# Factors Affecting User Clicks on Ads

Wenqi Li[1] and Ziyang Xu[2](✉)

[1] Department of International, Shenyang University, Shenyang 110003, China
[2] Department of Management, Hainan University, Hainan 570100, China
20192701310014@hainanu.edu.cn

**Abstract.** In recent years, e-commerce has become a new and popular industry, and the issue of click-through e-commerce advertising has received widespread attention. To increase product exposure and sales, more and more e-commerce companies are focusing on the issue of advertising. Displaying ads that are relevant to users' needs will greatly increase user satisfaction. Therefore, this study collected data on ads displayed on Taobao, user information, and ad clicks over eight days and conducted logistic regression and decision tree analysis on the data. Based on the perspective of user personality, the relationship between ad clicks and different user groups was analyzed to suggest effective strategies for ad operators. The findings show that ad clicks are influenced by age, gender and consumption level. It is suggested that more e-commerce companies can personalize their advertising operations to target different user groups, to more effectively increase the click-through rate of advertising, which in turn will increase the sales rate of their products and maximize their benefits.

**Keywords:** E-commerce · Taobao · Ad clicks · Logistic regression · Decision tree

## 1 Introduction

### 1.1 Background

Overall, the e-commerce industry in China continues to grow steadily. As Internet technology continues to improve, more professional services are being offered to platform users by e-commerce service providers, thus significantly reducing the costs required in the transaction process. In addition, with the rapidly developing technology of e-commerce, more and more offline businesses are looking for a change and transitioning onto the e-commerce development path. The e-commerce industry in China is still growing steadily, but the growth rate has gradually slowed, indicating greater competitive pressure within the e-commerce industry. Taobao, founded by Alibaba Group in May 2003, has the largest online retail and business community in Asia and is the most well-known online retail platform in China. According to statistics, Taobao has over 500 million registered users and more than 60 million regular visitors per day. The sheer

Wenqi Li and Ziyang Xu—contributed equally.


G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 2307–2318, 2022.
https://doi.org/10.2991/978-2-494069-31-2_272

volume of customers has resulted in a daily online product volume of over 800 million items, with an average of approximately 48,000 items sold every minute on Taobao. As of early 2012, the highest single-day Taobao transaction value reached a staggering 4.38 billion yuan, indirectly or directly generating 2.708 million jobs. With the rapid growth of Taobao and the increase of its user base, Taobao has also transformed from a single C2C network model to an integrated retail platform that includes C2C, group buying, distribution, auction and other e-commerce models. It has become one of the most successful e-commerce trading platforms.

### 1.2 Related Research

Shan et al. found that real-time bidding (RTB) provides an advertising ecosystem and faces similar challenges to sponsored search advertising. They compared the performance and runtime complexity of the approach with Tucker decomposition, canonical decomposition and other popular CTR prediction measures for CTR prediction on a real-world advertising dataset. The outcomes showed that the improved model had better prediction quality compared to other models [1]. To test the effect of age and gender on the click-through rate of online advertisements, Higgins et al. used multivariate tests and web-based techniques and found that personalized ads affect user engagement and increase CTR [2]. Keyword-based advertisements are becoming the dominant form of online advertising as they can be tailored and customized to the relevant information of the potential consumer. Gopal et al. studied the interaction between the two channels. The results showed that there was significant encroachment between the two channels, with a significant decrease in impression gain within each channel. This suggests that under certain conditions, two channels may be required to optimize advertising returns for both advertisers and service providers [3]. Gao and Gao used a Tencent click log dataset with millions of records (soso.com). First, they describe how so search engine advertising works. The system architecture is based on a click log dataset, in which ads with sufficient click log history are observed. The paper shows how ad-centric features can be used to discover models that can identify factors that affect CTR prediction performance [4]. Today, banner size, placement, and media planning factors are the focus of research on internet advertising. Chtourou et al. studied the effect of price and promotion in advertising on click-through rates (CTR), which predict user interest in a product and potential purchase behavior. A study of a database of approximately 1,200 ad insertions showed a significant interaction between price or promotion and display position [5]. Most classic search engines select and rank ads based on click-through rates (CTR). To forecast the click-through rate of an advertisement, it is often necessary to look at the historical click-through rate. Fang et al. developed a BN-based model to forecast the CTR of new advertisements. The keyword Bayesian network used to describe the advertisements in the domain is called KBN. Experimental results have shown the effectiveness and accuracy of this method [6]. Ads on search engines (SE) are usually ranked based on click-through rate (CTR). Accurately predicting the click-through rate of an ad is crucial to maximizing advertising revenue. A model for inheriting click-through information of rare/new ads from other semantically related ads to better predict click-through rate values was proposed by Kushal S. Dave et al. [7].

