Design of Smart Business-oriented Mining Engine

Neil Y. Yen and Jason C. Hung
School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan
Department of Information Technology, Overseas Chinese University, Taichung, Taiwan
neilyyen@u-aizu.ac.jp, jhungc.hung@gmail.com

Keywords: Data Fusion, Multi0-layered Fusion, Mining Engine, Planning-based Prediction, Smart Business.

Abstract: Keys to successful implementation of smart business require a wide spectrum of domain knowledge, experts, and their correlated experiences. Excluding those external factors – which can be collected by well-deployed sensors – being aware of user (or consumer) has the highest priority on the to-do-list. The more user is understood, the more user can be satisfied from an intuitive point of view, and thus, data plays a rather essential role in the scenario. However, it is never easy to achieve comprehensive understanding as the data requires further processing before its values can be extracted and used. So how the data can be properly transformed into something useful for smart business development is exactly what we pursue in this study. As a pioneer, three major tasks are focused. First, a mining engine is developed to be responsible for the universal collection of data which is primarily from real world, cyber world, and social world. Second, we go further into the fusion process of the collected data (e.g., the consumer purchase data shared by real-world company). A three-layer analysis and mining procedure is designed to enhance the mining engine through conventional RFM (Regency, Frequency, and Monetary Value) model and a set of fusion techniques. And in the end, we make planning-based predictions for a real-world company for expansion of the business interests.

1 INTRODUCTION

Smart business, by definition, indicates the ability to achieve goals which are set according to the development tendency of business (Watson et al., 2007). The key to successful implementation of the vision of smart business relies on a comprehensive understanding to the surrounded scenario in which wide spectrum of elements are concerned. Instances simply include vision of company, global economics situation, moving trends, targeted market and consumers, and etc. It is never difficult to find thousands of similar elements for consideration. But however, all these elements are useless unless they are well collected in form of data for further analysis (Cody et al, 2002).

Transforming data into meaningful and useful information (Parsons, 1996) that support the implementation of smart business is a long journey. Although rapid development in information communication technology makes it easy for data retrieval nowadays, sources where the data may be retrieved vary. The technology has also brought a tremendous change on our living world world – Hyper World (Kunii et al., 1996) – in which data is supposed to be from diverse channels and in unstructured formats. As such, how the data is retrieved, managed and processed becomes an open challenge when the issue concerning the comprehensive understanding is mentioned.

Collecting data, as much and complete as possible, is the first step to ensure enough and necessary information can be obtained. But this is, however, never taken as a practical way since decisions are made momentarily and sometimes only with limited information input. And thus, choosing one aspect as an entry point is a feasible action in the whole scenario. The end user, in general, is then considered a direct and intuitive way for this purpose.

Understanding the needs and preferences of users becomes complicated and requires more efforts than it used to be since users are spending more and more time on their activities like on-line shopping, interactions, communications, etc. in the Cyber world (Ma et al, 2011) via social media rather than face-to-face in the Real world. One of indispensable efforts is to collect their activity data in Hyper

World, mine their features, and discover their needs and preferences. This is a normal trilogy in the big data era. Data mining engine in this trilogy is essential for big data mining. In recent years, it has received great amount of attentions from academic society, industry, and business corporations. In particular, Google in 2011 released Data Mining Engine called Correlate, which enables users to find matching search trends. Oracle Data Mining Engine (DME) is the infrastructure that offers a set of in-database data mining functionality to its JDM (Java Data Ming) clients via a DME connection object. Amazon, the retail giant has been focusing on product recommendation engine, but recently, released Amazon Kinesis (Varia et al, 2014), which is streaming data real-time processing engine. It looks like almost data mining systems provide suites of data mining tools or software and put efforts on dealing with big and streaming data but how to efficiently meet application requirements and associated design approaches are not clearly mentioned or described.

Following the above-mentioned challenges and from the perspective of maximizing the benefits of business, this research pays the emphasis on the design of a universal framework for smart business support. This framework is instanced by a set of fusion techniques and a mining engine, and outcomes the planning-based predictions for a local company inside Japan. This smart business framework targets to provide services that best meet the needs of end users, retain the loyalty of existing users, and attract new users.

