Journal of Information & Computational Science 9: 16 (2012) 5039–5046
Available at http://www.joics.com

A Hybrid Approach for New Products Discovery of Cell Phone Based on Web Mining ⋆

Quanyin ZHU*, Jin DING, Yonghua YIN, Pei ZHOU

Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian 223005, China

Abstract

How to find new products is a challenge issue because of it can bring a new business opportunity for shopkeepers selling online. A hybrid approach for new products discovery of cell phones is adopted to verify the approaches. The participle algorithm and product news extracting algorithm are utilized to analyze the news on website. The selling information on the e-supermarkets is used to find the data and the price for new cell phones selling start. All the cell phone data and news are extracted from different website. The extraction functions coded by PHP and Python languages are designed for implantation the new products discovery extracting. Experiment demonstrates the proposed approach performance and proves the new products discovery is meaningful and useful for the cell phone shopkeepers selling online.

Keywords: New Products Discovery; Cell Phone; Participle Algorithm; System Functions; Web Mining

1 Introduction

The semantic analysis and its application are paid more attention in the modern e-government and e-commerce. The participle algorithm based on the semantic analysis get more and more research and application especially on the intelligent information systems. For example, Reference [1] researches the semantic gap between image contents and tags and Reference [2] researches the semantic analysis and organization of spoken documents. According to some researchers reports based on the semantic Web, we can know more about the use of intelligent application systems. However, for the application systems of semantic Web, the computational trust prototype [3, 4], classification approach [5, 6, 7, 8], search strategies [9, 10], semantic class learning [11], Web crawling approach [12], semantic link analysis [13], etc. are the important issues need to pay more attention to researched. On the other hand, 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. The knowledge discovery in database theory is applied to many field, such as electricity future market, products conceptual design and detection of faulty products [14,

⋆Project supported by the National Sparking Plan of China (No. 2011GA690190).
*Corresponding author.
Email address: hyitzqy@126.com (Quanyin ZHU).

December 1, 2012

15, 16]. There have been many studies on mining data. Building a mining target model and use process data extracting markup language to describe the model [17]. Classification of data mining is adopted to reach a framework that can map data mining techniques to data stream mining challenges and requirements [18]. Cluster analysis technique have been applied in the computing professions [19], 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 [20]. A keyword-based semantic perfecting approach is applied to internet news services and implements a client-side personalized perfecting system [21]. Shopkeepers want to know more information about new cell 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 an instance for a new products discovery of cell phones selling online.

2 Mathematical Notation

The extracted method is a Web extracting method based on the authors' previous work [22, 23, 24]. All the extracting design based on the Web mining technologies, it can not only extract product selling online but also news published in the websites.

2.1 Semantic analysis on products relative

Depend on the information theory, Pointwise Mutual Information (PMI) is a measure of the amount of information regarding the occurrence of one word when we observe an other. PMI between two words is defined as follows:

$$PMI(word_1, word_2) = log_2 \left| \begin{array}{c} p(word_1 \& word_2) \\ p(word_1)p(word_2) \end{array} \right|. \tag{1}$$

Eq.(1) is the probability that $word_1$ and $word_2$ appear together. $p(word_1)$ and $p(word_2)$ are the probabilities that $word_1$ and $word_2$ occur respectively. For this task, a web search engine and a method of processing the returned result are needed. Reference [25] extracts nouns and noun phrases as candidate product features which compose of two or more adjacent nouns, while the strategy may bring many noises. For the first aspect, they propose the boundary dependency (BD) algorithm to verify the phrase boundary of candidates as following equation:

$$BD(w_1, \cdots, w_n) = \frac{f(bdw + w_1 + \cdots + w_n) f(w_1 + \cdots + w_n + bdw)}{f(w_1, \cdots, w_n)^2}. \tag{2}$$

$w_1, \cdots, w_n$ denotes an extracted adjacent noun in the specific product reviews, and $f(w_1, \cdots, w_n)$ is its frequency. Reference [26] presented a context-aware information recommendation service model which can be also referred.