By reviewing counterfactual learning methods, Yuan et al. point out some difficulties in applying these methods for click-through rate prediction in realistic advertising systems. To overcome these difficulties, they proposed a new counterfactual CTR prediction framework to improve [8]. In the era of big data information, the key for companies in various fields seeking returns is how to effectively predict and analyze the click-through rate of infomercials. Zhu constructs predictive models based on neural network algorithms, extracts effective features, and performs predictive analysis based on simplified LSTM neural network time-serialized features. The prediction model is trained using CNN convolutional neural networks. This paper analyses the characteristics of traditional prediction methods [9]. Geoffrey Atkinson et al. Studied the relationship between specific elements of search engine advertising (SEA) and click-through rate (CTR) in Google AdWords campaigns and found that elements such as branding, value "hype", and price, promotion, and questioning were associated with CTR. Pros and cons, depending on where they appear in the ad. [10].

As mentioned earlier, most of the papers found in this paper focus on users' ad-clicking behaviors that side with search engines and social platforms such as Facebook. And many of the papers' studies also estimated CTR from different age and gender perspectives. (CTR is click-through rate, by the way.) Most of the previous research methods include logistic regression and multivariate tests. These methods are very useful in this field and support to some extent the direction of building models for data processing.

### 1.3 Objective

The objective of this study was to explore what factors influence the click-through rate of ads placed on the Alibaba platform over 8 days. People focus more on the benefits of advertising and ignore the factors that affect click-through rates, which makes it an interesting topic to study. The factors include the price, gender, age, consumption level, consumption depth, status, and city level, for a total of 7 factors. Some of these seemingly non-influential factors can have a significant impact, contrary to common perceptions.

## 2 Data and Method

The dataset was searched using the AliCloud Tianchi platform and the dataset was provided by the AliMama platform. The dataset describes Taobao's ads, user information, and ad clicks, and all data were desensitized. Three excel files were finally selected, including the original sample, the sample of basic ad information, and the sample of basic user information. The three excel files were combined based on the original samples to extract the desired user characteristics, and then the first graph is the generic dataset. Because there are some null values, they are meaningless. So they were removed with excel and then the new dataset in the second picture was obtained, which can be used directly to learn and build the model (Table 1).

In the organized dataset, variables 1, 2 and 3 were set and simplified by using the numbers 0,1,2,3. 5 variables were finalized: for clicks, set clicks equal to 1 and non-clicks equal to 0. For gender, set male to 1 and female to 2. For consumption level, set

**Table 1.** Features of the selected data

| Variable | Mean | Maximum Value | Minimum Value | Standard Deviation |
|---|---|---|---|---|
| Y Variable | 290.79 | 392.7 | 150.9 | 80.88 |
| X Variable 1 | 13,700 | 17,200 | 10,100 | 2350.18 |
| X Variable 2 | 7,293,067 | 7,507,400 | 7,024,200 | 155,171.53 |
| X Variable 3 | 193.04 | 201.4 | 174 | 7.30 |

**Table 2.** Relevant variables expressed in numerical terms

| Factor | Setup | | |
|---|---|---|---|
| click/non-click | click 1 | non-click 0 | |
| gender | man 1 | woman 2 | |
| consumption grade | low 1 | middle 2 | high 3 |
| Shopping level | new 1 | general 2 | regular 3 |
| occupation | undergraduate 1 | non-graduate 0 | |

low to 1, medium to 2, and high to 3. For shopping level, set new to 1, average to 2, and normal to 3. For occupation, set university student to 1 and non- daxue is 0 (Table 2).

Previous studies are mostly based on the advertisement feature, while the selected data are all from users' perspectives to study, analyze and offer operation strategy, and this is the data strength of this paper. For example, it not only retained age and gender which previous papers referred to, but also we added users' consumption depth, consumption level, occupation, and city level.