Meanwhile, descriptions to the proposed fusion techniques, data-data (D-D) fusion, algorithm-algorithm (A-A) fusion, feature-feature (F-F) fusion, data and algorithm (D-A) fusion, data and feature (D-F) fusion, and algorithm-feature (A-F) fusion, and data-algorithm-feature (D-A-F) fusion, will be elaborated. The input of the data mining engine is datasets and the output can be data, information, and knowledge, which are the input of Knowledge-Information-Data fusion engine or each of them can be used as a service (data as a service (DaaS), information as a service (IaaS), knowledge as a service (KaaS) directly to end user services.

Rest organization of this paper includes: Section 2 details the previous studies that relate to this study; Section 3 addresses the design of the fusion technique-based smart business framework; Section 4 gives a case study demonstrating the feasibility and preliminary results with the support of proposed framework; and Section 5 then concludes this paper and indicates potential extension of this work.

2 UNIVERSAL DESIGN OF FUSION TECHNIQUE-BASED SMART BUSINESS FRAMEWORK

Key to successful implementation of a smart business paradigm relies on many aspects. Excluding those that strongly require matured domain knowledge, the most common one is to satisfy the targeted audience, which is the user (or consumer as well), at the most. One significant instance is the service provision. A great amount of profits can be guaranteed if the service(s) to targeted consumers is right-to-the-needs.

A universal framework towards the implementation of smart business is then designed to this end. The proposed framework indicates an integrated approach, concerning the well transformation process from data to knowledge, together with a set of fusion techniques in interdisciplinary fields. A standard process that facilitates such process (i.e., diverse and unstructured data to well-defined information) is expected for future usage.

Figure 1 describes the image of our smart business framework. Five major portions are included:

(1) ***Data Acquisition*** is a universal entry for data collection. The data sources primarily contain the data in real world including weather information (e.g., temperature, atmosphere, quantity of rainfall, etc.), geographical information (e.g., coordinate, topography, etc.), human-related activities (e.g., supplies, equipment, manpower, etc.) that can be retrieved through deployed sensors; cyber data retrieved from social media such as a tweet/retweet from Twitter, a post (i.e., check-in, photo, message) on Facebook, an instant message via instant communication applications on smart devices; and other associated or related environmental data provided by third-party companies or organizations;

(2) ***Data Mining Engine*** is the fundamental component that connects the input data source and the follow-up data processes. It is composed by a set of mining techniques (e.g., statistics algorithms, practical machine learning tools) to meet all the necessary needs from users. It is responsible for the analysis of retrieved data, especially those multi-dimensional

data such as contextual, spatial, temporal, topical information with huge volume and high complexity, from available channels. Among all these methods, this engine is especially designed to incorporate possible fusion process, at the level of data, which may take place while specific requests are given by the users with heterogeneous data. Concerning the real-world situation, an integrated approach, i.e., the three-layer analysis and mining procedure, is proposed to cooperate with those existing, e.g., one-step, data fusion and mining algorithms. This approach, in particular, dynamically adjust the data for the fusion process, and select appropriate mining algorithms for execution;

(3) ***KID (Knowledge-Information-Data) Fusion Engine*** represents the second-step of fusion process in the framework. It especially concentrates on the fusion of processed input, which means, every stages of input, even knowledge, information, and raw data itself, may be fused in the case of necessary;

(4) ***Consumer Behavior Model*** contains a set of training and learning algorithms that continuously support the understanding of targeted users of a company. This model concentrates on the reuse of collected data correlated to users to shape the users and group them, depending on specific situations, as well. With several times of alternation, most explicit behaviors can be well predicted; and

(5) ***Open Platform*** is an universal portal that connects our proposed framework and external service, or data, providers. It enables the consumer behavior model to be built and grown, not only from the business point of view via the data mining engine and the data fusion engine but also from third-party contributions. It is designed to accept any trusted requests from the partners, and these accesses are also applied to enhance the proposed framework for better results provision.