2.2 The participle algorithm

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 approach of positive maximal matching algorithm is like following:

• **Step 1** Initialize and get short string $D1$;

• **Step 2** 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 is not in the dictionary, delete the last word of W;

• **Step 3** 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;

• **Step 4** 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.

3 New Products and Product News Extracting Algorithm

Let the view-source of webpage be defined as V, body of the page be defined as P:

$$P \subseteq V. \tag{3}$$

Let the normalized text be defined as $\hat{P}$

$$\hat{P} \subseteq P. \tag{4}$$

Keywords corpus consists of some self-learned keyword. Let the keywords corpus be defined as K, then could be presented as

$$K = \{k_1, k_2, \cdots, k_n\}. \tag{5}$$

In fact, the new products information field included in K. Let defined it as k_i, the position of k_i in $\hat{P}$ could be defined as l

$$l = (k_i, \hat{P}). \tag{6}$$

The potential new products information field could be defined as m, define a constant as con then m maybe equals to the normalized text range between m and $m + con$:

$$m = \hat{P}(m, m + con). \tag{7}$$

Finally, because of other fields effecting, we must remove the effect. The positions of K except k_n could be defined as N:

$$M = \{m_1, m_2, \cdots, m_n\}. \tag{8}$$

Let define the minimum position in M as m_{min}define the real new products information as k_{real}, we can conclude that the real expert information filed:

$$k_{real} = f(0, m_{min}). \tag{9}$$

For product news, using Eq. (3) to Eq. (9), we can achieve extracting the information of cell phone from the new product news published in the websites, 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. 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 *cell phone*, then select the news that published today, find the brand and type in the news by using Eq. (3) to Eq. (9) algorithm, then compare it with records in the database to judge whether it is a new product or not, if it is a new product, add it into the database, if not go on extracting.

4 Main Functions Design

The MVC architecture is used to develop the new products extracting system of cell phone. The main functions are more then 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 less than or equal four

Achieve method: Judge the length is one or not, if it is one, means it is not a word, if not judges the length is two or not, if it is two check it in the database, if finds it return true, if not judges the length is three or not, if it is three check the first two words and last two words in the

database, if finds it returns true, if not judges the length is four or not, if it is four check first two words, middle two words, last two words, first three words and last three 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) 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.

(6) 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.

5 System Implementation

We programmed the application system which coded by PHP and Python language. The new products news on the website is shown Fig. 1. The (a) shows a new product of Motorola Titan and (b) shows a new product of Samsung S3778 which will be to the market at Jun 2011 respectively; The (c) shows a new product of Nokia WP7 and (d) shows a new product of Sony Xperia Leon which will be to the market at 2012 respectively. Table 1 lists the ten kinds of new products we have discovered from May 19 to Jan. 31 2012 on the news of www.sina.com. Another approach is to find the selling information on the different e-supermarkets for the new cell phone products. Fig. 2 shows one of new cell phone product on different e-supermarkets. Table 2 lists the ten kinds of the new products time to market we have extracted from April 14 to Jan. 20 2012 on the three e- supermarkets.

6 Conclusion and Future Work

In order to find new cell phones, we built an integrated application system. The semantic analysis theory and participle algorithm are used to segmentation products characteristics which include name, type, price, time to market etc. The entire source extracted from websites. The proposed approach can not only discovery the new cell phones from news of websites but also find the time to market from e-supermarkets.The market of cell phone changes very fast. So the new products mining are paid more attention by the shopkeepers. We have discovered more than 400 kinds of new products from different websites. Our future interesting work is on the decision support system for the cell phone selling online.

