2011 International Conference on E-business, Management and Economics
IPEDR Vol.25 (2011) © (2011) IACSIT Press, Singapore

A Multilayer Data Mining Approach to an Optimized Ebusiness Analytics Framework

Masoud Pesaran Behbahani[(1,2,3) +], Islam Choudhury[2] and Souheil Khaddaj[2]

[1] Azad University (IR) in Oxford, Oxford, UK

[2] School of Computing and Information Systems, Kingston University London, UK

[3]Azad University, Khorasgan Branch

Abstract. The aim of this paper is to introduce a new methodology for providing high level of business intelligence and optimization strategies for enterprises. The paper proposes a new optimized Ebusiness Analytical Framework (EBAF) and shows how the proposed business-oriented multilayer data mining methodology can be used in developing the framework. EBAF will serve as a practice blueprint for any Small Medium Enterprise (SME) wishing to enhance its customer relationship management and lead generation, and will give competitive advantages to SMEs that utilize this framework. EBAF presents a conversion model as a template for wide range of business models to create layers of required mining structures. The resulted multilayer mining structures will be served as the platform for applying the associated mining models.

Keywords: Business Intelligence, Data Mining, Ebusiness

1. Introduction

Business transactions are producing huge amount of data on a daily manner. Every website hit, every webpage visit, every payment are creating new rows in database tables. The current paradigm for using this data fails to maximize the business profit. The reason is that sometimes a decision improves the performance of a specific conversion metric, but exacerbates the overall conversion efficiency. As an example, a bank puts a service that loses money on the chopping block, inadvertently deteriorating the profitability problem by causing the best customers to look elsewhere for better service. Therefore, the first hypothesis in this research is that an integrated ebusiness framework which provides an information view of business activities yields more effective business intelligence than the existing stand-alone solutions. In this paper, a new optimized E-Business Analytical Framework (EBAF) is developed by a new approach. EBAF is highly scalable and can be extended to be even more useful in big enterprises. The approach is more business-driven, rather than current software-driven ones. In fact traditional views of business activities, like that of Kotler and Kelly [1] have mainly focused on the physical and human aspects of the organization. The information view of them started getting conceptualized with contributions from Holland and Naude [2] and Kumar Kar et al. [3] by emphasizing on marketing activities. The paper introduces and explains the new multilayer data mining approach. A multilayer data mining methodology needs multilayer mining structure. To identify the components of these mining structure layers, the organization should be analyzed. The resulted multilayer mining structure can be a platform for the multilayer data mining models. The multilayer data mining strategy can be generalized and utilized in any kind of organization to provide high levels of optimization.

2. Multilayer Data Mining Methodology

[+] Tel.: 00447407110444; Fax: 00442084172972. Email: Masoud@Kingston.ac.uk

For understanding EBAF multilayer data mining core, concepts of mining structure and mining model must be clearly defined. Mining structure generally specifies the number and type of attributes and optionally partitioning the source data into training and testing sets. Data mining structures can even contain nested tables to provide additional detail. In EBAF mining structures, EBAF Conversion model components are being used as data source to prepare the structure for the middle level of optimization. While mining structure stores information about the data source, mining model stores information derived from statistical processing of the data. Multiple mining models can be derived from a single mining structure. Each mining model includes metadata and patterns. The metadata includes a list of the attributes from the mining structure that is used to build the model, the description of optional filters that are applied during process, and the algorithm that is used to analyze the data. The main content of mining model i.e. patterns, can be in quite a few forms such as if-then rules that describe how objects are grouped together in a transaction, decision trees that can segment objects into groups, mathematical models with equations that describes patterns and can be used to forecast the future, and a set of clusters that define the characteristics of objects in the dataset.