A wide range of elements for the sake of better improvement in fusion techniques are considered while this framework is designed. This framework, and thus, identifies a general design to the whole scenario that take place in the implementation phrase of smart business. In other words, this framework is applicable to be further exploited to meet any specific purposes and cases.

In order to examine the feasibility, this paper especially concentrates on its usage to advance three essential issues, which are also the basics in the whole scenario, of smart business. The data mining engine of this framework is expected to lead preliminary solutions for a real-world retail company to:

a) **Find out the motivation of consumers and keep them connected:** The data, such as the personal information, preference, records of browsing and purchasing, activities on the Internet, device(s) used, and any possible activities on the social network (Moutinho, 1987), are collected and analyzed to provide better purchasing experience (i.e., personalized product and browsing on the website).

b) **Find out the elements that best attract consumers:** The above-mentioned data is further translated into information for self-training and learning processes. This information is expected to lead the element that creates the motivation of consumers. Monthly discount, free gifts, and jumping sales are taken as instance.

c) **Find out the thinking pattern of consumers:** For this purpose, the most efficient way is to allow seamless participation of consumers. No matter the comments or information shares over the social media or other related platforms by consumers shall be considered. With the trained information, the company may present new products that best meet the consumers, or a specific portion of them, to increase the business profits.

These three issues are taken as the primary concerns in data mining engine. Details of the design are introduced from the next section on.

3 DESIGN OF DATA MINING ENGINE

As stated above, this paper mainly focused on the design of data mining engine and its underlying fusion techniques for the planning-based product prediction for a real-world retail company though five core components are mentioned in the framework. As we know, collection of data with variety of types, huge volume, and high complexity

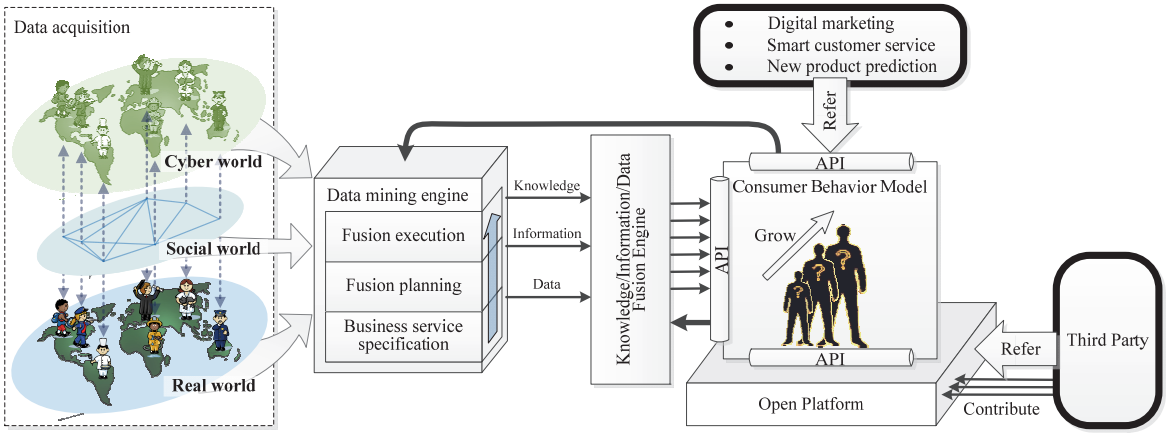

Figure 1: The Smart Business Framework.

is never easy and often requires corresponding hardware/software supports.

The fundamental concern to design a new type of data mining engine is that one single data mining method (or algorithm) or one-step mining procedure by conventional platforms or software tools are limited and not easy to meet all the possible needs in different phrases of service provision in smart business. Built-in fusion algorithm is important in the data mining engine. Our data mining engine, to this end, is featured by its dynamic mining and self-learning process with continuous input from the possible data sources. A three-level fusion technique for different business objectives and a set of fusion-based learning algorithms for prediction are developed.

