DATA SCIENCE
JOURNAL

Huang, W et al 2015 A Novel Trigger Model for Sales Prediction with Data Mining Techniques. *Data Science Journal,* 14: 15, pp. 1–8, DOI: http://dx.doi.org/10.5334/dsj-2015-015

PROCEEDINGS PAPER

A Novel Trigger Model for Sales Prediction with Data Mining Techniques

Wenjie Huang[1], Qing Zhang[1], Wei Xu[1], Hongjiao Fu[1], Mingming Wang[1] and Xun Liang[1]

[1] School of Information, Renmin University of China, Beijing, China
xhywcj@163.com, sunny890824@126.com, fuhj@ruc.edu.cn, mingmwang@263.com, xliang@ruc.edu.cn
weixu@ruc.edu.cn

Previous research on sales prediction has always used a single prediction model. However, no single model can perform the best for all kinds of merchandise. Accurate prediction results for just one commodity are meaningless to sellers. A general prediction for all commodities is needed. This paper illustrates a novel trigger system that can match certain kinds of commodities with a prediction model to give better prediction results for different kinds of commodities. We find some related factors for classification. Several classical prediction models are included as basic models for classification. We compared the results of the trigger model with other single models. The results show that the accuracy of the trigger model is better than that of a single model. This has implications for business in that sellers can utilize the proposed system to effectively predict the sales of several commodities.

Keywords: Sales prediction; Trigger model; Data mining; E-commerce

1 Introduction

Sales prediction is playing a growing and important role in many fields, such as economic forecasting, electric power forecasting, resource prediction, etc. Sales prediction is an important prerequisite for enterprise planning and correct decision making, allowing companies to better plan their business activities (Schroeder, Klim, Heinz, et al., 2010).

Sales prediction is important for offline businesses, especially car sales, real estate, and other everyday ventures (Baehr & Williams, 1968). The predictions are generally done by applying statistical methods, such as regression or the autoregressive–moving-average (ARMA) based on historical sales data. However, these methods only work for particular data. So many factors with complex interrelationships influence sales and probably include ones with a fair degree of uncertainty. Using data mining, we can identify potential models and development regularity from the masses of data. Therefore, an increasing number of researchers focus on how to make full use of data mining to process historical data and handle trends in sales prediction.

Previously, sales prediction research on online sales has been less studied because of the scarcity of real data on the subject. With the popularity of smart mobile terminals, E-commerce, especially B2C (Business-to-Customer), has been booming in recent years. Thus, the appropriate sales prediction method in the field of E-commerce to promote efficiency in online sales operations is a significant issue. In comparison with offline sales, e-commerce has its own sales characteristics, such as detailed basic user information and Web browsing information.

The paper proceeds with a new perspective that focuses on how to choose an appropriate approach to forecast sales with higher effectiveness and more accurate precision. The data for this paper have been provided by a well known, competitive Chinese online shopping company that is part of the B2C market in e-commerce book sales. We delve into a new research field, e-commerce, and apply real sales data to several classical prediction models, aiming to discover a trigger model that could select the appropriate forecasting model to predict sales of a given product. There is no doubt that it will effectively support an enterprise in making sales decisions in actual operations. It will enrich the theoretical basis and research methods in the background of big data.

The remainder of the paper is organized as follows. Section 2 discusses several existing classical sales prediction research methods and models, which are the theoretical background of our study. Section 3 presents the data analysis and processing method and then the trigger model. Section 4 verifies this model through empirical analysis. Section 5 contains the conclusion and discussion.

2 Literature Review

In this section, we will briefly review the previous studies on sales prediction and several classic prediction models. More than 200 kinds of prediction methods have been developed, which can be divided into two categories, subjective and objective methods.