In the proposed model which we name it multilayer data mining, a business is modelled into a multilayer data structure in which every layer can include several mining models involving various mining algorithms. The outcomes of each layer of mining models will be included in the mining structures of the next layer. Source data used in first layer data structures depend on the business model of the company. In a purchase business model, first layer of mining structures can be generated from the leaf level data such as company's products and services data, customers' demographic data, geographic data, behavioural data, recency/ frequency/ monetary transactional data, preferences/ interests/ hobbies data, psychographic data, propensity model, and media interaction data. In a bottom-up process, first layer of mining models are being produced. Based on these models, first level of business conversion rate influencers which are introduced as EBAF conversion model components can be individually optimized. At the second layer of mining structures, these components are used as datasets to optimize the whole performance of the organization. The point in EBAF optimization is that conversion model components such as traditional media advertising, email advertising, SEO, PPC and so forth which are individually optimized through previous data mining process are the inputs of next level of data mining practice. The inspiration in this methodology originated during optimization process of a multivariable business environment, and then developed by an academic research project. Multilayer data mining model is functional in enterprises with any size, and obviously bigger enterprises might need more layers to reach the desired point.

3. Middle Layer Mining Structures in EBAF

A five-stage conversion model including awareness, contact, engagement, conversion and retention is proposed to help identify mid-level mining structures in business domain. We desire each step to have the highest possible yield because they also act like funnels that each one feeds into the next. A simple overview of EBAF conversion model is presented in Fig. 1:

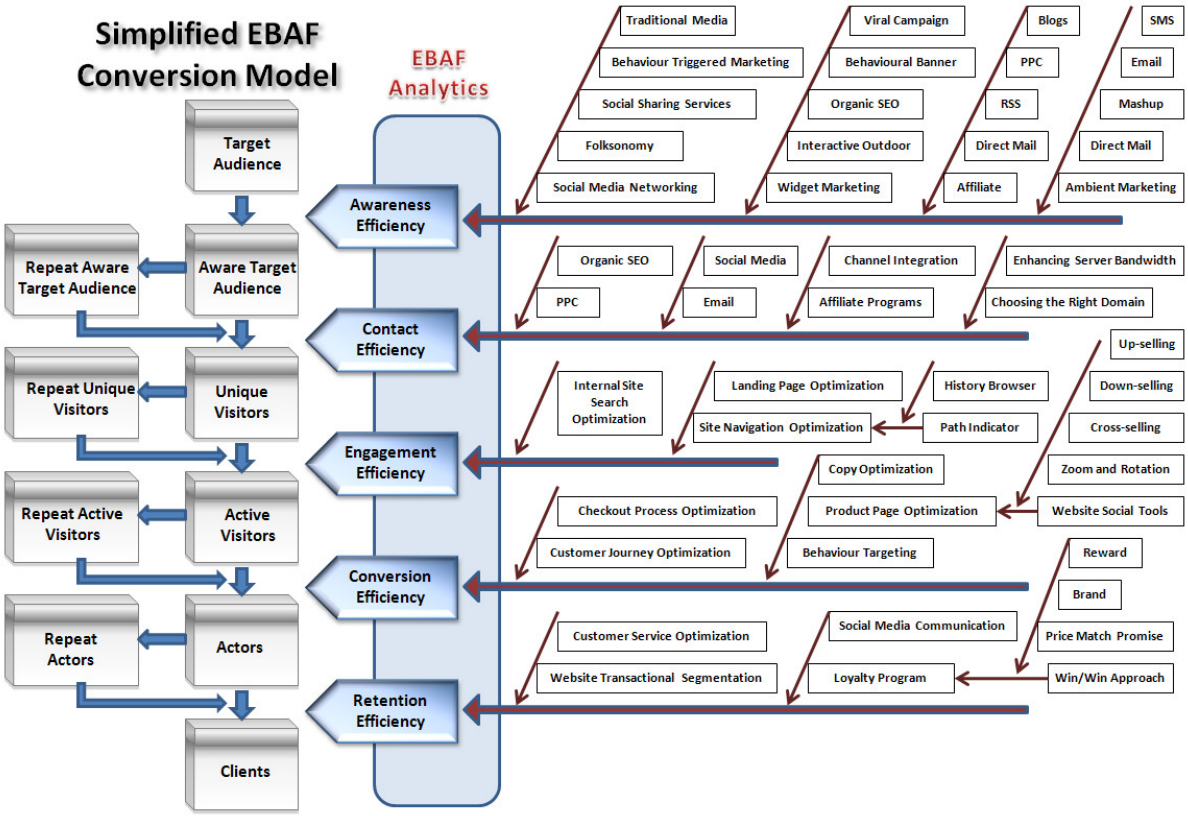

Fig. 1: Simplified proposed EBAF Conversion Model for Building Middle Layer Mining Structures