3.1 Three-level Fusion Technique

The fusion techniques may be applied to three levels of fusion: data level (D-D fusion), algorithm level (A-A fusion), and feature level (F-F fusion). In addition, it may be necessary to merge the outcomes from three different level fusion, i.e., combinational D-A fusion, D-F fusion, A-F fusion, and D-A-F fusion. To a specific business objective or expected features from the data mining engine, an objective oriented fusion management agent is designated for fusion planning.

As shown in Figure 2, the data mining engine performs a 3-step process, i.e., the business service specification, the fusion planning, the fusion execution. It includes the following components: an objective-feature_label table, a dataset pool, an algorithm pool, feature_label discovery function, data_attribute discovery function, data_fusion function, and algorithm_fusion function.

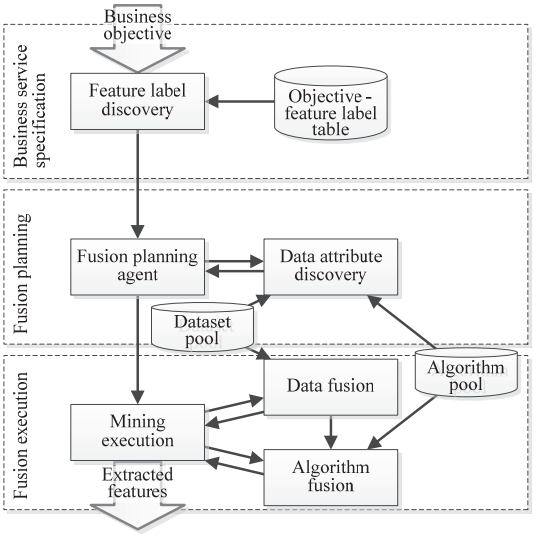

Figure 2: Process flow of fusion technique-based data mining engine design.

The data mining engine starts with a specified business objective, and retrieves its associated feature labels from the objective_label table. For a feature label or a set of feature labels, the fusion planning engine is to discover and schedule a sequence of algorithms in the algorithm pool triggered by the feature label(s). The data attribute discovery engine checks if the data attributes required by the triggered algorithms are contained in the dataset pool. If yes, data attributes are fused from different datasets as the input of the triggered algorithm. If not, unfound data attribute(s) are regarded as feature label(s), the above process is recursively repeated until all data attributes are directly found or indirectly created by applying for data fusion algorithm(s). Each triggered algorithm is

associated with its required data attributes as its input and the data attributes are associated with datasets in the dataset pool. As a result, a sequence of algorithms is retrieved and scheduled for execution.

The fusion planning agent is to trigger a sequence of algorithms in the algorithm-pool and discover their required data attributes are in datasets in the data-pool. The data attributes discovery is a backward chaining recursive procedure as given below. The resulting fusion plan implicitly indicates a suite of fusion algorithms (all or a part of 7-type fusion algorithms) triggered.

Algorithm 1 Backward Chaining Fusion algorithm

Inputs:
a set of feature labels: L
Outputs:
a set of data attributes: B
a plan as a collection of Algorithms: P

```
BCF()
  if L is empty ∨ recursive-depth = d then
    return
  else
    Get each feature label l in L
    Check each algorithm A in algorithm-pool
    if a(i) in datasets in data-pool then
      add a(i) → B
      add A → P
    else
      a(i) → L
      call BCF()
    end if
  end if
end
```

3.2 Fusion-based Learning Method

The fusion-based learning method is proposed to obtain correlations, especially the implicit patterns, between the preference of consumer and their periodical purchase. Most of the cases nowadays pay attentions on the discovery of user preference and further recommend potential products to the specific consumers so as to attract their interests. But this approach, however, faces to the user side. From the perspective of provider side (i.e., the company in this study), it is better to achieve comprehensive understanding of the cycle of purchase in order to present appropriate products to their consumers. For instance, a company will never sell heavy jacket in summer, or tank shirt in winter. What actions the company may take is to avoid unnecessary expenses (e.g., number of stocks, etc.) by prediction.