The subjective prediction method is based on the experience of experts who judge and estimate. It is strongly subjective and flexible. Examples are the Delphi method (Linstone & Turof, 1975), the brain storm method (Tremblay, Grosskopf, & Yang, 2010), the subjective probability method (Hogarth, 1975), and so on. These methods use the experience of experts or the integration of predicted results. In contrast, the objective prediction method uses raw data to build models based on mathematics and mathematical statistics methods. It is reusable but not flexible. The objective prediction method includes mainly regression analysis and time series analysis. These methods use actual sales data, establishing a reusable model in order to predict future sales. Regression analysis methods include a simple regression model, a multivariate regression model, etc. (Kleinbaum, Kupper, Nizam, et al., 2013). The time series analysis forecast model includes the moving average model, the exponential smoothing model, the seasonal trends model, the autoregressive-moving-average model, the generalized autoregressive conditional heteroscedastic model, etc. (Box, Jenkins, & Reinsel, 2013)

Most conventional sales prediction methods introduce either factors or time series to determine the forecast. McElroy and Burmeister (1988) applied Arbitrage Pricing Theory into a multivariate regression model. Lee and Fambro (1999) used the autoregressive-integrated-moving-average model to do traffic volume forecasting. In 2003, Huang and Shih (2003) forecast short-term loans using ARMA. Tay and Cao (2001) studied time series forecasting. However, the relationship between influence factors or past time series data and sales prediction results is quite complicated. Therefore the predictions obtained from the aforementioned methods are often not satisfactory. As a consequence, many new intelligent model methods have recently been put to use in the area of forecasting; these perform better in terms of control and recognition. Some of the most representative new models are, for example, artificial neural networks (ANN) and support vector machines (SVM), the hot spots of forecasting research in recent of years. Kuo and Xue (1998) put forward a decision support system for sales prediction using fuzzy neural networks. Hill, Marquez, and O'Connor (1994) reviewed the artificial neural network models for forecasting and decision making. Cao (2003) combined SVM with time series for sales prediction while Gao et al. (2014) recommend extreme learning machine for sales prediction. Finally, Yuan (2014) proposed an online user behavior-based data mining method to predict sales in e-commerce.

However, the above research focused mainly on improving the accuracy of sales prediction via optimizing a single model algorithm or analyzing the factors that influence sales. For special cases, such as when the sales volume was zero, the single prediction model didn't perform well. In addition, most of the previous methods only predicted results for one object, for example, one kind of book's sales. In actual situations, the approach needs to cover a large scale of products. Thus, the traditional single model optimization method has significant limitations in sales prediction.

We built a trigger model system instead of depending on a single model algorithm. Based on data about factors that influence sales, "the system" triggers one of the prediction models discussed previously, leading to better prediction results than before. Also, our method can be used for a much larger scale of sales prediction. Therefore, we provide a new proposal for sales prediction research, which has been proven to be a significant improvement over past methods through our validation.

3 The Proposed Sales Prediction System

In this section, a trigger system is proposed for on-line sales prediction. An overview of the proposed framework is illustrated in **Figure 1**.

As can be seen from **Figure 1**, raw data are first manipulated into available forms, and then a trigger model is proposed to do the classification. Next, the classification result shows the potentially best prediction model for each SKU. Finally, by use of the most appropriate model, the prediction is accomplished.

