
<td>HTC</td>
<td>965</td>
<td>965</td>
<td>0.00%</td>
<td>952</td>
<td>965</td>
<td>1.37%</td>
<td>952</td>
<td>960.67</td>
<td>0.91%</td>
</tr>
<tr>
<td>Nokia</td>
<td>1066</td>
<td>1066</td>
<td>0.00%</td>
<td>1066</td>
<td>1066</td>
<td>0.00%</td>
<td>1066</td>
<td>1066</td>
<td>0.00%</td>
</tr>
<tr>
<td>MOTO</td>
<td>948</td>
<td>949.7</td>
<td>0.07%</td>
<td>999</td>
<td>925</td>
<td>3.34%</td>
<td>999</td>
<td>982</td>
<td>1.70%</td>
</tr>
</tbody>
</table>
The four months price forecasting can be achieved more than 96% and it only depend on incompletely extracting data. So it can be touch very useful application for the shopkeeper’s selling online.

Figure 3 and Figure 4 show the two kinds of mobile phone dynamic trend analysis for different e-supermarket from March 17 to May 13, 2011. It also support us a very kind belief of our research work.

V. CONCLUSION AND FUTURE WORK

The market of mobile phone change very fast. So the price dynamic trend analysis is pay more attention by the shopkeepers. All data mining is as same importance as individual analysts from enterprise standpoint. We believe that in the future, dynamic trend analysis system using the data mining technology can have greater development. Although our work achievement the very useful effect that touched the shopkeepers mind. But future interesting work is waiting on our hard work such as the decision support system for the commodities selling online.

ACKNOWLEDGMENT

This work is supported by the Industry-University Collaboration Project of Huai'an City, China (HAC201002); the fund of Huai'an Industry Science and Technology, China(HAG2010064).

REFERENCES

[3] Zhen Peng,Bingru Yang,Wu Qu. One knowledge discovery approach for product conceptual design. 2009 International Joint Conference on Artificial Intelligence,Page(s):109-112