## 2.1 Logistic Regression

The specific objectives of the forecast are clearly defined as ad click-through rates. And there are also seven influencing factors regarding the click-through rate of ads, which are price, gender, age, consumption level, consumption depth, status, and city level. Calculations are based on historical statistics of the independent and dependent variables, based on which the regression analysis equation is established.

Predictive modeling:

ad click-through ratesi = $\alpha$ + $\beta$0 price + $\beta$1 gender + $\beta$2 age + $\beta$3consumption level + $\beta$4consumption depth + $\beta$5status + $\beta$6city level + $\gamma$Controlsi + i.

## 2.2 Decision Tree

Machine learning was performed on these data and a decision tree model was constructed using a growth approach (CHAID) with 70% of the learning sample and 30% of the testing sample.

Logistic regression analysis was carried out on the overall data, with 'click or no click' as the dependent variable and seven other data as covariates: price, gender, age, depth of consumption, level of consumption, occupation, and city level. The method used was Forward: LR.

## 3 Results and Discussion

This paper uses SPSS 26.0 to analyze the above data, which is divided into three main sections.

### 3.1 Covariance Test

A covariance test was performed on the data before data analysis to ensure the stability of the model in the regression analysis.As shown in Tables 3 and 4, all data tolerances were greater than 0.1 and VIF was less than 10; indicating that there was no covariance between the variables.

### 3.2 Logistic Regression

The significant variables obtained in the results are gender, age, and consumption level. According to the OR values in the analysis results, it can be seen that: 1. Women tend to click on ads pushed by e-commerce platforms more than men, with the ratio of male to female clicks being about 1:1.071; 2. in the sample of users aged 0–69, the chances of users clicking on ads show a curve of first decreasing and then increasing as their age increases. Taking the 0–9-year-old user group as the benchmark, the data reached the lowest (0.895) in the 20–29-year-old age group, and for the first time exceeded the benchmark value of 1 (1.019) for the user sample aged over 50, reaching the highest (1.215) when the user was over 60 years old; 3. In the 3 different consumption levels of low, medium, and high users, the click rate of users on advertisements showed a decreasing trend with the increase in consumption level. Using the click-through rate of users at low consumption levels as a benchmark value1, the figures for medium and high consumption levels are 0.989 and 0.890 respectively (Figs. 1, 2 and 3; Tables 5 and 6).

### 3.3 Decision Tree

From the model constructed, it can be seen that the ads pushed to users are divided into five price bands according to the price of the product. In the price range below \$12.60, a total of 6.8% of users clicked, the highest percentage of users and low consumption level in the whole price range; while in the adjacent (price range), the city class of the user plays a decisive role. In addition, among users in the price (price range), the gender factor plays an important role in the classification, i.e. the ratio of men to women is 1: 2.023; and in this price range, 67.29% of the female group is at a medium to high consumption level (Fig. 3).

**Table 3.** Coefficient

| Model | Unstandardized coefficients | | Standardized coefficient | | | Collinearity statistics | |
|---|---|---|---|---|---|---|---|
| | B | Standard error | Beta | t | Salience | Tolerance | VIF |
| Constant | .044 | .005 | | 8.988 | .000 | | |
| Ad_feature.price | − 4.738E-10 | .000 | .000 | − .220 | .826 | 1.000 | 1.000 |
| User_profile.gender | .003 | .001 | .007 | 4.408 | .000 | .979 | 1.021 |
| User_profile.age rating | .002 | .000 | .010 | 5.624 | .000 | .885 | 1.130 |
| User_profile.consumption level | − .002 | .001 | − .005 | − 3.290 | .001 | .941 | 1.063 |
| User_profile.shopping level | − .001 | .002 | − .001 | − .922 | .357 | .988 | 1.013 |
| User_profile.whether college students | − .001 | .002 | − .001 | − .858 | .391 | .919 | 1.088 |
| User_profile.city level | 8.023E-5 | .000 | .000 | .215 | .830 | .979 | 1.021 |