By the term awareness we mean the marketing activities that identified as influencers to aware the target audience of the ebusiness and can be used in middle layer mining structures. These activities target the first stage of the EBAF analytics model which is illustrated at the left side of the figure 1 and increase awareness efficiency. Damani and Damani [4] illustrate and describe pros and cons of 21 different awareness channels including traditional TV commercials, on-demand/ IPTV commercials, radio, Internet radio, podcasts, traditional outdoor, interactive outdoor, newspapers, magazines, direct mail, public relation, ambient or guerrilla marketing and street graffiti, traditional online banner, behavioural media banner and rich media banner that can even allow full shopping within the banner, organic search engine optimization, pay per click, affiliates, email to Internet list, email to 3rd party list, SMS, RSS. Still there are additional channels including widget and gadget marketing, micro sites, and viral campaigns. These are good candidates to be optimized through data mining models and then being included in mid level mining structures to provide more business intelligence.

By the term contact, we mean marketing activities that are identified to be influencers to ease it for aware target audience to hit the website or generally contact the business. There are several channels that an interested consumer can contact the business. In Internet channel, activities such as enhancing server speed and bandwidth, choosing suitable names for the site that can be easily guessed, using multiple names, affiliate programs and embedding hot links in sponsored websites, banner ads on search engines, and organic search engine optimization are of most importance. Channel integration approach is essential for success of the business. That means from the customer's perspective, all contact channels are gates to the same place and a multi-channel retailing is just a single retail organisation that has multiple touch points, in the form of call-centre, physical store, mail, interactive TV, main website, web service, kiosk, and mobile.

Many of the website visitors may leave it immediately after viewing the landing page. In fact, on average, a website has less than 10 seconds to capture the visitor's interest [5]. By the term Engagement, we mean identified activities to engage the unique visitors and prevent them from getting back. We define these engaged page viewers as active visitors. Although the operational definition of an active visit is to some extent dependent on the business model, the distinctive feature of it is some interaction between the surfer

and the webpage that could be as simple as viewing the offers or querying a database. There are some activities that may be useful to increase engagement efficiency, including internal site search optimization, landing page optimization, and site navigation optimization. Landing page optimization activities include upgrading web servers to increase bandwidth, keyword follow-through, multiple browsers and screen resolution testing, creating trust by illustrating press recognition and awards. Site navigation optimization activities include keeping navigation consistent, keeping navigation clear by using path indicator and history browser, prioritizing navigation links, and offering a variety of navigation themes.

Conversion deals with strategies and activities that are identified to be effective to persuade the active visitor to take the desired action and can be used in related mining structure. These activities include checkout process optimization, customer journey optimization, copy optimization, product page optimization, behaviour targeting, space management, and stock management. The last stage in the model is retention. Influencer datasets in this stage include loyalty program and social media communication and a mining model based on transactional segmentation might be significantly effective.

4. Experimental Study

In this section three different sample algorithms, Logistic Regression, Naive Bayes, and Clustering, are selected [6] to clarify and support the idea. In a real situation, every layer may utilize various different models for achieving better optimization in that layer, but here for simplicity only one algorithm is used per layer.