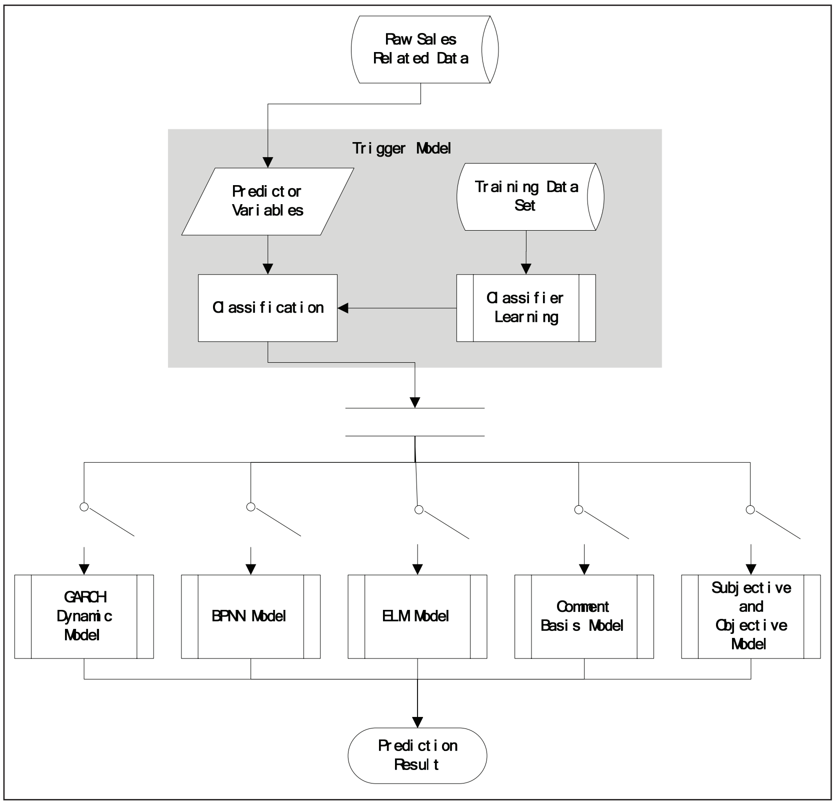

Figure 1: The framework of our Trigger Model System.

3.1 Data Collection

As it is difficult to collect a large quantity of data, we selected books to be our research objects. Sales records for books are more stable than for other commodities such as household appliances, which fluctuate with the seasons.

In the book sales market, there are significant differences between the sales data in entities trading and that in B2C. They both have common basic information (name, category, author, etc.) and trading information (time, sales volume, etc.), but special information, such as total attention, reference price, dealing price, and comments, is very important in the B2C platform. Total attention means the sum of "clicking", "searching", and "following" quantities. When customers search for a book, click a web page, or follow a book, the behavior is recorded. A product's total attention represents its degree of popularity. The reference price refers to the book's original price while the dealing price refers to the price the book actually sold for. "Comments" refers to customers' comments on the book.

The raw data we used was taken from 5 different database tables. For reasons of privacy, personal information was excluded. We used Excel to manipulate the raw data, aiming to filter redundant data and irrelevant attributes. **Figure 2** shows the entity relationship diagram for the raw data.

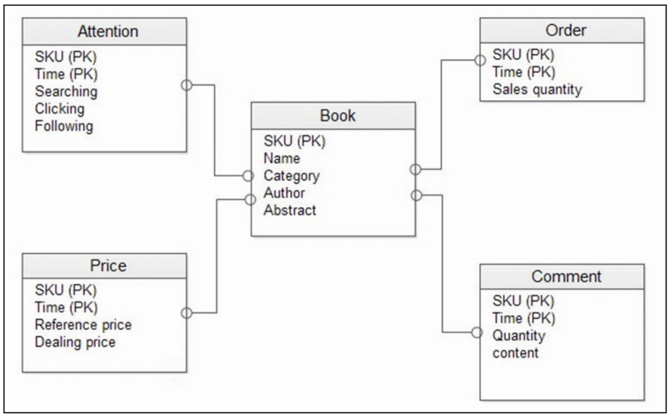

Figure 2: Entity relationship diagram.

3.2 Basic Prediction Models

Some basic models are the foundation of our research. We first divided each book's data into two parts. The data from Feb 1st to Apr 30th were used for training the prediction models, and the data from May 1st to May 31st were used for testing the prediction models. We used each basic model to do the prediction for the whole data set. Then we proposed a trigger model to choose which model performs the best for a certain SKU. The technique for the basic models is as follows.

The Artificial Neural Network is a mathematical model which imitates the distributed parallel information processing animal behavior characteristic of neural networks. It is widely used in the prediction field. The BPNN, created by MATLAB, is a feed forward neural network.

The Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model is a regression model specifically for financial data. It makes a further model for error variance, which is especially suitable for the analysis and forecast of volatility.

The Extreme Learning Machine (ELM) is an algorithm of NN, which is a generalized single-hidden layer feed forward network. It performs faster than any other model. Besides regression, it can also be used as a classification model. Thus, we use ELM not only as a basic prediction model but also as the classification model for our trigger system.

3.3 Variables for the Trigger Model

We selected 4/5 of the total dataset as the training set of our trigger model. The data were selected randomly; thus, the property features were evenly distributed. 3/4 of the training set was treated as the base of the trigger model while the other 1/4 of the training set was used to test the model's effects. **Figure 3** presents the detailed research procedure.

Before the experiment, we hypothesized that many properties, such as total sales and sales variance, were related to model selection. We made full use of all provided data and found that the following are closely related to classification.