**Table 4.** Collinearity diagnosis

| Dimension | Eigen value | Condition indicator | Constant | Ad_ feature. Price | User_ profile. Gender | User_ profile. Age rating | User_ profile. Consumption level | User_ profile. Shopping level | User_ profile. Whether college students | User_profile.city level |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.708 | 1.000 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2 | 1.000 | 2.389 | .00 | 1.00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 3 | .961 | 2.437 | .00 | .00 | .00 | .00 | .00 | .00 | .90 | .00 |
| 4 | .130 | 6.625 | .00 | .00 | .00 | .05 | .22 | .00 | .01 | .55 |
| 5 | .094 | 7.812 | .00 | .00 | .53 | .07 | .09 | .00 | .02 | .21 |
| 6 | .074 | 8.807 | .00 | .00 | .00 | .65 | .49 | .00 | .04 | .03 |
| 7 | .031 | 13.631 | .04 | .00 | .45 | .20 | .19 | .07 | .03 | .17 |
| 8 | .003 | 44.534 | .96 | .00 | .01 | .02 | .01 | .92 | .00 | .04 |

**Table 5.** Variables in the equation

| | B | Standard error | Wald | Degree of freedom | Salience | Exp(B) |
|---|---|---|---|---|---|---|
| User_profile.gender(1man,2woman) | .068 | .016 | 18.618 | 1 | .000 | 1.071 |
| User_profile.age rating | | | 51.775 | 6 | .000 | 1.000 |
| User_profile.age rating(1) | –.029 | .461 | .004 | 1 | .949 | .971 |
| User_profile.age rating(2) | –.111 | .459 | .059 | 1 | .809 | .895 |
| User_profile.age rating(3) | –.051 | .459 | .012 | 1 | .912 | .951 |
| User_profile.age rating(4) | –.018 | .459 | .002 | 1 | .968 | .982 |
| User_profile.age rating(5) | .019 | .459 | .002 | 1 | .967 | 1.019 |
| User_profile.age rating(6) | .195 | .461 | .179 | 1 | .672 | 1.215 |
| User_profile.consumption level (1 low 2 medium 3 high) | | | 17.726 | 2 | .000 | 1.000 |
| User_profile.consumption level (1 low 2 medium 3 high) (1) | –.011 | .017 | .433 | 1 | .510 | .989 |
| User_profile.consumption level (1 low 2 medium 3 high) (2) | –.117 | .029 | 16.491 | 1 | .000 | .890 |
| Constant | –2.967 | .459 | 41.761 | 1 | .000 | .051 |

