From Online Browsing to Offline Purchases: Analyzing Contextual Information in the Retail Business

Simon Chan, Licia Capra
University College London
London, United Kingdom
{mhchan,l.capra}@cs.ucl.ac.uk

ABSTRACT

Accurate recommender systems can enhance consumers' shopping experiences. In retail and many other business environments, extra contextual factors are usually available for building even more accurate recommender systems. The influence of some factors is controversial in the industry. For instance, consumers' recent online exposure to products can decrease the chance of in-store purchase as consumers may choose to purchase products online. On the other hand, online exposure can be seen as an evidence of consumers' preference on products, which implies a higher chance of in-store purchase. The understanding of true influence is important for product recommendation in-store in this case. The question is how to evaluate the relevance and the influence of potential factors for prediction. Existing literature focuses on applying machine learning techniques to identify relevant contextual factors. While these methods are proven to be effective in some experiments, an alternative approach that can provide easy-to-interpret analysis on relevance and influence is preferred in many situations. The paper introduces a computationally inexpensive approach to conduct preliminary relevance and influence analysis for contextual information in retail business. Statistical techniques from medical research field are applied to analyze relationship between consumers' online exposure to retailer's e-commerce website, i.e., a contextual factor, and their offline in-store purchase decisions, i.e., the outcome to be predicted, based on a retail dataset provided by a large UK retail business with both online and offline presence. Unlike machine learning approaches, this analysis can be done even before a recommender system is built by using the proposed approach. This research further shows that the influence of this contextual factor depends on extraneous attributes, such as consumers' ages and gender. This papers serves as a preliminary step to analyze relevant contextual factors for building context-aware recommender systems.

Keywords

contextual information, odds ratio, stratified analysis, retail

CARS-2012, September 9, 2012, Dublin, Ireland.

1. INTRODUCTION

Recommender systems are increasingly important for retail businesses. Retailers commonly provide personalized product recommendations to consumers through various channels, for instance, web advertisement, email vouchers advertisement and in-store location-based mobile advertisement. Recommender systems try to predict the outcomes accurately so that they can recommend products according to the best expected scenario. Research showed that the use of proper contextual information in recommender systems can improve the accuracy of prediction in some situations [1]. For instance, contextual factors such as purchase intent have shown to improve prediction [14]. How could we, however, know the intent of purchase of a consumer who just enters a retail store without requiring additional user interaction? In an offline in-store environment, other than time and location context, the type of user context we can obtain without intruding consumers' experience is limited. To address this issue, this paper explores the possibility of using consumers' recent online behaviors on the retailer's e-commerce website to derive contextual factors for their in-store purchase decisions. The type of consumer behaviors that can be collected non-intrusively online is usually richer than those can be collected in-store.

In this paper, we consider consumers' recent online exposure to brands at a retailer's website as a potential, but controversial, contextual factor. Consumers' online exposure could imply consumers' tendency to purchase products online, so they are less likely to purchase them in-store. Oppositely, it could be seen as a user context that represents consumers' recent preference on brands, which implies a higher chance of in-store purchase. The understanding of the influence of this contextual factor is important for a number of recommendation scenarios. For simplicity, we focus on a specific scenario: A large retailer that has both online e-commerce website and offline stores presence wants to predict which product brands the customers are going to purchase when they enter offline stores. For a brand that a targeted consumer has browsed online recently, a recommender system can possibly consider three cases of influence: 1) If it is strong positively, the consumer will likely to purchase products of this brand in-store anyway, so no recommendation is needed; 2) If it is negative or if there is no influence, the chance of purchase in-store is not high; 3) If it is medium positively, the system may try to nudge the consumer to purchase products of this brand in-store. The design of this recommender system is part of our future work.

