Research on the New Products Discovery based on Web Mining

Jin Ding, Quanyin Zhu, Luijiang Zhou and Jin Qian
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, Jiangsu Province, China
hyitzqy@126.com

Abstract—The new products discovery is a very interesting data which can bring new business opportunity for shopkeepers selling online. By using the Web mining can get more and more data in everywhere such as e-supermarkets and e-commerce. This paper shows a case study for a new products discovery of mobile phones selling online. Extracting the Web news and the e-supermarkets to get the new products using the participle algorithm is researched. The discovery algorithm and its application are described in detailed. The functions coded by PHP are designed for implantation the new products discovery extracting. Experiment demonstrates its performance and proves the new products discovery is meaningful and useful for the shopkeepers selling online.

Keywords- new products; discovery; mobile phone; selling online; Web mining

I. INTRODUCTION

With the fast and continuous development of the electronic commerce, how to extract the price of the new products are 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, products conceptual design and detection of faulty products [1,2,3].

There have been many studies on mining data. Building a mining target model and use Process Data Extracting Markup Language to describe the model [4], Oriented-Object technology is applied in defining the mining target and E-R figure 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].

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 mobile phones. How to extract the new products is a task for mining data. Our approach is based on semantic analysis method and Participle algorithm. We present a instance for a new products discovery of mobile phones selling online.

II. THE SYSTEM DESIGN

In this section, the system data flow, data mining process and the system functions are designed for the new products discovery. All the extracting design based on the Web mining technologies, it can not only extract product selling online but also news published in the website.

A. DFD Design

The first process, reading webpage at the beginning of the mobile phone information extracting, the user needs to find the URL and read the webpage, identify which is the mobile phone and which is not, this process needs the data in D5-new website and D2-brand.

The second process is adding products when the user finds a kind of mobile phone. The system will compare it with the information store in D5 of new products, if it exists in the database means it is not a new product, if not will update the database.

B. Flow Chart Design

After user logins in the system and selects in the data mining model, first select the website that user want to extract from and then enter the key word “mobile phone”, then go to the next page, select the brand of mobile phone which you want to extract. You can select one brand or more than one, if you select nothing, system will display error message and return beginning, if not, the user can extract the information. In the process of extracting information, the
system will detect the network environment automatic at any
time. After the system accesses a webpage, system will
record the URL into the database, if network interrupts at
this time, user only needs refresh the page, system can find
the last URL before the network interrupts and extracts
information continue. Figure 2 shows flow chart for data
mining process.

![Flow chart for data mining process](image1)

In data analysis model, the user can select the website
and destination to extract the new product: one is from
storage rack of the new product; the other is from news
published in the website. After user selects the website and
destination, the system will access the website to extract new
product. When system extracting a product, it will compare
with data in database in order to judge the information is the
new product or not. If the information existed in the database
the system will not insert this product in to the database, the
other way, the system will insert this product information in
the database. Figure 3 shows flow chart for new products
extracting process.

C. Main Functions Design

The MVC architecture is used to develop the new
products extracting system of mobile phone. The main
functions are more than forty; some of them are introduced
as follow:

1) get_product_url($url) function
   Belong to: JingDong, TaoBao
   Function: Getting the URL of all the product of current
   webpage.
   Achieve method: invoke file_get_contents() to get the
   source code of current webpage, invoke eregi() to find the
   string between string “<ul class="list-h clearfix">” and “<div
   class="m clearfix>”, then find the hyperlink insert into the
   database.
   Return: URL of all the products of current webpage

2) get_product($brand) function
   Belong to: JingDong, TaoBao
   Function: Extracting the information of cell phone.
   Achieve method: Invoking the above all functions to
   finish the extract of product.
   Return: Null.

3) get_nextpage($url) function
   Belong to: JingDong, TaoBao
   Function: Getting the URL of the next webpage of
   current webpage.
   Achieve method: Invoking file_get_contents() to get the
   source code of current webpage, invoke eregi() to find the
   string between string “var_product_addTime” and “<!--filter
   end-->”, then find the hyperlink that there is a “nextpage”
   behind it.
   Return: URL of nextpage of current webpage

4) lessthan($str,$array,$MaxLength) function
   Belong to: get_url
   Function: Using participle method to divide the string
   into word when length of string <=4
   Achieve method: Judge the length is 1 or not, if it is 1,
   means it is not a word, if not judges the length is 2 or not, if
   it is 2 check it in the database, if finds it return true, if not
   judges the length is 3 or not, if it is 3 check the first 2 words
   and last 2 words in the database, if finds it returns true, if not
   judges the length is 4 or not, if it is 4 check first 2 words,
   middle 2 words, last 2 words, first 3 words and last 3 words
   in the database, if finds it return true, if not returns false.
   Return: if finds the word return true, if not returns false.

5) geturldata($url) function
   Belong to: get_url, get_info;
   Function: Getting the source code of webpage;
   Achieve method: Invoking curl_setup(),curl_init() and so
   on to get the source of current webpage;
   Return: source code of current webpage.