It is true that different approaches meet different kinds of consumer in running a business. For those low-royalty consumers, the recommending products may reach better profits than presenting periodical products there and waiting for consumers' notification to a company. But for those high-royalty consumers, a company may better stand a passive position with products that may attract the consumers to obtain better profits.

4 A CASE STUDY ON A RETAIL BUSINESS

4.1 Retail Data

The two types of retail data were collected through the online retail business and kept in the dataset pool. One is the customer profile which is composed of 23 attributes including customer ID, registration date and site, date of birth, region, gender, received mail magazines, reward points, etc. The number of registered customers is over 250,000. Another is the purchase record, which is composed of 38 attributes including order ID, item ID, color, size, order date and time, order site, customer ID, rough address, amount of purchase, etc. The number of records is

Figure 3: Retail Data at a glance.

over 450,000, about one year data from August 2011 to September 2012 as given in Figure 3.

4.2 Business Objectives

For any company, it is necessary to set their business goal in various levels from abstract to concrete business projects such as predicting, developing, adverting new products, attracting new customers, rewarding old customers, etc. Two business goals below are taken as case study in this research, however, due to the 4-page limitation, only case-1 is used to explain our fusion techniques based approach in this section.

Case-1: Awarding top customers: Objective feature label table lists top/best customers and associated customer value, customer segmentation and customer scoring as feature labels.

Case-2: Predict new product tendency: Objective feature label table lists new product tendency prediction and associated classification of the products in the top customer records as feature label.

4.3 Apply Backward Chaining Fusion Technique

Let us apply the backward chaining fusion to the case-1. As show in Figure 4, its associated two feature labels are top_instances and customer_value. Searching through the Algorithm pool, the algorithm, Extracting_top_instances() is triggered.

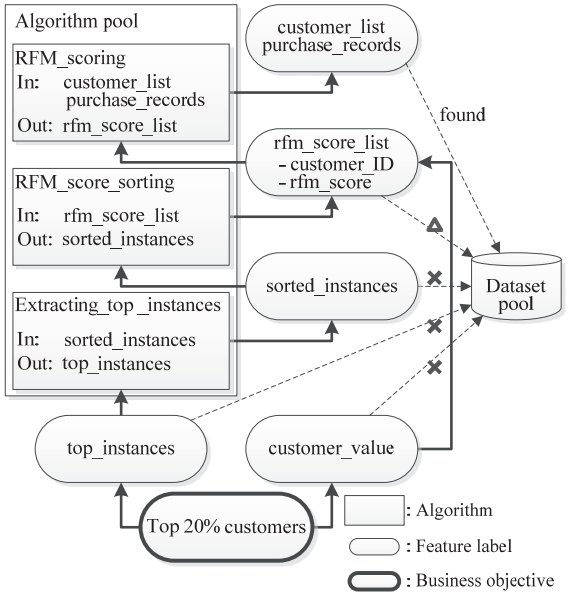

Figure 4: Applying BCF algorithm to case-1.

However, sorted_instances is not contained in the Dataset pool but it as a feature label triggers the algorithm, RFM_score_sorting() in the Algorithm pool. Again, rfm_sort_list including rfm_score are not in the Dataset pool but customer_ID can be retrieved from the Dataset pool while rfm_sort_list including rmf_score as a feature label triggers the algorithm, RFM_scoring(). Finally, required two attributes, customer_list and purchase_records are found in the Dataset pool. The backward chaining fusion algorithm terminated and a sequence of triggered algortihms, RFM_scoring(), RFM_score_sorting(), Extracting_top_instances(), with associated attributes, customer_list and purchase_records are the output of the fusion algorithm.

4.4 Analysis and Remarks

In implementing RFM_scoring(), it further requires R_scoring(), F_scoring(), and M_scoring() algorithms in the algorithm pool. These three algorithms can be called in parallel or sequence. Their results are fused, which is F-F fusion. Three algortihms, RFM_scoring(), RFM_score_sorting(), and Extracting_top_instances() are implemented in a sequence, which is A-A fusion. The customer_list and purchase_records are merged as RFM_scoring algorithm's input, which is D-D fusion. In other cases, other 4 types of fusion techniques may be applied.