From the point of view of utilizing contextual information, this paper studies the relevance and influence between consumers' online exposure to brands at the retailer's website, i.e., a contextual factor, and the probabilities of their in-store purchasing decisions, i.e., the outcome to be predicted, for product brands, stratified by different consumer groups. We propose the use of statistical techniques to analyze contextual factor. Unlike machine learning techniques implemented in the literature, this approach is independent from the prediction model of the system. Besides, unlike traditional correlation analysis techniques, such as sign test and chi-squared test, our approach can estimate the influence of the factor on the probability of in-store purchase on products of brands. This information can possibly be used to improve the prediction model directly in our future work. A challenge of using statistical techniques, namely the issue that basic probabilistic measurement is sensitive to external noise, is presented by a numerical example in this paper. More robust techniques, such as odds ratio and stratified analysis, are then proposed. The dataset for experiments used in this research is provided by a large UK retail business. Ten product brands are analyzed. It is a 1-year anonymized records of loyalty card holders who have browsed products of the selected brands online at the retailer's website and of those who have purchased products of the selected brands in any store of the retailer in the UK. All data is collected in a real non-experimental setting. In our experiments, whether the consumers have browsed any product page of a targeted brand at the retailer's website within a month is regarded as a binary contextual factor and the outcome to be predicted is consumers' in-store purchase decisions. Results show that the influence of this contextual factor on the outcome to be predicted varies with consumers' attributes such as age and gender. The rest of the paper is organized as follows. Section 2 is a review of related literature about techniques that identify relevant contextual information in recommender systems. Section 3 describes the challenges of data analysis in a retail scenario. Solutions are then proposed to analyze contextual factors. Section 4 presents experiments conducted based on real retail dataset and section 5 is the discussion and future work.

2. RELATED WORK

Our literature review focuses on techniques that identify and evaluate relevant contextual information in recommender systems. The technical goal of a recommender system can generally be seen as the problem of predicting ratings, or any other outcome, for the items that have not been seen by a user [2]. The outcome to be predicted in retail recommender systems, for instance, may be consumers' purchase decisions instead of ratings. The use of contextual information in recommender systems have proven to improve the accuracy of prediction in some situations [1]. Context is a multifaceted concept that is defined differently in multiple research disciplines [3]. Various kind of attributes can be defined as context. For instance, there are context of users, context of items and context of interactions or situations [4]. Regardless of the definition, the selection of relevant contextual factors to be used in recommender systems is a critical issue. To deal with this issue, some literature applies machine learning techniques to identify relevant factors automatically. Decision trees and feature selection techniques are used to rank the relevance of user preferences and system settings in a news recommender system for accurate recommendations [5]. Another feature selection technique, Las Vegas Filter algorithm, has been applied in a more recent work to identify relevant factors [17]. A pre-filtering algorithm that pre-processes and selects contextual segments offline has also been described in details [3]. In [10], users are clustered based on the value of some contextual factors. The predictive accuracy of each cluster is then compared with the one of the whole dataset which is non-contextual in order to understand whether and where the performance improves. The advantage of this kind of algorithm is that factors are considered only in situations where contextual method outperforms the standard non-contextual algorithm. In current literature, relevance of contextual factors is measured based on their effects in the system's predictive accuracy. Recent research, however, shows that recommendation accuracy of context-aware recommender systems can be affected by conditions other than the contextual factors themselves, such as the task requirement and the overall number of items in the recommended list [15]. As a result, contextual factors may be omitted simply because they are not integrated into the system or the prediction model properly. Other literature proposes the use of statistical methods to evaluate the relevance of contextual factors. In contrast to a machine learning approach, a statistical approach is fast to compute and is independent from the prediction model implemented by the system. Not all type of data fulfills the assumptions of these statistical models though. For instance, Pearson Correlation Coefficient, or its binary form, Phi Coefficient, expects a linear relationship between the two variables. Paired t-test discussed in previous literature [1] is not suitable for binary data with binomial distribution. They are, therefore, not suitable for our scenario. Although other statistical methods, such as Sign Test and Chi-squared Test, could be suitable for our binary data, this paper presents an alternative statistical methodology that, not only evaluates the relevance of a contextual factor, but also estimates the influence of it on the probability of expected outcomes at the same time. By knowing the influence in probability, it is possible to make use of this contextual information to improve the prediction models directly in our future work.