6) morethan($str,$array,$MaxLength) function
Belong to: get_url
Function: Using participle method to divide the string into word when length of string > 4.
Achieve method: Dividing the first 4 word from the string then invoke lessthan() to find the word.
7) search_url($url) function
Belong to: get_url
Function: Finding the URL of webpage that show the all product or contain the “nextpage”.
Achieve method: Invoking eregi() to find the webpage that contain the key word “nextpage” if find means current webpage is the basement, it can be the first webpage to start to extract.
Return: webpage to start extract.
8) goto_mobile_index ($str) function
Belong to: newsina, newsohu
Function: finding level2 webpage from the $str.
Achieve method: invoking preg_match() to find the hyperlink of key word “cell phone” from the $str and invoking geturldata() to get the source code of that webpage.
Return: source code of the hyperlink of key word “cell phone”.
9) goto_new ($str) function
Belong to: newsina, newsohu
Function: finding level3 webpage from the $str.
Achieve method: invoking preg_match() to find the hyperlink of key word “new product” or “more new product” from the $str and invoking geturldata() to get the source code of that webpage.
Return: source code of the hyperlink of key word “new product” or “more new product”.
10) get_news($arr, $str) function
Belong to: newsina, newsohu
Function: finding all the corresponding URL of today’s news.
Achieve method: first getting all the date string from the source code of current webpage then find all the date that match today’s date, then invoking preg_match() to find the URL of hyperlink before that date.
Return: arraying combine all the corresponding URL of today’s news.
11) get_parameter($i, $data) function
Belong to: newsina, newsohu
Function: finding all information of current cell phone
Achieve method: invoking the get_net(), get_money(), get_shape(), get_OS(), get_screen() to find all the detail parameter of current cell phone
Return: shape of current cell phone.

III. NEW PRODUCT MINING ALGORITHM

According to introduce of the life cycle of mobile phone, a conclusion can be got is that new products mining functions are very important for user to know the market information of mobile phone.

There are two directions in new products mining model: one is extract new products of cell phone information from goods shelf published in the “new product” webpage. The other one is extracting the news of new products on website that can be extracted from the news to find the new products.

A. New Products Mining Algorithm

This method can extract the information of the mobile phone in the goods shelf published in the website, and then compares with the records in the database to judge whether it is new product, if it is new product, add it into the database, if not go on extracting. Detail methods and steps are as follows:

Firstly, capturing the URL of website that we need, then get the catalogue tree of current webpage, find the catalogue that related with “mobile phone”, enter that catalogue, get all the brands, find the new products in each brands. After extracted the information compare with the records in the database, if that information is not existed, the system consider it is the new product of today and add it in the database, or it is the old product.

The algorithm is as follows:
Input: URL A, Brand B
Output: New product name N
1: initialize A
2: S=get the source code of A
3: D=divide S
4: T=get the date of today
5: WHILE(i<=K)
6: IF(D[i]==T)
7: WHILE(i-1>0)
8: IF(D[i-1] is null)
9: CONTINUE WHILE
10: ELSE
11: F=D[i-1]
12: FOR(e: element in B)
13: IF(F include e)
14: N=N+e+" 
15: j=index of last words in F
16: WHILE(j-<L)
17: IF(F[j] is letter or F[j] is number or F[j] is ".")
18: N=N+F[j]
19: ELSE
20: CONTINUE WHILE
21: ENDWHILE
22: ELSE
23: go to row 5
24: ENDIF
25: find new product
26: go to row 5
27: ENDIF
28: ELSE
29: CONTINUE FOR
30: ENDIF
31: ENDFOR
32: ENDIF
33: ENDWHILE
34: ELSE
35: CONTINUE WHILE
36: ENDIF
37: ENDWHILE
38:print N
B. Product News Extracting

This method can achieve extracting the information of mobile phone from the new product news published in the website, then compare with records in the database to judge whether it is the new product. If it is a new product then add it in the database or pass it go on extracting. The method of getting the brand and type of mobile phone from the news catalog is like the price mining of using semantic analysis method. Detail methods and steps are as follows: getting the URL of website that we need, then get the catalog tree of current webpage, find the catalog that related with the “mobile phone”, then select the news that published today, find the brand and type in the news by using participle method in that news, then compare it with records in the database to judge whether it is new product, if it is a new product, add it into the database, if not go on extracting. The algorithm is like the part A listed above.

IV. SYSTEM IMPLEMENTATION

We have programmed the application system which coded by PHP language. The news of new products on the website like Figure 4 and Figure 5. The Figure 4 shows a new product of Samsung S3778 which will be to the market at Jun 2011, and Figure 5 shows a new product of Nokia C9 which will be to the market at 2012. Table 1 lists the ten kinds of new products we have discovered from May 19 to Jun 12 2011 on the news of www.sina.com.

![Figure 4. The new product of Samsung S3778 on the www.sina.com](attachment:image1.png)

![Figure 5. The new product of Nokia C9 on the www.sina.com](attachment:image2.png)

V. CONCLUSION AND FUTURE WORK

The market of mobile phone changes very fast. So the new products extracting is paid more attention by the shopkeepers. We have discovered more than 200 kinds of new products from website. The system achievement is very useful effect that touched the shopkeepers mind. Our future interesting work is on the decision support system for the mobile phone selling online.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Time to market</th>
<th>Price</th>
<th>Network Format</th>
<th>Shape</th>
<th>Screen</th>
<th>Web URL</th>
<th>Extracting Time</th>
</tr>
</thead>
</table>

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, HAG08009).
REFERENCES