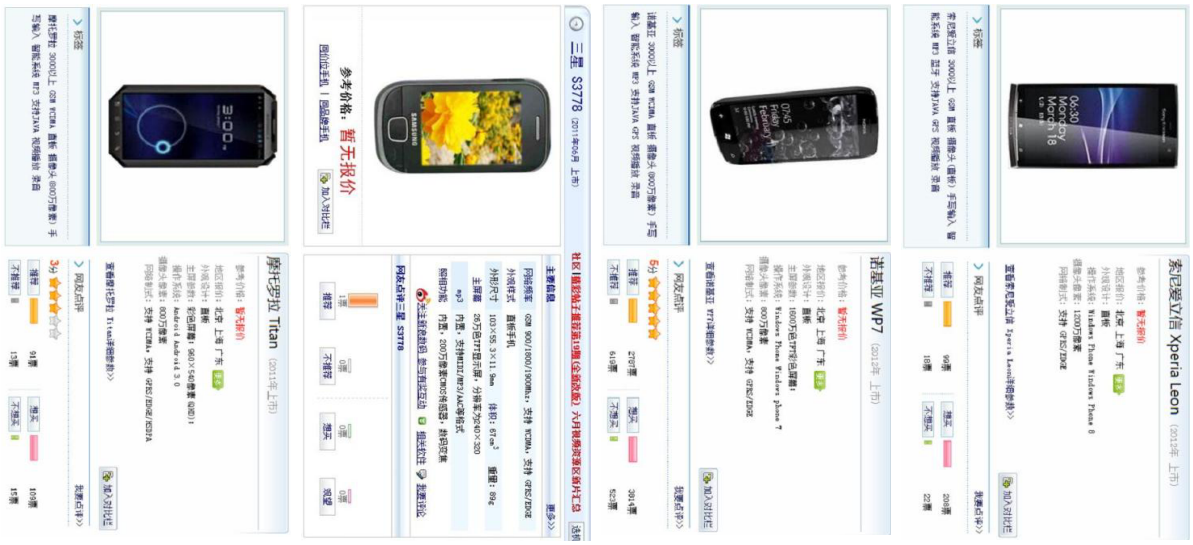

Fig. 1: The new products on the www.sina.com (a) Motorola Titan (b) Samsung S3778 (c) Nokia WP7 (d) Sony Xperia Leon