3. STATISTICAL ANALYSIS

3.1 Problem Formulation

In this paper, we consider a retailer that operates both an online e-commerce website and physical retail stores. Consumers' recent online exposure at the retailer's website is the contextual factor to be evaluated. In particular, we define whether the consumers have browsed at least a page of a targeted brand at the retailer's website within a month as the contextual condition, $browse = 1$ if the condition exists, 0 otherwise. For illustration purposes, we assume that the outcome to be predicted is the purchase decision of any product of the targeted brand at any physical store (in-store purchase), which is a binary variable: $purchase = 1$ if purchased, 0 otherwise. In order to evaluate the relevance and influence between this contextual factor and the outcome, we need to compare the probability of in-store purchase of consumers who have browsed and of those who have not, i.e. $p(purchase = 1|browse = 1)$ and $p(purchase = 1|browse = 0)$. In a population of N potential consumers, we can construct a table to represent the online browsing and in-store purchasing situation of the dataset:

	purchase=1	purchase=0	
browse=1	a_{11}	a_{10}	a_{1*}
browse=0	a_{01}	a_{00}	a_{0*}
	a_{*1}	a_{*0}	N

where a_{11},a_{10},a_{01} and a_{00} are the number of consumers for the corresponding purchasing and browsing situations. $a_{*1} = a_{11} + a_{01}$ is the number of consumers who have purchased in-store, $a_{*0} = a_{10} + a_{00}$ is the number of consumers who have not purchased in-store, $a_{1*} = a_{11} + a_{10}$ is the number of consumers who have browsed online and $a_{0*} = a_{01} + a_{00}$ is the number of consumers who have not browsed online. A direct way to express the relationship is to compare the two probabilities with *relative correlation* (RC), where

$$RC = \frac{p(purchase = 1|browse = 1)}{p(purchase = 1|browse = 0)} = \frac{a_{11}/a_{1*}}{a_{01}/a_{0*}} \quad (1)$$

There is no correlation if $RC = 1$, the influence is positive if $RC < 1$ and negative if $RC > 1$. This approach, however, suffers from two problems when the data is collected from a non-experimental retail environment. First, RC is sensitive to the total number of consumers who have purchased and also to the total number of consumers who have not purchased. These two numbers, unfortunately, can be affected by external irrelevant factors, such as marketing campaigns or product promotions, which should be isolated from this analysis. This problem can be illustrated with a numerical example. Suppose the data looks like the following table when there is no sales promotion:

	purchase=1	purchase=0	
browse=1	5	500	505
browse=0	60	25,000	25,060
	65	25,500	25,565

Let us assume that a sales promotion successfully attracts new consumers to purchase and the number of purchase increases 10 times as shown in the following table:

	purchase=1	purchase=0	
browse=1	50	500	550
browse=0	600	25,000	25,600
	650	25,500	26,150

When all things being equal, a temporary sales promotion should not affect the relationship between the contextual factor and the outcome. In reality, however, $RC = \frac{5/505}{60/25060} = 4.14$ in the first case while $RC = \frac{50/550}{600/25600} = 3.88$ in the second one. In another words, RC is sensitive to the change of number of consumers who purchase (a_{*1}). This problem presence in many real-world environments since businesses can always attract new consumers to stores or website dynamically, which affects N, and thus a_{*1} and a_{*1} can be manipulated. An odds ratio technique to estimate RC that is insensitive to the change of N is proposed later in this paper. The second problem is the existence of extraneous attributes, such as age and gender, that potentially affect the influence of the targeted contextual factor on the outcome to be predicted. This problem occurs when an attribute is associated with the contextual factor and at the same time such attribute affects the outcome dependently or independently. This kind of extraneous attribute is called a *confounder* in the statistics discipline. Stratified analysis is proposed to evaluate the impact of possible confounding attributes.