5 CONCLUSIONS

This paper is mainly focused on our fusion technique based data mining engine which is the core component in the smart business framework. In this paper, 7-type fusion algorithms are listed, the fusion technique based data mining engine is described, the backward chaining based fusion planning engine as the heart of DME is explained. The case study on a practical retail business, a number of customer purchase record datasets is employed to show our design ideas and explain working principles of the proposed data mining engine.

Compared with other related work on data mining engine, our approach is the brand new in terms of building in 7-type fusion algorithms and having corresponding the backward chaining based fusion planning in the data mining engine.

ACKNOWLEDGEMENTS

If any, should be placed before the references section without numbering.

REFERENCES

T. L. Kunii, J. Ma and R. Huang, "Hyperworld Modeling", in proceedings of the *International Conference on Visual Information Systems*, pp1-8, Australia, February 1996.

J. Varia, S. Mathew, Overview of Amazon Web Services, http://media.amazonwebservices.com/AWS_Overview.pdf, January 2014, Amazon.

Jianhua Ma, Jie Wen, Runhe Huang, Benxiong Huang, "Cyber-Individual Meets Brain Informatics", IEEE Intelligent Systems , *Special Issue on Brain Informatics, Vol.26, No.5*, pp. 30-37, September/October 2011.

H. J. Watson, Barbara H. Wixom, "The Current State of Business Intelligence," Computer, 40(9), 96-99, 2007.

W. F. Cody, J.T. Kreulen, V. Krishna, W.S. Spangler, "*The Integration of business intelligence and knowledge management," IBM Systems Journal*, 41(4), 697-713, 2002.

N. Sun, J.G. Morris, J. Xu, X. Zhu, M. Xie, "iCARE: A framework for big data-based banking customer analytics," *IBM Journal of Research and Development, 58(5/6)*, 4:1-4:9, 2014.

S. Parsons, "Current approaches to handling imperfect information in data and knowledge bases," *IEEE Transactions on Knowledge and Data Engineering, 8(3),* 353-372, 1996.

L. Moutinho, "Consumer behaviour in tourism," *European journal of marketing, 21(10)*, 5-44, 1987.

Y. S. Wang, H.H. Lin, P. Luarn, "Predicting consumer intention to use mobile service," *Information Systems Journal, 16(2)*, 157-179, 2006.

J. C. Anderson, J.A. Narus, "Business marketing: understand what customers value*," Harvard business review, 76,* 53-67, 1998.

D. Boyd, K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information," *Communication & Society, 15(5),* 662-679, 2012.

B. Xiao, I. Benbasat, "E-commerce product recommendation agents: use, characteristics, and impact," *Mis Quarterly, 31(1)*, 137-209, 2007.

D. F. Duhan, S.D. Johnson, J.B. Wilcox, G.D. Harrell, "Influences on consumer use of word-of-mouth recommendation sources," *Journal of the Academy of Marketing Science, 25(4)*, 283-295, 1997.

W. H. Delone, E.R. Mclean, "Measuring e-commerce success: Applying the DeLone & McLean information systems success model," *International Journal of Electronic Commerce, 9(1),* 31-47, 2004.

S. M. S. Hosseini, A. Maleki, M.R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty," *Expert Systems with Applications, 37(7*), 5259-5264, 2010.

R. Kohavi, L. Mason, R. Parekh, Z. Zheng, "Lessons and challenges from mining retail e-commerce data," *Machine Learning, 57(1-2),* 83-113, 2004.

H. U. Bauer, K.R. Pawelzik, "Quantifying the neighborhood preservation of self-organizing feature maps," *IEEE Transactions on Neural Networks, 3(4),* 570-579, 1992.

W. Hoeffding, "A class of statistics with asymptotically normal distribution," *The Annals of Mathematical Statistics, 293-325*, 1948.