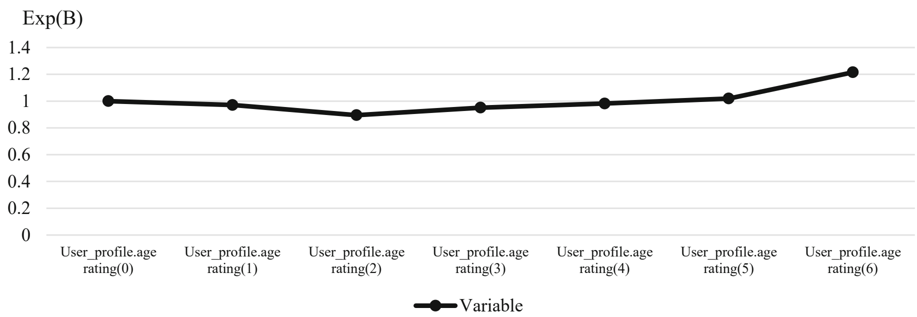


**Fig. 1.** Relationship between age rating and Exp(B)

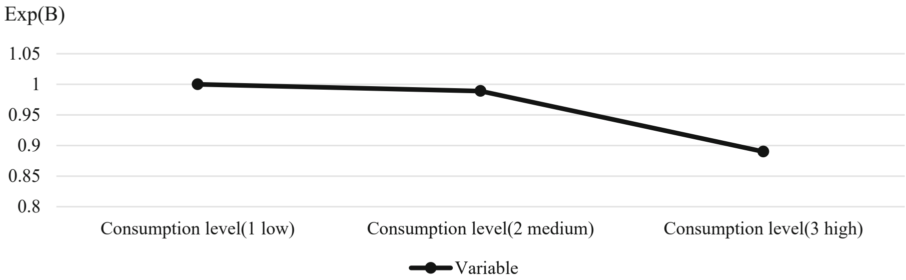


**Fig. 2.** The relationship between consumption level and Exp(B)

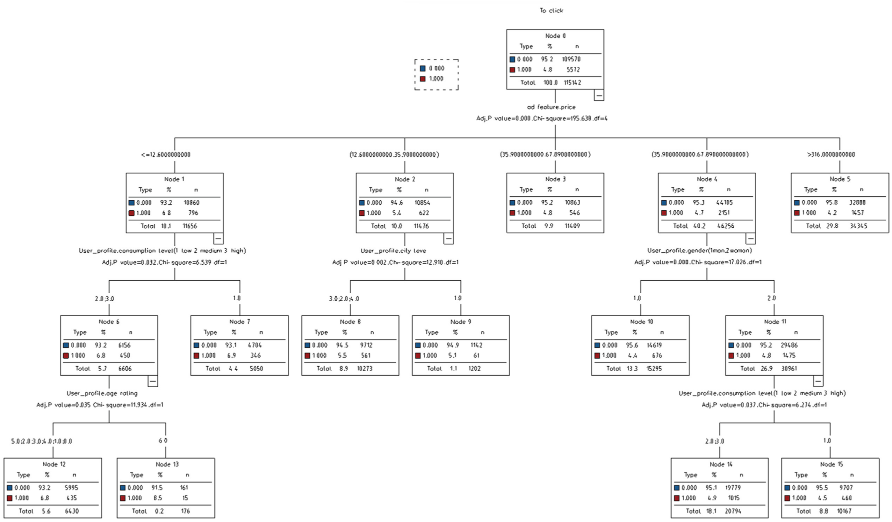

**Fig. 3.** Decision tree

## 4 Robust Analysis

In order to verify the robustness of the regression conclusion, some data were extracted from it and a robust regression was performed through the SPSSAU data analysis platform to check whether the regression results were robust. Robust regression is a method in statistical robust estimation that applies robust estimation methods to regression models.

It can be seen from the table that user_profile.gender (1 male and 2 females), user_profile.age rating, user_profile.consumption level (1 low, 2 medium, and 3 high) are used as independent variables, and clicks are used as dependent variables for Robust regression analysis (M estimation method), the regression coefficient value of user_profile.gender (1 male and 2 females) is 0.000 ($t = 9.068$, $p = 0.000 < 0.01$), which means that user_profile.gender (1 male and 2 females) will have a significant positive effect on clicking affect the relationship. The regression coefficient value of user_profile.age rating is 0.000 ($t = 9.363$, $p = 0.000 < 0.01$), which means that

**Table 6.** Robust regression analysis results (n = 23742)

| | Regression coefficients | Standard error | t | p | 95% CI | $R^2$ | Adjust$R^2$ | F |
|---|---|---|---|---|---|---|---|---|
| Constant | 0.000 | 0.000 | 13.712 | 0.000** | 0.000 ~ 0.000 | − 0.043 | -0.043 | F(3.23738) = − 326.945, p = 1.000 |
| User_profile.gender (1male 2female) | 0.000 | 0.000 | 9.068 | 0.000** | 0.000 ~ 0.000 | | | |
| User_profile.age rating | 0.000 | 0.000 | 9.363 | 0.000** | 0.000 ~ 0.000 | | | |
| User_profile. consumption level (1low 2medium 3high) | 0.000 | 0.000 | -2.552 | 0.011* | − 0.000 ~ -0.000 | | | |

user_profile.age rating will have a significant positive impact on clicks. The regression coefficient value of user_profile. Consumption grade (1 low, 2 medium, and 3 high) is -0.000 ($t = -2.552$, $p = 0.011 < 0.05$), which means that user_profile. Consumption grade (1 low, 2 medium, and 3 high) will click produce a significant negative impact relationship. The summary analysis can be obtained: user_profile. Gender (1 male and 2 female), user_profile. Age rating will have a significant positive impact on clicks. And user_profile. The consumption profile (1 low, 2 medium, 3 high) will have a significant negative impact on clicks. The final result is still stable.