3.2 Odds Ratio

Odds ratio is commonly used as a estimator of RC in medical and epidemiological research for case-control studies where disease cases are not easy to be obtained [6, 13, 7]. Similar to our problem, N is also adjustable in medical studies because the number of people with and without diseases in the dataset are determined by the design of the case-control studies artificially. In our case, OR can be calculated as:

$$OR = \frac{a_{11}/a_{01}}{a_{10}/a_{00}} = \frac{a_{11}a_{00}}{a_{10}a_{01}} \quad (2)$$

Identical to RC, there is no correlation if $OR = 1$, the influence is positive if $OR < 1$ and negative if $OR > 1$. Unlike RC, OR is insensitive to the row and column scaling operations of the data table. Using the same example above, $OR = \frac{5X25000}{60X500} = 4.17$ when there is no sales promotion, $OR = \frac{50X25000}{600X500} = 4.17$ as well when there is a sales promotion. OR is a good estimator statistically if a requirement is fulfilled: For the two groups of consumers, i.e. those who have browsed online and those who have not browsed online, separately, the number of consumers who have purchased in-store must be a small percentage (less than 10%) of the total number of consumers in the group. This requirement is reasonably fulfilled in most retail situations. Confidence interval (CI) is used to determine the reliability of the results. The larger the range of CI, the less reliable the result is. The CI of odds ratio [12] can be approximated with:

$$CI = \frac{a_{11}a_{00}}{a_{10}a_{01}} \exp\left(\pm z\sqrt{\frac{1}{a_{11}} + \frac{1}{a_{10}} + \frac{1}{a_{01}} + \frac{1}{a_{11}}}\right) \quad (3)$$

where z is the score of the standard normal distribution associated with the confidence level. $z = 1.96$ for a 95% confidence interval.

3.3 Stratified Analysis

Extraneous attributes, such as consumers' age and gender, potentially affect the influence of the contextual factor on the outcome. Stratified analysis is a computationally inexpensive solution to reveal their effects. This technique is commonly used in medical research when setting up control group experiments is not feasible and so the existence of extraneous factors is common [9]. It analyzes subgroups (strata) of the study population separately according to the attributes. For instance, two strata are created for the gender attribute: female consumers and male consumers. Odds rate is measured for each strata separately. Stratified analysis provides an independent view for each strata, each comes with its own odds ratio. The difference is then comparable among these strata. In addition, a common strata-adjusted odds ratio is estimated by Mantel-Haenszel (MH) method [11]. This adjusted value represents a weighted average of the stratum-specific odds ratio which is an approximation to the maximum likelihood estimation. According to [8], the formula of approximation can be written as:

$$OR_{MH} = \frac{\sum_{i=1}^{k} \frac{a_{11i}a_{00i}}{N_i}}{\sum_{i=1}^{k} \frac{a_{01i}a_{10i}}{N_i}} \quad (4)$$

where k is the total number of strata in an analysis and i represents one of them. For this Mantel-Haenszel method of

estimation to be accurate, the overall sample size must be large. [16] provides a more robust but complicated approximation method for data with small sample size. Confidence interval (CI) can again be used to indicate the reliability of the result:

$$95\%\ \text{CI for}\ OR_{MH} = Exp[(lnOR_{MH} \pm SE(lnOR_{MH})] \quad (5)$$

where

$$SE(lnOR_{MH}) = \sqrt{(\frac{\sum_{i=1}^{k}(\frac{a_{10i}a_{01i}}{N_i})^2 v_i}{\sum_{i=1}^{k}(\frac{a_{10i}a_{01i}}{N_i})^2})}$$

and

$$v_i = \frac{1}{a_{11i}} + \frac{1}{a_{10i}} + \frac{1}{a_{01i}} + \frac{1}{a_{00i}}$$

4. EXPERIMENT

4.1 Dataset

Our dataset, which is provided by a large UK retail business, is a 1-year anonymized records of loyalty card holders who have browsed the selected products online on the retailer's website and of those who have purchased the selected products in any store of the retailer in the UK. It contains 10,217,972 unique loyalty card holders and 2,939 unique products under 10 selected brands. There are 21,668,137 in-store purchase transaction records and 299,070 online browsing records. We associate consumers' online browsing and in-store purchasing behaviors with unique loyalty card numbers. All data is collected in a real non-experimental setting.