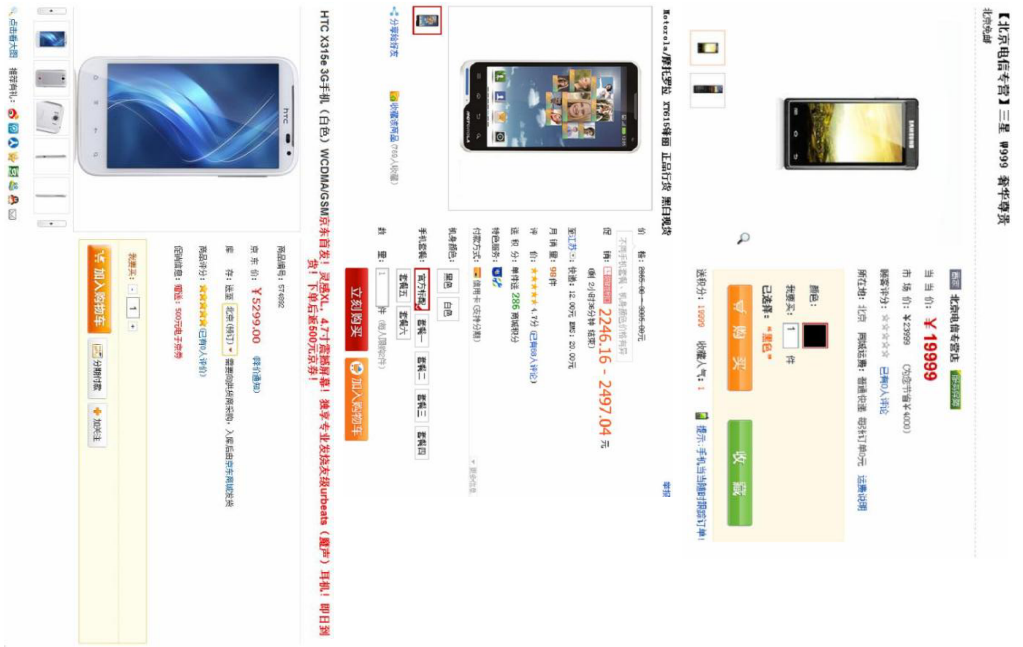

Fig. 2: The new cell phone product on different e-supermarkets (a)HTC X315e on the www.360buy.com (b) Motorola XT615 on the www. tmall.com (c) Samsung W999 on the www. dangdang.com

References

[1] H. Ma, J. Zhu, M. R. Lyu, T. King, Bridging the semantic gap between image contents and tags, IEEE Transl. J. Multimedia, 12 (2010), 462-473.

[2] S. Y. Kong, L. S. Lee, Semantic analysis and organization of spoken documents based on parameters derived from latent topics, IEEE Transl. J. Audio, Speech, and Language Processing, 19 (2011), 1875-1889.

[3] Y. Zhang, H. J. Chen, Z. H. Wu, X. Q. Zheng, Develop a computational trust prototype for the

Table 1: Ten kinds of new products discovered from website news

Name	URL	Extracting Time
Sony Severine	http://tech.sina.com.cn/mobile/models/11797.html	2011-5-19
Motorola EX130	http://tech.sina.com.cn/mobile/models/11798.html	2011-5-19
Sharp SH-12C	http://tech.sina.com.cn/mobile/models/11824.html	2011-5-27
Nokia C9	http://tech.sina.com.cn/mobile/models/11812.html	2011-5-27
Samsung E2652W	http://tech.sina.com.cn/mobile/models/11840.html	2011-6-09
Motorola EX212	http://tech.sina.com.cn/mobile/models/11872.html	2011-6-10
Samsung S3778	http://tech.sina.com.cn/mobile/models/11869.html	2011-6-10
Nokia WP6	http://tech.sina.com.cn/mobile/models/11682.html	2012-1-31
Nokia WP7	http://tech.sina.com.cn/mobile/models/11681.html	2012-1-31
Sony Xperia Leon	http://tech.sina.com.cn/mobile/models/11808.html	2012-1-31

Table 2: Time to market for ten kinds of new products find from e-supermarkets

Name	e-supermarket	Schedule	Time to market	Price (CNY)
Sony Xperia Arc	http://product.dangdang.com	2011-03	2011-04-19	4980
Samsung W999	http://product.dangdang.com	2011-11	2012-01-05	19999
Nokia 701	http://product.dangdang.com	2011-08	2012-1-19	2498
Samsung I9220	http://www.360buy.com	2011-09	2012-1-17	5999
HTC 315e	http://www.360buy.com	2012-01	2012-1-18	5299
IPhone 4S	http://www.360buy.com	2012-01	2012-1-13	5399
LG E730	http://www.360buy.com	2011-09	2011-12-21	2699
Motorola XT615	http://spu.tmall.com	2011-11	2011-12-02	2898
Samsung I9200	http://item.tmall.com	2011-11	2011-12-3	5124
Motorola XT928	http://spu.tmall.com	2011-11	2011-12-2	5680

semantic Web, in: Proc. 22nd International Conference on Data Engineering Workshops, 2006, 57-63.

[4] J. L. Zhang, Q. L. Shen, J. Y. Wu, E-Businessmen credibility mining based on web reviews, in: Proc. International Conference on E -Business and E -Government, 2011, 1-4.

[5] S. Q. Yin, Y. H. Qiu, J. Ge, F. Wang, A Chinese text classification approach based on semantic Web, in: Proc. Fourth International Conference on Semantics, Knowledge and Grid, 2008, 497-498.

[6] H. Zhou, B. W. Liu, J. Liu, Research on methods of ontology-based class label semantic similarity computation, in: Proc. International Conference on Computational and Information Sciences, 2011, 276-280.