To sum up, people's lives are full of traces of e-commerce after it has been strongly supported by national policies. The e-commerce industry started to enter infomercial and effect advertising, hoping to make their products reach users more directly through the form of ad placement. The search for customers is supported by big data and once the match is precise, then its commerciality is realized and the customer's willingness to buy becomes the final sales.

For advertisers, advertising effectiveness is the key concern. The Click-through rate is the first level of effectiveness of an advertisement. It reflects whether or not the user is interested in the ad, which in turn leads to a subsequent conversion. If there is no click-through, there is of course no further subsequent behavior. For the user, it is about the accuracy of the ad. The click-through rate reflects the user's attitude towards the ad. A high click-through rate indicates a better match between the user and the ad and is less likely to cause resentment.

## 5 Conclusion

To sum up, by using logistic regression and a decision tree to study the factors affecting ad clicks, the p-values of gender, age, and consumption level are all less than 0.01, so the difference is statistically significant. First, women click more frequently; second, middle-aged users have the lowest click frequency, and users over 60 have the highest click rate; and finally, users with the lowest spending are more likely to click on ads. At the same time, if you want to increase the click-through rate of an advertisement, these influencing factors cannot be discussed separately, and whether the product of the advertisement corresponds to the target audience should also be considered.

In terms of advertising operations, Taobao can provide different types of advertisements for different user groups to increase the click-through rate. The cheapest product advertisements are recommended to young groups with intermediate consumption levels; the middle-priced product advertisements are placed in second and third-tier cities; the more expensive product advertisements are recommended to female users in the middle and high consumption levels. This study shows that Taobao ad operators would be a worthwhile investment if they spend more time personalizing their ad placements. From the perspective of user characteristics, it is relatively easy to customize online ads for different user groups.

## References

1. L. Shan, L. Lin, C. Sun, X. Wang, Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization, Electronic Commerce Research & Applications, no. 16, 2016, pp.30-42. DOI:https://doi.org/10.1016/j.elerap.2016.01.004
2. S. F. Higgins, MD Mulvenna, R. B. Bond, A. Mccartan, S. Gallagher, D. Quinn, Multivariate testing confirms the effect of age–gender congruence on click-through rates from online social network digital advertisements, Cyberpsychology, Behavior, and Social Networking, vol.21, no.10, 2018, pp.646-654. DOI: https://doi.org/10.1089/cyber.2018.0197
3. R. D. Gopal, X. Li, R. Sankaranarayanan, Online keyword based advertising: impact of ad impressions on own-channel and cross-channel click-through rates, Decision Support Systems, vol. 52, no. 1, 2012, pp.1-8. DOI:https://doi.org/10.1016/j.dss.2011.04.002
4. Z. Gao, Q. Gao, Ad-centric model discovery for prediciting ads's click-through rate, Procedia Computer Science, no. 19, 2013, pp.155-162. DOI:https://doi.org/10.1016/j.procs.2013.06.025
5. M. S. Chtourou, J. L. Chandon, M. Zollinger, Effect of price information and promotion on click-through rates for internet banners, Journal of Euromarketing, 2002. DOI: https://doi.org/10.1300/j037v11n02_02
6. Z. Fang, K. Yue, J. Zhang, D. Zhang, W. Liu, Predicting click-through rates of new advertisements based on the bayesian network, Mathematical Problems in Engineering, 2014, pp.1–9. DOI:https://doi.org/10.1155/2014/818203
7. K.S. Dave, V. Varma, Learning the click-through rate for rare/new ads from similar ads, In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp.897–898. DOI:https://doi.org/10.1145/1835449.1835671
8. B. Yuan, J. Y. Hsia, M. Y. Yang, H. Zhu, C. Y. Chang, Z. Dong, C. J. Lin, Improving ad click prediction by considering non-displayed events, In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp.329–338. DOI:https://doi.org/10.1145/3357384.3358058
9. D. Zhu. Advertising click-through rate prediction based on cnn-lstm neural network, Computational Intelligence and Neuroscience, no. 12, 2021. pp.1-10. DOI:https://doi.org/10.1155/2021/3484104
10. G. Atkinson, C. Driesener, D. Corkindale, Search engine advertisement design effects on click-through rates, Journal of Interactive Advertising, vol.14, no.1, 2014, pp.24-30. DOI:https://doi.org/10.1080/15252019.2014.890394