4.2 Experimental Design

This experiment investigates the relevance and influence between consumers' recent online browsing behaviors and the probabilities of their in-store purchase decisions for ten product brands carried by a large UK retail business nationally. These ten brands are selected randomly, some of them are luxury brands while the others are mid-range brands. We define whether a consumer has browsed at least a page of a targeted brand at the retailer's website within a month as the context of the consumer, $browse = 1$ if the condition exists, 0 otherwise. Odds ratio is used to compare the influence of this contextual factor on the probabilities of consumers' binary purchase decision of any product of the targeted brand at any physical store (in-store purchase). We pre-process the dataset to filter out consumers who have not visited any page at the retailer's website at least once in the past year. This process ensures that the remaining N consumers have at least successfully accessed the retailer's website recently. We start with a hypothesis that age and gender are two attributes of consumers that may confound the influence. We conduct monthly strata-specific measurement of odds ratio based on these two attributes for each brand. Practically, age and gender information is missing in *some* records. In each analysis, therefore, we analyze a population size of N_{age} or N_{gender} which represent the total number of consumers *with* age information or *with* gender information respectively. In these experiments, we calculate the monthly crude (unadjusted) odds ratio for each strata for each brand. If the odds ratio for a strata of a brand is X, it means that, in this strata and in this particular month, the probability to purchase at least one product of this brand in-store by consumers who have browsed at least a webpage of this brand online is X times higher than the probability for those who have not browsed so. We also calculate the monthly common strata-adjusted odds ratio as well as the 95% confidence interval (CI).

4.3 Results

Results of only three stratified odds ratios analysis are presented in this paper due to length constraint. All figures show that odds ratio measurements are well above 1, i.e., the influence is positive for all brands. It means that the probability to purchase at least one product of a selected brand in-store by consumers who have browsed at least a webpage of that brand online at the retailer's website is higher than the probability for those who have not browsed so. The values and patterns are different for each brand though, which means that the impact of this contextual factor of online exposure varies with brands.

Figure 1 represents gender-stratified analysis of brand A. The influence on female consumers is much stronger than the one on male consumers. An interesting discovery is that the odds ratio measurements for both genders follow a very similar up and down monthly pattern. Both strata have peaked odds ratio in February. This finding hints that time is a contextual factor that should also be considered in future work. Figure 2 shows that the odds ratio range of different age groups for brand B are separated clearly. The influence for consumers of age 18-25 is the highest while the one for consumers of age 26-35 is the lowest. It means that the probability for consumers of age 18-25 is higher than the one for consumers of age 36-45 and both of them are higher than the one for consumers of age 26-35. This finding implies that, for brand B, the age attribute itself can be correlated to consumers' in-store purchase decisions. Figure 3, on the other hand, draws a different conclusion for brand C. In this case, the odds ratio measurements of these age groups mixed together in a close range. There is no clear monthly pattern either. It means that age, gender and month are not confounding factors for this brand.

5. DISCUSSION AND FUTURE WORK

This paper derives a contextual factor from consumers' recent online browsing behaviors on the retailers' website for the prediction of their offline in-store purchase. A statistical approach is presented to conduct a preliminary analysis on the relevance and influence between this factor and the offline purchase decisions on brands using odds ratio and stratified analysis techniques. The initial uncertainty that consumers who browse online on retailer's website tend to

Figure 1: Odds Ratio by Gender (Brand A)

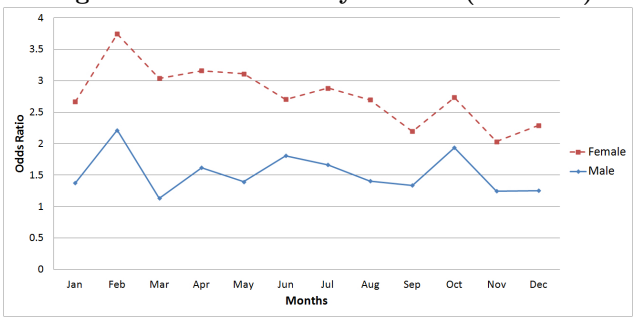


Figure 2: Odds Ratio by Age (Brand B)