4.1. Logistic Regression in a First Layer Optimization

Logistic regression is a regression technique that is optimized for binary models in which dependent variable refers only to two variables. Examples of these yes-no models can be expressed as response to questions like: Is the customer loyal to the business? Is the customer a high value customer? Will the customer buy this product? In these cases, when the dependent variable refers to two values, standard multiple regression cannot be used. In this algorithm, a transformation of the dependent variable is going under prediction. This transformation is called the logit transformation. We mention the transformation as logit (p) and define its formula as: Logit (p) = Ln (p/ (1-p)) = Ln (p) – Ln (1-p). In this formula, p is the proportion of objects with a certain characteristic e.g. the probability for a customer to remain loyal to a company or brand. The logistic transformation of any number of z which like probabilities, always takes on values between zero and one, is given by the inverse-logit: Logistic (z) = $Logit^{-1}(z)$ = exp (z)/ (1+exp (z)) = 1/ (1+ exp(-z)). The value of the transform rapidly approaches to zero or one, making the result suitable for binary predictions. Now a Multiple Linear Logistic Regression can be defined as: Logit (p) = b0 + b1*X1 + b2*X2 + … + bn*Xn. The probability p of outcome variable can be derived by the equation:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 * X_1 + b_2 * X_2 + \cdots + b_n * X_n)}}.$$

As an example, the researcher evaluates the algorithm in optimizing a loyalty program. To minimize the cost, logistic regression algorithm is used to predict the outcome of the activity for each customer. Minimum requirement for determining threshold include false positive cost, false negative cost, true positive profit, and true negative profit. These values help the algorithm to find out the best threshold to maximize profit, as shown in fig. 2. This threshold later will be used to evaluate next customers and predict about them.

The figure also shows a table of relative impact of each influencer factor in prediction report. This table is just for getting a better understanding about the process. A small relative impact like having only one child shows that the related factors has only marginal effect on prediction. A zero relative impact like having no children shows that the factor does not affect the outcome. A factor with big relative impact like having more than one child is a strong indicator that the influencer is effective. The threshold obtained from profit diagram then can be used in companion with a calculator form for each object. Each factor in the form has a point assigned to it which has derived from the analysis by logistic regression algorithm. If the total score reaches the threshold, the marketing activity would be successful.

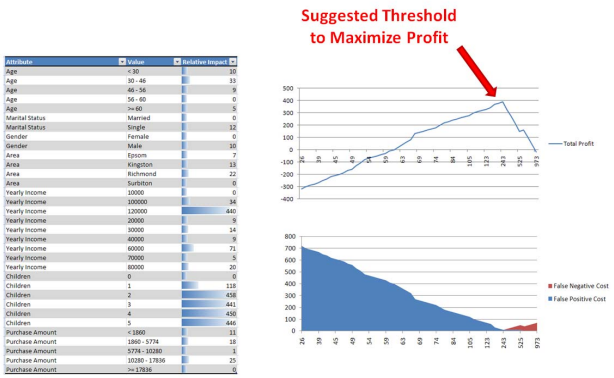

Fig. 2: Prediction by Logistic Regression Algorithm in First Layer

4.2. Naive Bayes in a Second Layer Optimization

In the second layer of optimization, the dataset includes items that have been previously optimized as targets. In a case study a classification algorithm is used for this layer. Given an object with attributes {A1, A2... An}, we wish to classify it in class C. The classification is correct when the conditional probability Pr(Ck|A1, A2... An) reaches its maximum among other classes:

$$\mathbf{Pr(Cj|A1, A2... An) = Pr(A1, A2, ..., An \mid Cj) * Pr(Cj) / Pr(A1, A2, ..., An)}$$

Assuming mutual independence of attributes for a given class C, simplifies the algorithm. By estimating all the probabilities Pr (Ai|Cj) for all attributes Ai and classes Cj, a new object can be classified to class Ck if the probability associated to it is maximized among the other classes:

$$\mathbf{Pr(Ck) * \prod_{i=1}^{n} Pr(Ai|Ck)}$$

One of main advantage of this algorithm is that it can easily handle irrelevant input attributes. Other advantages include robustness to noise and missing values. The input attributes through the case study including components of a low level of conversion model components e.g. Facebook marketing as a social media marketing or Google Adwords as a kind of PPC. A sample dataset for this layer is shown in fig. 3:

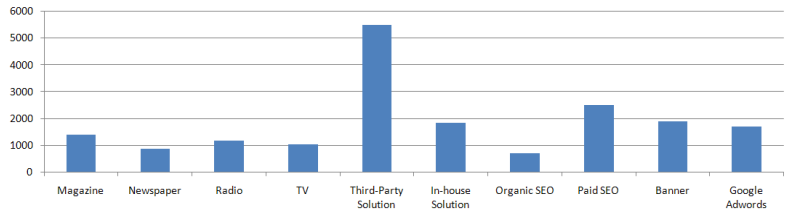

Fig. 3: Dataset Sample for Second Layer of Mining Structures

When the algorithm analyze the data for identifying key influencers, it creates predictions that correlates each column of data with the specified outcome, and then uses the confidence score for the predictions to identify the factors that are the most influential in producing the targeted outcome [7].

4.3. Clustering in a Third Layer Optimization

This layer includes components that previously optimized in second layer of optimization. As an example the dataset can contains information about customers' main awareness and contact channels. Before applying the clustering algorithm, it is useful to explore the dataset.

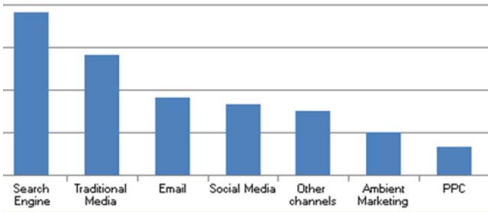

Fig. 4: Dataset Sample for Third Layer of Mining Structures

Initial results of applying algorithm shows the clusters and also shows that which object belong to which cluster. The category characteristics results show details about the similarities that were found in each category. The results also shows the relative importance that indicates how important the attribute and value pair is as a distinguishing factor for the category. We would then explore the generated customer segments in further detail to better understand the overall attributes of customers who belong to this category [8,9,10]. It is obvious that segments change over time as well as customers tend to change behaviours. Therefore, the algorithms should be recalculated on a more recent EBAF data source to check for correctness.

5. Conclusion

It appears from the preceding discussions and experimentations that the proposed multilayer data mining approach to an ebusiness framework may increase overall amount of business intelligence that an enterprise can gain. The concept contains a new methodology and its associated mining structures and mining models. The paper used this novel methodology and introduced an optimized framework called EBAF, to provide intelligence for SMEs and help them to gain competitive advantages. To support the theory, an experimental study consisting of three algorithms each applied on a different mining structure layer presented to provide a better understanding of the concept. The next step of this research is planned to integrate the methodology into multidimensional data and cube structures.

6. References

[1] P Kotler and K L Keller, *Marketing Management. 12th Edition*. New York: Prentice hall, 2006.

[2] P C Holland and P Naude, "The metamorphosis of marketing into an information-handling problem," *The Journal of Business & Industrial Marketing 19(3)*, pp. 167-178, 2004.

[3] A Kumar Kar, A Kumar Pani, and S Kumar De, "A Study On Using Business Intelligence For Improving Marketing Efforts," *Business Intelligence Journal*, pp. 141-150, 2010.

[4] R Damani and C Damani, *Ecommerce 2.0: The Evolution of Ecommerce*. London, United Kingdom: Imanco plc, 2007.

[5] J Palmer, Ecommerce Roadmap, Best Practices of Today's Successful Ecommerce Sites. Palmer Web Marketing, LLC, 2010.

[6] Q Yang and X Wu, "10 challenging problems in data mining research, " *Internatinl Journal of Information Technology and Decision Making 5(4)*, pp. 597–604, 2006.

[7] J Zhang, D K Kang, and A Silvescu, "Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data," *Knowlegde and Information Systems 9(2),* pp. 157-179, 2007.

[8] H Koga, T Ishibashi, and T Watanabe, "Fast agglomerative hierarchical clustering algorithm using Locality-Sensitive Hashing," *Knowlegde and Information Systems12(1),* pp. 25-53, 2006.

[9] R Jin, A Goswami, amd G Agrawal, "Fast and exact out-of-core and distributed k-means clustering," *Knowlegde and Information Systems10(1),* pp. 17-40, 2006.

[10] J R Chen, "Making clustering in delay-vector space meaningful," *Knowlegde and Information Systems11(3),* pp. 369-385, 2007.