[7] Y. Li, F. Tian, F. Ren, S. Kuroiwa, Y. X. Zhong, A method of semantic dictionary construction from on-line encyclopedia classifications, in: Proc. International Conference on Natural Language Processing and Knowledge Engineering, 2007, 82-89.

[8] J. Y. Xuan, X. F. Luo, S. X. Zhang, Z. Xu, H. M. Liu, F. Y. Ye, Building hierarchical keyword level association link networks for Web events semantic analysis, in: Proc. IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, 2011, 987-994.

[9] Y. W. Yang, Personalized search strategies for spatial information on the Web, IEEE J. Intelligent Systems, PP (2010), 1.

[10] Z. L. Su, H. S. Chen, L. Zhu, Y. H. Zeng, Framework of semantic Web service discovery based on Fuzzy logic and multi-phase matching, Journal of Information and Computational Science, 9 (2012), 203-214.

[11] X. X. Zhang, Z. F. Sui, Semantic class learning with deep coordinate structures in Web pages, Journal of Computational Information Systems, 8 (2012), 1245-1254.

[12] Y. Q. Dong, Q. Z. Li, A deep Web crawling approach based on query harvest model, Journal of Computational Information Systems, 8 (2012), 973-981.

[13] Z. J. Liu, Y. J. Du, Focused crawling based on semantic link analysis, Journal of Computational Information Systems, 8 (2012), 1213-1220.

[14] W. L. Feng, S. D. Liu, M. Y. Lai, X. H. Deng, Empirical research on price discovery efficiency in electricity futures market, in: Proc. Power Engineering Society General Meeting, 2007, 1-6.

[15] Z. Peng, B. R. Yang, W. Qu, One knowledge discovery approach for product conceptual design, in: Proc. International Joint Conference on Artificial Intelligence, 2009, 109-112.

[16] M. A. Karim, G. Russ, A. Islam, Detection of faulty products using data mining, in: Proc. 11th International Conference on Computer and Information Technology, 2008, 101-107.

[17] S. N. Liu, X. T. Tian, Z. M. Zhang, A process data extracting method in process planning knowledge discovery, in: Proc. IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009, 517-521.

[18] M. Kholghi, H. Hassanzadeh, M. R. Keyvanpour, Classification and evaluation of data mining techniques for data stream requirements, in: Proc. International Symposium on Computer, Communication, Control and Automation, 2010, 474-477.

[19] C. Litecky, A. Aken, A. Ahmad, H. James Nelson, Mining for computing jobs, IEEE J. Software, 27 (2010), 78-85.

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

[21] C. Z. Xu, T. I. Ibrahim, A keyword-based semantic prefetching approach in Internet news services, IEEE Transl. J. Knowledge and Data Engineering, 16 (2004), 601-611.

[22] Q. Y. Zhu, Y. Y. Yan, J. Ding, J. Qian, The case study for price extracting of cell phone sell online, in: Proc. 2nd IEEE International Conference on Software Engineering and Service Sciences, 2011, 282-285.

[23] Q. Y. Zhu, Y. Y. Yan, J. Ding, Y. Zhang, The commodities price extracting for shop online, in: Proc. International Conference on Future Information Technology and Management Engineering, 2010, 317-320.

[24] J. P. Deng, F. W. Cao, Q. Y. Zhu, Y. Zhang, The Web data extracting and application for shop online based on commodities classified, in: Y. W. Wu (Ed.), Communications in Computer and Information Science, Springer-Verlag Berlin Heidelberg, 234 (2011), 189-197.

[25] Q. Su, X. Y. Xu, H. L. Guo, Z. L. Guo, X. Wu, X. X. Zhang, Hidden sentiment association in Chinese Web opinion mining, in: Proc. WWW 2008, 2008, 959-968.

[26] Z. M. Zeng, Research on context-aware information recommendation service for the ubiquitious Web, Journal of Computational Information Systems, 9 (2012), 3715-3722.