- CV Sales (CVS): coefficient of variation for sales.
- CV Attention (CVA): coefficient of variation for attention.
- Sold Price Variation (SPV): the variation of sold price, which is different from the market price. It includes the price difference for promotions. We analyzed both, and the SPV gave better results.

The SPV calculation is shown as follows.

$$SPV = (S_{max} - S_{min})/S_a \tag{1}$$

where max and min S are the highest and the lowest sold price during the 3 months, and S is the average sold price.

Besides the prediction model we mentioned above, another prediction model can also be used as the basic model. With this model the prediction for the base data is done, and then the result is added to the system.

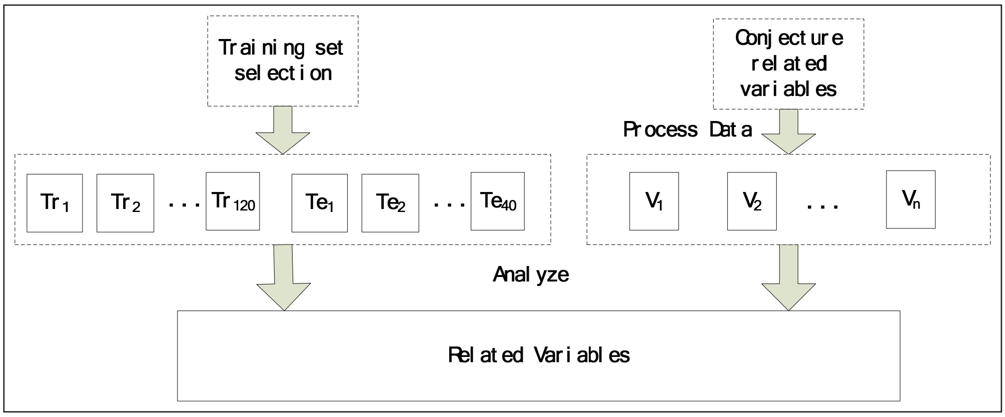

Figure 3: Trigger model research procedure.

4 Empirical Analysis

4.1 Data Description

The raw data were collected from the professional B2C online shopping market. Sales related data from February 2013 to May 2013 of 199 different books were provided. The books were randomly chosen from the top 1000 sales.

We identify several important and representative attributes to support our study. Data are processed in order by SKU and time. The involved attributes are listed in **Table 1**.

Attribute	Description
SKU	the unique identification number of the book
Searching record	how many times the book is searched daily
Clicking record	how many times the book's webpage is clicked daily
Following record	how many times the book is followed daily
Comments	due to time limits, we use only a daily number of comments in this research
Reference price	the book's original price
Dealing price	the book's actual sold price
Previous sales	past sales

Table 1: Details of attributes.

In particular, numbers for searching, clicking, and following data were sparse. Thus, we used total attention, the sum of the three parts, as degree of attention. All these factors may be used in prediction models.

Among the 199 SKUs, we selected 160 to be the training set of our trigger model. Because the data were selected randomly, the property features were evenly distributed. 120 items from the training set were treated as the base of the trigger model while the other 40 items from the training set were used to test the model.

We divided every book's data into two parts, the training set and the testing set. The data from Feb 1st to Apr 30th were for training the prediction models, and the data from May 1st to May 31st were for testing the prediction models. Next, we applied the training set data (160 books' attributes and sales data) and the 3 selected prediction models to train our trigger model in the appropriate environment. The details are shown in **Table 2**.

Model	Prediction Method	Variables	Environment
Model 1	BPNN Model	previous sales, total attention, daily average price	MATLAB
Model 2	GARCH Dynamic Model	previous sales, total attention, daily average price	EVIEWS
Model 3	ELM Model	10 Variables	MATLAB

Table 2: Details of prediction models.

The Mean Absolute Percentage Error (MAPE) was used to evaluate the performance of our proposed method. Because the books had significant sales volume disparities, MAPE could measure more accurately. The calculation of MAPE is shown as follows.

$$\mathrm{MAPE} = \left(\sum_{i=1}^{N} |\mathrm{F_t} - \mathrm{F_t}'| / \mathrm{F_t} \right) / N \tag{2}$$

where Ft is the actual sales, Ft' is the forecast sales, and N is the number of items.