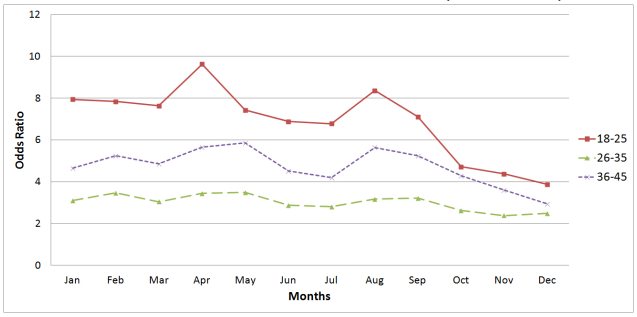
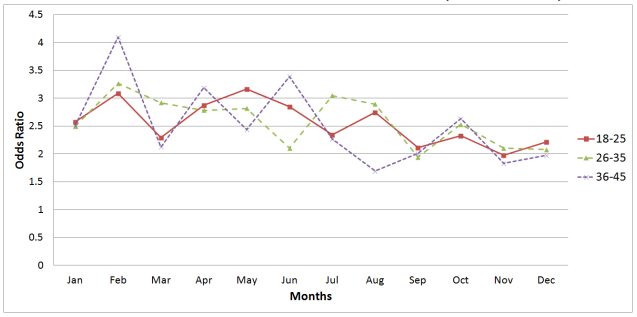


Figure 3: Odds Ratio by Age (Brand C)



purchase online and therefore they have lower chance to purchase in-store has been proven untrue for the brands we have analyzed. In addition, as expected, the influence of online exposure on offline purchases varies with brands and consumers' ages and gender. In our future work, the analysis for non-binary contextual factor will be illustrated. Besides the factor we have evaluated in this paper, it is interesting to see whether other relevant contextual factors can be derived from consumers' recent online behaviors for their in-store purchase decisions. Future work is to build a context-aware recommender system for in-store product recommendation based on these findings. We are interested in using the OR value directly to improve prediction. Also, a comparison of predictive performance of recommender systems using contextual factors selected by this approach and by existing machine learning techniques is part of our future work.

6. REFERENCES

[1] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145, Jan. 2005.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[3] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. *Recommender Systems Handbook*, pages 217–253, 2011.

[4] M. Bazire and P. Brezillon. Understanding context before using it. *Modeling and using context*, pages 113–192, 2005.

[5] A. Bellogín, I. Cantador, P. Castells, and A. Ortigosa. Discovering relevant preferences in a personalised recommender system using machine learning techniques. In *Proceedings of the ECML-PKDD 2008 Workshop on Preference Learning*, 2008.

[6] J. Cornfield et al. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6):1269, 1951.

[7] A. Edwards. The measure of association in a 2× 2 table. *Journal of the Royal Statistical Society. Series A (General)*, pages 109–114, 1963.

[8] W. Hauck. The large sample variance of the mantel-haenszel estimator of a common odds ratio. *Biometrics*, pages 817–819, 1979.

[9] D. Kleinbaum, L. Kupper, and H. Morgenstern. *Epidemiologic research: principles and quantitative methods*. Wiley, 1982.

[10] S. Lombardi, S. Anand, and M. Gorgoglione. Context and customer behavior in recommendation. *In RecSys09: Workshop on context-aware recommender systems (CARS-2009)*, 2009.

[11] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *The Challenge of Epidemiology: Issues and Selected Readings*, 1(1):533–553, 2004.

[12] J. Morris and M. Gardner. Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British medical journal (Clinical research ed.)*, 296(6632):1313, 1988.

[13] F. Mosteller. Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63(321):1–28, 1968.

[14] C. Palmisano, A. Tuzhilin, and M. Gorgoglione. Using context to improve predictive modeling of customers in personalization applications. *Knowledge and Data Engineering, IEEE Transactions on*, 20(11):1535 –1549, nov. 2008.

[15] U. Panniello and M. Gorgoglione. Does the recommendation task affect a cars performance? *In RecSys10: Workshop on context-aware recommender systems (CARS-2010)*, 2010.

[16] J. Robins, N. Breslow, and S. Greenland. Estimators of the mantel-haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, pages 311–323, 1986.

[17] B. Vargas-Govea, G. González-Serna, and R. Ponce-Medellín. Effects of relevant contextual features in the performance of a restaurant recommender system. *In RecSys11: Workshop on context-aware recommender systems (CARS-2011)*, 2011.