4.2 Experimental Results

We used the testing set data (data for 39 books) in the trigger model's validation experiment. The experiment results are shown in **Figure 4**.

We can see from this figure that the total prediction trend of the trigger model is fairly flat. Most of the books' prediction results are less than 1. About half of the books' results are less than 0.5. After calculation, the average MAPE is 0.540844.

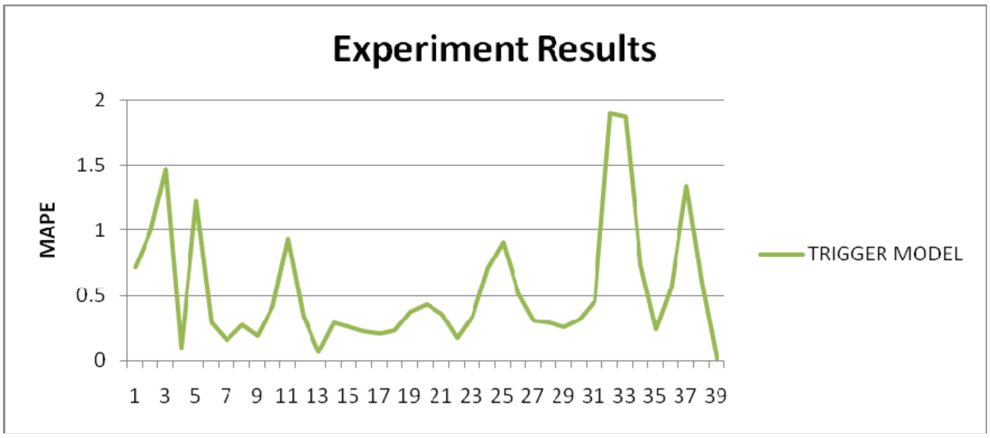

Figure 4: Trigger model prediction results.

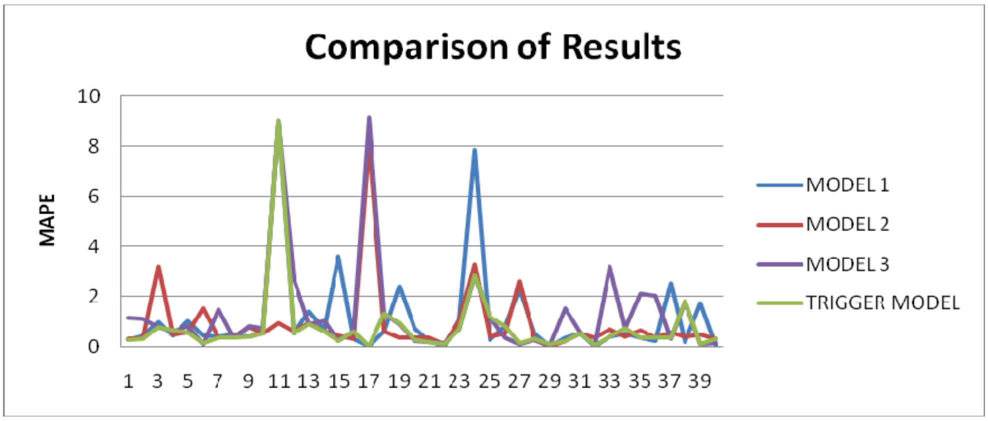

Figure 5: Testing data set comparison results.

MODEL	Model 1	Model 2	Model 3	ARMA	Trigger Model
MAPE(AVE)	0.99771	0.797165	0.985958	1.144498	0.540844

Table 3: Testing data set MAPE comparison.

4.3 Comparison

In this section, we contrast the predictive effects of Model 1, Model 2, and Model 3, and the trigger model. What is more, we select ARMA as the baseline model and introduce its prediction result into the result comparison.

As can be seen from **Figure 5** and **Table 3**, the baseline model (ARMA)'s result, which is 1.144498, is not ideal. Each of the models we select performs better than the baseline model. Model 1, Model 2, and Model 3 are 0.99771, 0.797165, and 0.985958, respectively. It is obvious that our trigger model performs the best among them, with a result of 0.540844. Compared with the other prediction models selected, the trigger model plays a prominent role in improving sales prediction accuracy in the field of large-scale product sales data.

5 Conclusions

This paper presents a new approach, building a trigger model for forecasting selection, to improve accuracy and efficiency in the area of e-commerce. We applied two typical forecasting models and several dimensions to the trigger model through training and testing the classification model with real sales data. Finally we obtained more accurate forecasting results than could be obtained by executing a single model. However, the study has some weak points. First, the amount of raw data is not enough. The forecasting accuracy needs to be increased further; moreover, we only selected two forecasting models to classify. More models need to be introduced to broaden the trigger model's application scope.

In conclusion, we present the idea of using a "trigger model" in the area of sales prediction. This focuses on the correlation of two subjects and ignores the causal relationship between them. It reflects the basic idea of "Big Data". In the future, the trigger model could be made smarter and more mature. If successful, the trigger model is likely to have a considerable impact on sales prediction.

6 Acknowledgments

This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No.71001103, 91224008, 91324015), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075), Beijing Natural Science Foundation (No. 9122013), Beijing Social Science Fund (No. 13JGB035), Beijing Nova Program (No.Z131101000413058), and Program for Excellent Talents in Beijing.

7 References

Baehr, M. & Williams, G. (1968) Prediction of sales success from factorially determined dimensions of personal background data. *Journal of Applied Psychology 52*(2), pp 98.

Box, G., Jenkins, G., & Reinsel, G. (2013) *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Cao, L. (2003) Support vector machines experts for time series forecasting. *Neurocomputing 51*, pp 321–339.

Gao, M., Xu, W., Fu, H., Wang, M. & Liang, X. (2014). A novel forecasting method for large-scale sales prediction using extreme learning machine. *The Seventh International Joint Conference on Computational Sciences and Optimization*, pp 602–606.

Hill, T., Marquez, L., O'Connor, M., et al. (1994) Artificial neural network models for forecasting and decision making. *International Journal of Forecasting 10*(1), pp 5–15.

Hogarth, R. (1975) Cognitive processes and the assessment of subjective probability distributions. *Journal of the American statistical Association 70*(350), pp 271–289.

Huang, S. & Shih, K. (2003) Short-term load forecasting via ARMA model identification including non-Gaussian process considerations. *IEEE Transactions on Power Systems 18*(2), pp 673–679.

Kleinbaum, D., Kupper, L., Nizam, A., et al. (2013) *Applied Regression Analysis and Other Multivariable Methods*. Cengage Learning.

Kuo, R. & Xue, K. (1998) A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights. *Decision Support Systems 24*(2), pp 105–126.

Lee, S. & Fambro, D. (1999) Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transportation Research Record: *Journal of the Transportation Research Board 1678*(1), pp 179–188.

Linstone, H. & Turoff, M. (Eds.)(1975) *The Delphi method: Techniques and applications.*

McElroy, M. & Burmeister, E. (1988) Arbitrage Pricing Theory as a Restricted Nonlinear Multivariate Regression Model Iterated Nonlinear Seemingly Unrelated Regression Estimates. *Journal of Business & Economic Statistics 6*(1), pp 29–42.

Schroeder, G., Klim, A., Heinz, G., et al. (2010) System for predicting sales lift and profit of a product based on historical sales information: U.S. Patent 7,689,456.

Tay, F. & Cao, L. (2001) Application of support vector machines in financial time series forecasting. *Omega 29*(4), pp 309–317.

Tremblay, C., Grosskopf, S., & Yang, K. (2010) Brainstorm: Occupational choice, bipolar illness and creativity. *Economics & Human Biology 8*(2), pp 233–241.

Yuan, H., Xu, W & Wang, M. (2014) Can online user behavior improve the performance of sales prediction in E-commerce? *IEEE International Conference on Systems, Man, and Cybernetics*, pp 2377–2382.

How to cite this article: Huang, W, Zhang, Q, Xu, W, Fu, H, Wang, M and Liang, X 2015 A Novel Trigger Model for Sales Prediction with Data Mining Techniques. *Data Science Journal,* 14: 15, pp. 1–8, DOI: http://dx.doi.org/10.5334/dsj-2015-015

Published: 22 May 2015

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS