Multimed Tools Appl (2017) 76:11509–11530
DOI 10.1007/s11042-014-2387-2

# Towards next-generation business intelligence: an integrated framework based on DME and KID fusion engine

**Runhe Huang · Atsushi Sato · Toshihiro Tamura · Jianhua Ma · Neil. Y. Yen**

Received: 25 April 2014 / Revised: 29 August 2014 / Accepted: 30 October 2014 /
Published online: 7 December 2014


**Abstract** Advances in information technology prompt a tremendous usage growth of the Internet. Online activities, such as e-commerce, social interaction, etc., have drawn increasing attentions in regard to the provision of personalized services which require best and comprehensive understanding of users. As an approach, this study outlines a general framework based on human (or consumer) contexts for the discovery and creation of business intelligence. Three major portions are discussed. First, the collection of human contexts, including activity logs in both cyber and physical worlds, is modeled. Second, data analysis was performed via proposed mining algorithms that concern potential fusion at different levels according to situations and ultimate purposes. Third, sustenance of developed model is then concentrated. An open platform was developed to support the evolutionary process of human models, and to allow contributions (e.g., data sharing, accessing, etc.) from third parties.

**Keywords** Business intelligence · Fusion techniques · Consumer behavior model · Personalization · Service provision · Cyber-I

## 1 Introduction

Over the years, the Web has presented itself with tremendous extensibility and flexibility. It serves as a public environment (e.g., platform, testbed, etc.) for information to be delivered,

R. Huang (✉) · A. Sato · T. Tamura · J. Ma
Faculty of Computer and Information Sciences, Hosei University, Tokyo, Japan
e-mail: rhuang@hosei.ac.jp

A. Sato
e-mail: atsushi.sato.3r@stu.hosei.ac.jp

T. Tamura
e-mail: toshihiro.tamura.3k@stu.hosei.ac.jp

J. Ma
e-mail: jianhua@hosei.ac.jp

N. Y. Yen
School of Computer Science and Engineering, University of Aizu, Fukushima, Japan
e-mail: neilyyen@u-aizu.ac.jp

technologies to be developed and services (and applications) to be performed, which are mostly unanticipated when it was introduced at the first time. The Web has been transferring quickly, and a dramatic growth concerning the interdisciplinary intelligence and potential hybridization of technologies is envisioned through the ways people act upon it in the coming future.

A considerable number of media (e.g., social media, active media, ubiquitous media, etc.) has been developed to facilitate the sharing and delivery of information. One typical instance is E-Commerce. According to one annual report "Announcement of the results of the 2011 e-Commerce Market Survey" by the Ministry of Economy, Trade and Industry of Japan, the size of B2B (Business to Business) e-commerce (EC) market grew 8.6 % (about 8.5 trillion Japanese Yen) in 2010.

In addition, the market shares of B2C (Business to Consumer) e-commerce increased by 0.3 % over the previous year. These figures reveal the increasing frequency of online shopping, and to the present, over 50 % consumers purchase items over the Internet at least once per week. The development of products that meet the needs of consumers becomes important. In order to provide appropriate products (and services as well), a variety of resources, especially finance, is required to investigate behaviors (both explicit and implicit) within specific consumer groups. Achievement of essential personalization and adaptation relies on an adequate level of understanding of the needs and preferences of consumers. It is obvious that development of a successful product is led by comprehensive understandings of needs.

Recently, techniques for data mining have been widely applied in e-commerce in order to extract features of consumers and provide corresponding services [6, 11] and products. Taking Amazon as an instance, it offers a verity of recommendations to its users through an item-based collaborative filtering algorithm [17]. In addition, this algorithm has been revised, and enhanced as well, to serve specific purposes, where one typical instance of this algorithm is the integration of ontology and association rule for mining transaction logs of e-commerce [27].

Although issues concerning personalization and adaptation have been studied, few are left for investigation. One open issue is the isolated consumer management system. Each company owns separate system that causes the difficulty to collect complete user profile. That is, personalization on a specific aspect of a consumer is achieved (partial, not comprehensive). In this case, tedious profile-filling process is always required if a consumer expects a highly relevant outcomes.

To understand or especially to well understand an individual consumer, it is required to keep continuous monitoring of the consumer's purchasing activities on the Web and in the real world, and extraction of the consumer's features from collected data to form a more and more comprehensive model of the consumer's behavior. Three major parts are involved. First, a general model (CBM: consumer behavior model) is defined through the related contexts (e.g., browsing, clicking, inquiry, etc.) led while conducting the online purchase. Second, a dynamic evolutionary mining approach is employed to extract characteristics from the contexts, transform them into multi-dimensional set, and represent the knowledge structure. Third, an open platform is developed to share and collect information at different aspects. Any third-party applications (or services) are applicable to retrieve and share data via the platform.

The rest of the paper is organized as follows: related studies are discussed in Section 2; the proposed framework is introduced in Section 3; the features and applied approaches are addressed in Section 4; the scenario is given in Section 5; the implementation is in Section 6; Section 7 details the experiment results and findings; and the work is then concluded in Section 8.

## 2 Related work

Two issues concerning the main theme of this study are discussed in this section. The essential of data for prompting the business is first addressed, and the techniques that support user context analysis is then discussed.

(1) Data-Empowered Business Service Provision

Processing big data is computationally challenging problem. Ankur et al. [1] proposed a parallel implementation for distributed co-clustering and collaborative filtering algorithms on existing big data sets. Nakada et al. [22] presented a novel way to merge the static big data with new coming stream data based on MapReduce.

Discovering useful knowledge from big data is another important issue. Some studies have put attentions and efforts on combining mining algorithms, fusion knowledge, etc. for efficient and effective features extraction and knowledge discover. Wang and Tang [28] proposed a new data mining method based on k-means algorithm and density-based method for Geographic Information System. Milenova and Campos [21] proposed an analytic data based information fusion architecture, in which the data mining is performed inside the database and the resets can be further combined with spatial processing compounds to execute additional analysis.

Two main topics were addressed in this section. First, phenomenon of online marketing (or e-commerce) and its related issues were summarized. Second, potential supports from ubiquitous computing paradigm were then discussed.

Issues concerning online marketing have been discussed since the idea [35] was first put forward in 1979. A wide spectrum of applications and services as well has been developed to meet the needs of both retailers and consumers. One emerging topic is online advertising. At present, the Behavioral Targeting Advertising (BTA) model [5] is considered the main approach to attracting the attention of consumers. It makes use of collected behaviors, which are often implicit and the collection of which users themselves are not aware, to initially classify consumers into separate categories, and alter the classification in accordance with activities performed by the users. Recently, Chandramouli proposed a novel framework called TiMR (pronounced timer), that combines a time-oriented data processing system with an M-R framework in order to present advertisements through real-time streams naturally [30]. Through this model, services that may interest specific consumers can be easily located and delivered. However, an open issue - What kind of products and services do consumers need? - requires an equivalent perception of users that still remains a challenge. A situation may occur in which advertisements related to products/services that consumers are no longer interested in (e.g., a product that consumer bought already) are presented to consumers. On the other hand, the Consumer Development Model [3, 26] has been proposed to establish the directionality of products and services by modeling targeted users based on their activity logs, such as lifestyles and senses. However, this model may not provide relevant services to each consumer since it aims to provide services to a specific group by picking representative consumer(s). In this context, this research pays attention to the provision of services to individuals, and can help judge how best to answer the questions “What should the company do for the consumer?” and “What are consumers demanding?”

(2) Techniques for User Data Analysis

As supports to e-commerce through ubiquitous computing paradigm, scenarios regarding the mobility and flexibility are always envisioned. For examples, researches concerning the extraction

of data from real world via available sensors have been studied. This process helps detect the contexts around users, the knowledge discovery from large dataset, and the prediction of status of space (e.g., a room for shopping, etc.). Another instance is the framework, named Cyber-I [18], which refers to a comprehensive model of a human being in cyber space as a counterpart of himself/herself, and aims at facilitating the elderly-care process, healthcare process, and etc. With such kind of model, it becomes easier for companies to understand the intentions, purposes, and demands of potential consumers and furthermore predict how the actions of consumers, where services or products are easily evaluated and reduce the costs.

In order to deal with considerable number of data that generates from the end services, a conceptual scenario to data collection, management, and re-production was proposed [35]. A cycle for data processing in a hybridization of worlds, named Hyper World, (i.e., cyber world, real world, and social world) was addressed, which claims that collected data (via available sensors) shall be a continuous and circulated process while the data is being produced to a knowledge for the provision of services to users. With this process, efficient data management process and high quality services can be delivered.

In summary, an emerging model which collects the behaviors of users continuously, analyzes the data accordingly, and produces solutions correctly is envisioned in the development of a smart e-commerce environment. The concept of Cyber-I is put forward as the main actor and implemented as the core mechanism of this study.

## 3 Overview of smart business framework

The success of a smart business is based on satisfying consumers by providing them appropriate services so as to retain their regular consumers and attract new consumers. In order to achieve this, the most important point is to understand each individual consumer. Face-to-face interaction is a traditional and easy way of getting to know consumers. However, with the appearance of the Internet and advanced IT and Web technologies, online business has become more and more popular, reducing the amount of face-to-face interaction in the real world whilst increasing consumer activities in the Cyber digital world. We are actually living in a Hyperworld including the Cyberworld and the real world, and depositing huge of activity data in the Hyperworld.

Although face-to-face interaction is becoming less and less frequent in the business process, each consumer's footprint left in the Hyperworld provides valuable data for understanding consumers through data mining, information extraction, and knowledge discovery. This paper proposes a new business paradigm, i.e., a consumer behavior modeling based smart business framework. An overview (Fig. 1) of the proposed framework includes five parts: (1) data acquisition; (2) data mining engine; (3) knowledge/information/data (KID) fusion engine; (4) consumer behavior model; and (5) open platform.

The data acquisition is to collect three categories of data sources, shopping or other physical activities in the real world; online shopping or SNS social activities on the Web; other associated or related environmental data and information. The data mining engine is for mining processes through a combination of mining techniques (e.g., statistics algorithms and practical machine learning tools). The KID fusion engine is to provide a suit of fusion techniques such as algorithmic data fusion, information fusion, knowledge fusion and their hybrid fusion approaches, both to provide a specified service or advertisement to a consumer, or prediction of future products. The consumer behavior model is a digital description of a consumer in the real world, which is based on the concept of Cyber-I and particularly from the business viewpoint. The open platform is to enable the consumer behavior model to be built

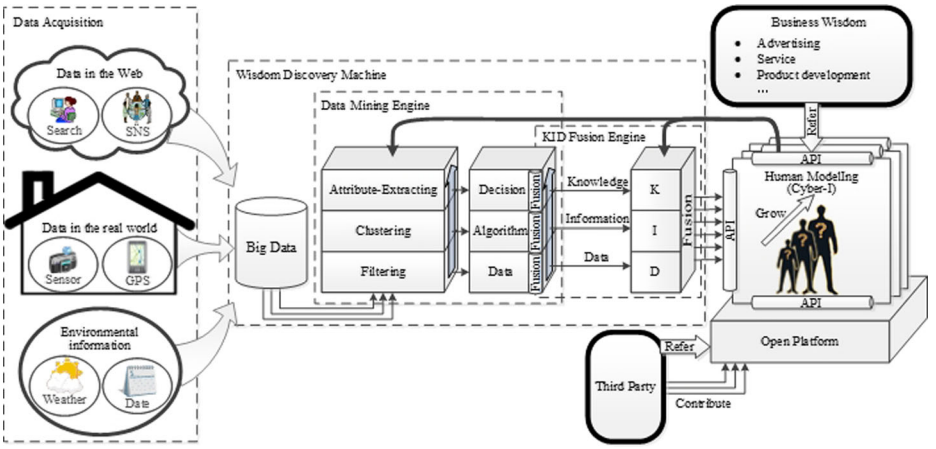


**Fig. 1** Overview of an integrated framework for next-generateion business intelligence

and grow, not only from the business viewpoint via the data mining engine and the data fusion engine but also from third-party contributions. Any third party can benefit from the framework by referring to the consumer model to provide better services to the consumer. Three goals of this business framework are to provide the best and most appropriate services to consumers, to provide personalized advertisement; and to predict products based on CBM.

## 4 Features of smart business framework

Five major components, namely Data Mining Engine, KID Fusion Engine, Growing Consumer Behavior Model, Open Platform, and Computational Psychological Model, of the proposed smart business framework are introduced in this section, where each of them is featured by its specific purpose and function in the framework.

### A. Data Mining Engine

This component can be interpreted as an engine (or a computational engine) for mining the data collected from different sources. The data can be imported from external, and existing as well, repositories, generated by daily users via various channels (e.g., social media), collected from deployed smart objects (e.g., sensors, mobile), or other available sources, and the aim of data mining engine is to re-create the value from such data.

Although conventional techniques for data mining have been discussed, they were defined and applied in various ways according to specific purposes. That means, one solution, i.e., mining technique (algorithm), for one problem is the general scenario except one conventional characteristic that these techniques are the automated, exploratory data analysis on massive data sets. Techniques for data mining fall under the general heading of unsupervised learning; other data mining activities, such as searching for specific patterns and supervised learning already presuppose some knowledge about the data. Many unsupervised learning methods can be identified as algorithms for filtration, classification, and abstraction that attempt to partition the data into classes of similar items. The underlying motivation is to approximate the empirical distribution of the data so that this empirical estimate can be studied as a demised proxy for the original data.

In this study, we primarily concentrate on multi-dimensional data collected from the Internet users while some business-related, especially from the viewpoint of B2C (Business to Customer), activities were taken places. Note that mentioned activities indicate the multi-dimensional, data such as behavior information, action logs, environmental contexts, etc., with variety of types, huge volume, and high complexity. In order to efficiently deal with the growing data, it is obvious that a single algorithm (or one-step approach) for data mining is no longer available. As such, a new approach, i.e., *objective-oriented fusion mining approach*, that automatically applies appropriate mining techniques, incorporates them, and optimizes them to meet the ultimate task. The outcome-driven (goal-driven) pre-processing includes:

1) *Data-filtering process* removes redundant or unwanted data from a collected dataset or datasets towards mining structures for expected outcome specification.
2) *Data-clustering process* classifies the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) and get ready subsets of a dataset for different data mining algorithms.
3) *Attribute-extracting process* then extracts attributes (or features) from the resulting clusters (subsets of a dataset) which are used for selecting a data mining algorithm or algorithms by matching their features.

The proposed approach can be briefly separated into three levels of fusion (as illustrated in Fig. 2), data level (data fusion), algorithm level (algorithm fusion), and result level (result fusion), and they were defined as:

1) **Data level fusion** is a multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats and their significance. It also identifies the process of organizing, merging and linking disparate information elements (e.g., map features, images, text reports, video, etc.) to produce a consistent and understandable representation of an actual or hypothetical set of objects and/or events in space and time. We employed the methods [7] for the business scenario (see Fig. 3). The process can be simply formulated by:

$$C(t)_k \leftarrow x(t)_i \cup y(t)_i \; z(t)_j$$
$$A(t)_i \leftarrow \cup x(t)_i \; y(t)_i$$
$$B(t)_j \leftarrow \cup x(t)_i \; z(t)_j$$

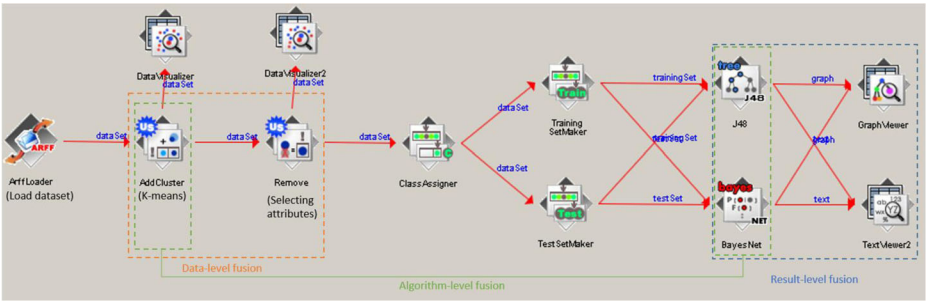


**Fig. 2** Illustration of three level fusion process in our data mining engine

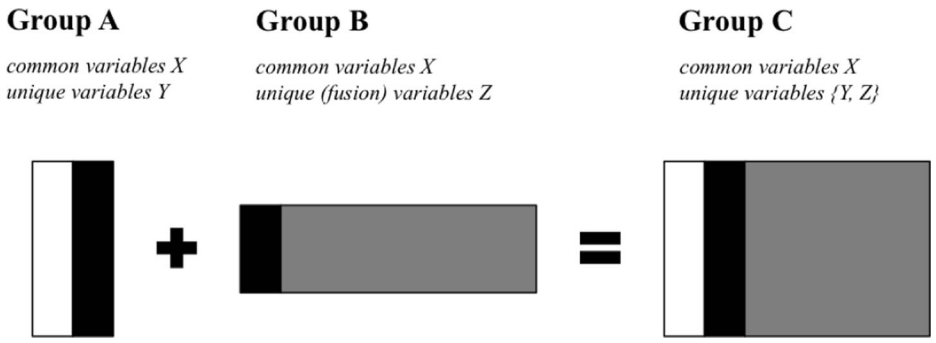


**Fig. 3** Concept of data fusion

where $C(t)_k$ , Group C, denotes the union set $k$ of Group A and Group B at time $t$ , $A(t)_i$ denotes the set of unique characteristics, $i$ , of Group A, and $B(t)_j$ is for unique characteristics, $j$ , of Group B at time $t$ . Note that the fusion process here is to create new data sets for further analysis. Data sets after fusion are considered new input and can be fused for other purposes.

<<Example>>

It is believed that "80 % of your sales come from 20 % of your clients" in Pareto principle (from Wikipedia). Let us take mining top 20 % regular customers as a data mining application, it requires the customer profile and the customer segment. However, what we have are two datasets: $D_1$ is the customer shopping record, having the attributes: ID, DATE, PRODUCT and PRICE, and $D_2$ is the costumer profile having the attributes: ID, NAME, SEX, BIRTHDAY and JOB as given in Fig. 4. It is obvious that only using $D_2$ cannot meet the requirement of the algorithm and simply combining $D_2$ and $D_1$ is also not a good solution. Here, a better solution is to use RFM (recency, frequency, monetary) analysis which is a marketing technique for determining quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). Therefore, the data mining engine at first performs RFM analysis algorithm by applying the customer purchasing history dataset $D_1$ to generate RFM dataset, further uses a clustering algorithm to segment the customers, and then the resulting customer segment is merged into the dataset $D_2$ as an attributes. Finally the generated dataset is

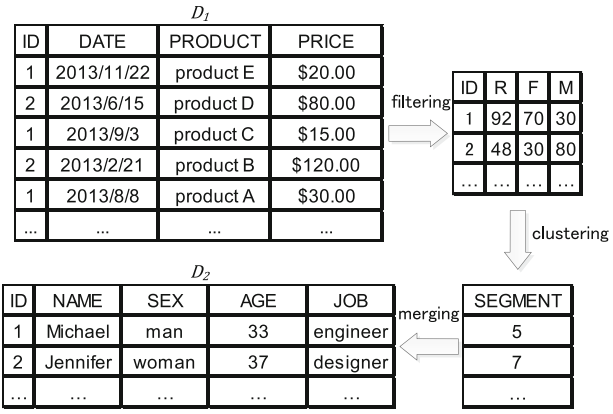


**Fig. 4** Example on data level fusion

composed of ID, NAME, SEX, BIRTHDAY, JOB and SEGMENT and sent to algorithm-level fusion process.

2) **Algorithm level fusion**, as its name implies, considers the fusion among algorithms which are applied onto the fused data sets so as to generate the decisions (or candidate solutions for specific tasks as well). For the purpose, types of fusion at this category are defined in accordance with different purposes. We simply separate the fusion process at this category into following categories considering the general business scenario. For user behavior analysis, the Bayesian network [8] is the base. Xiao's credit scoring algorithm [29] is added to predict donors, or companies, through the computation of features evolution of consumer attributes. The association rules, Apriori algorithm [24], are then employed to extract the patterns, which simply refer to special promotion activities in the business scenario, from features of consumers. Consumer's tendency is extracted through Principal component analysis [12]. Regression analysis [4] is used to understand what attributes are related with sales. Before the algorithm fusion is performed, it is critical to determine filtration, clustering, and extraction process for data towards a specific mining objective which are instanced as

   *Filteration process:* The noise filteration is the main theme of this process. Multi-level filterations are applied to transfer raw data into clean data;
   *Clustering process:* The data after filteration is then classified into different categories according to their unique characteristics;
   *Extraction process*: The features (or hidden patterns) are then extracted from the classified data sets.

   In addition, it may be necessary to merge or combine the outcomes from different mining algorithms for the value increase referring to a specific business objective. It is important to the value increase by taking into account context from consumer profile, environmental data and information, such as date, weather, events, etc. Whether it is necessary or whether it is important are governed by the objective oriented mining rule manager.

   <<Example-Cont.>>

   In order to find out what kinds of customers belong to the top regular customers, a single algorithm is difficult to achieve the task, it requires a number of algorithms to perform in both sequential and in parallel. As shown in Fig. 4, the algorithms, K-means, J48 and Bayesian Network are applied [10].

3) **Result level fusion** then determines a higher level fusion other than the above-mentioned fusion processes. It employes a set of classifiers to provide a better and unbiased result. The classifiers can be of same or different type and can also have same or different feature sets. Techniques for decision fusion [13, 15] have been studied, this study revises Kuncheva's work [14] as the core of decision fusion techniques, and Ho et al.'s work [9] for the classifier (i.e., produced decision candidate) optimization.

   As studied, features are vital and important cues for any successful classification or clustering task. Noisy, overlapping (based on class-wise distribution), distorted, and confusing features which may lead to inhibited performance in any case. Hence the challenge falls onto the selection of a proper set of features. Three steps to achive the fusion at this level includes:

   *Feature selection*: In the filter approach, the feature set is evaluated at once which is independent of any clustering algorithm or classifier. On the other hand, the wrapper method calls the clustering algorithm or the classifier for each subset evaluation to

find the final subset. While the filter method is unbiased and fast, the wrapper method gives better results for a particular clustering algorithm or classifier. Hybrid methods, such as complete, sequential, and random search algorithms, distance, dependency, consistency measure algorithms, and information gain, is an applied fusion of both filter and wrapper methods in our framework.

*Feature extraction*: The whole feature space is projected into another dimensional space for a better analysis of the features [2, 32]. In the reduced dimension space, the scatter of the clusters, with or without the class information, is observed to rank the reduced number of features. The number of dimensions can be reduced by principal component analysis (PCA) [33], kernel PCA (KPCA) [16], linear discriminant analysis (LDA), and independent component analysis (IDA).

*Feature combination*: Once the best feature subset is obtained from the original set, we can use the derived set or can derive a new feature based on two or more of the selected features for the task of classification. Based on this concept, there are two existing techniques of feature combination: serial and parallel combination. We will first calculate the weight of features (preprocessing phrase), and proceed the serial and parallel combination.

Above mentioned three process can be briefly formulated by:

$$H = \sum w_i v_i > T$$

where $w_i$ is user-defined weight, the $v_i$ is the set of generated decisions, and $T$ is a user-defined threshold and can be updated automatically through the analysis of user behavior. Through the formulation, the decision sets are expected to meet users' (including both business and end-user sides) goal if the value of $H$ is greater than pre-defined threshold $T$ , and vise versa.

<<Example-Cont.>>

Two results from J48 and Bayesian Network are merged, where Fig. 5 reveals the clustering results. Totally eight groups (i.e., segments) are obtained in this case, and each segment possesses its own attributes visualized by a graph (see "sex," "job," and "income"). Among these results, we can find that most of users in this group (Cluster4) are women (91 %); jobs (70 % of them) fall within Marketing, Real Estate, Food Service, Retail, Sales, and Telecommunications; and with yearly-income (over 90 %) above 584 in Japanese Yen. However, this only reflects the phenomenon of Cluster4, and the purpose of result-level fusion is to hybridize different characteristics among different segments and their outcomes insight corresponding to specific situations.

B. Knowledge/Information/Data (KID) Fusion Module

The KID (Knowledge, Information, and Data) fusion engine aims to generate appropriate outcome or decisions which can bring business organizer and business customer mutual maximum benefits with applying the fusion techniques to the following three categories of entity: data, information, and knowledge. They, in context, refer to:

1) Data: they are structured after going through filtering and clustering processes from raw data. The structured data should be efficient and effective for the re-structured, the reusable, and the scalable so as to be packed for provision of data service.

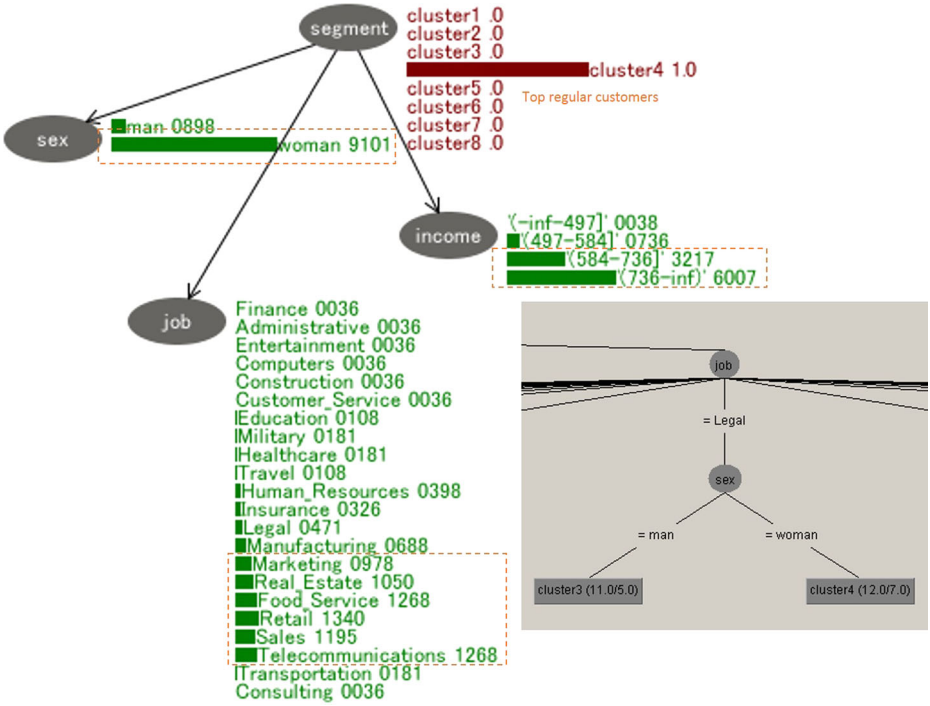


**Fig. 5** Illustration of customers for reward through the result level fusion

2) Information: it is a collection of organized and interpretable data being meaningful to and binding with a specific application according to its syntax and semantics. An information can be a service to a query request from a user or an application.
3) Knowledge: it is a collection of appropriate information such as facts, common sense truths, derived rules, etc. which are logical and rational to a specific task or domain problem solving. Knowledge as a service can be used for deriving new knowledge in a problem solving.

*Knowledge object* is a basic measurement unit for data mining in the scenario, where each knowledge object represents a specific target (e.g., consumer, product, etc.) in the dataset. This module interprets knowledge object to the knowledge representation that identifies precise and detail characteristics of consumers. Following methods [31] were applied to produce a new knowledge object from collected data.

1) *Enhancement process:* Strong correlations among collected data (e.g., characteristics in CBM) are analyzed to be additional desctipions of knowledge objects (i.e., metadata). It is applied to enhance existing knowledge objects based on predefined fusion rules.
2) *Extension process:* To obtain characteristics of consumers objectively, knowledge objects from data mining engine are applied to multi-dimentional attributes in CBM by fusing knowledge objects from data mining engine and multiple personal information in consumer behavior model. Certain attributes are updated from data mining engine based on predefined fusion rules.

Generated knowledge objects are then assimilated into the consumer behavior model to efficiently prompt the profiling/modeling the consumer. This process, focusing on the individual, improves the accuracy than conventional one for a group of consumers.

According to the definition of fusion engine, we go further to define the formulation of it. Let D be a data item, I be an information item and K be a knowledge item.

A data item is a set which consists of one or more elements. An information item is a data item which includes two data items, and mapping between them, which is expressed as $\mathrm{I} = (\mathrm{D}_1, \mathrm{D}_2, f : \mathrm{D}_1 \rightarrow \mathrm{D}_2)$ . A knowledge item is an information item which includes two information items mapping between them, which is expressed as $\mathrm{K} = (\mathrm{I}_1, \mathrm{I}_2, f : \mathrm{I}_1 \rightarrow \mathrm{I}_2)$ .

Let D be a set of data, I be a set of information and K be a set of knowledge. These definitions give an expression:

$$\mathrm{K} \subseteq \mathrm{I} \subseteq \mathrm{D}$$

In particular, if a set of data $D^k$ with the binary operation $f_k$ is a monoid, a data item of $D^k$ is called a data of type $k$ .

A simple example is given like: String (“aaa” + “bb”) + “c” = “aaa” + (“bb” + “c”) → “aaabbc”, (Associativity), “” + “aa” = “aa” + “” → “aa”, (Identity element).

We define 6 patterns (combinations) of the fusion $F$ ($F$ = {*KK, KI, KD, II, ID, DD*}). Detail descriptions on each of them are given below.

*KK* represents the self-fusion of knowledge objects, and can be simply expressed by $F : K \times K$ , where $K = \{k | k \in |K|\}$

*KI* represents the fusion of knowledge object and information object, and is expressed by $F : K \times I$ where $K = \{k | k \in |K|\}$ and $I = \{i | i \in |I|\}$ .

*KD* represents the fusion of knowledge object and data. It is considered a direct connection among data and its derived knowledge object, and is denoted by $F : K \times D$ , $K = (i_n, i_m, f : i_n \rightarrow i_m)$ and $D = \{d | d \in |D|\}$ , where $n, m \in \mathbb{N}$ .

*II* represents the fusion action at the level of information. Based on the data fusion results, two conditions (in-coming information and out-going information) that identify the directions of information are given. It is then formulated as $F : I \times I$ and $I = (d_n \leftarrow d_m)$ where $n, m \in \mathbb{N}$ . If the identifier $n$ is greater than $m$ than $I$ is considered as in-coming information, and vice versa (using $I'$ for the representation).

*ID* represents the fusion action between object at data and information levels. It can be formulated as $F : I \times D$ , where $I = (d_n \leftarrow d_m)$ and $D = \{d | d \in |D|\}$ .

*DD* then represents the fusion action at the level of data itself. The super identifier is applied to defined the order of given data objects. It is formulated like $F : D \times D$ , $D \leftarrow D^k$ where $k \in \mathbb{N}$ .

C. Growing Consumer Behavior Model

CBM grows dynamically according to continuous assimilation of activity logs and contexts related to consumers. Two processes are included: (1) CBM grows through the system’s data transmission process from raw activity and context data to extract details (e.g., behavior, hidden attributes, knowledge objects, etc.) of consumer. Key technologies in the data mining engine and data/information/knowledge fusion engine are applied; and (2) CBM grows via the open platform by the knowledge about the consumer from multiple perspectives obtained from the services (and applications as well) implemented by third parties on an

open platform. The experience and activity data of the consumer in exploiting services and applications provided by third parties is important resource for further well-understanding the consumer, which can be enabled by the open platform which enables the third parties' involvement and contributions. The model, the digital description of the consumer behaviors gets closer and closer to the real world's consumer through its growth.

D. An Open Platform

Data such as personal information, activities, etc. are kept by business companies in a profile-like file. CBM is applied with two main modules: an essence module and a growing module. As we know it is important to well understand a person from different aspects and angles, the growing module enables the consumer model evolving with the business process and obtaining contributions from the external third parties. Of course, a third party can share the consumer behavior model upon a mutual beneficial agreement and authentication.

These are carried out on the open platform as shown in Fig. 1. The consumer behavior model is shared with third parties on the open platform. A suit of APIs is provided for a third party to make contributions via providing services or applications to the consumer. A suit of APIs is provided for a third party to share the model for provide their own better and active services to the consumer. The more interactions the consumer in this business process and in the services and application provided by third parties, the more pieces of activity and behavior data left, the business paradigm can better understand the consumer through the data cycle from raw data to smart services. Although the open platform is a crucial part to implement CBM, challenges on privacy and security issues need further consideration.

E. Computational Psychological Model

In order to provide consumers personalized service/product recommendation, we attempt to not only understand what a consumer is interested in, but also take account of their emotional feeling how one feels by embedding a computational personality associated emotion model [23] in the consumer behavior model for consumer shopping behavior modeling.

The consumer behavior model is based on the concept of Cyber-Individual [18], which is the counterpart of real individual in the digital world and it is also so-called digital clone of real individual. A Cyber-Individual is a comprehensive digital description of its real individual. A consumer behavior model is one aspect of the Cyber-Individual from viewpoint of shopping.

The personality associated emotional model in the consumer behavior model takes one's personality into account. It describes both emotional state in PAD space [20] and plots one's personality in PAD space as well by converting personality's big five values [19, 25] to three coordinate values of a point in PAD space, To a same event, different consumers may be more or less differently triggered and have different emotional reaction, which is quite important for a success business to have slightly different ways for handling different consumers in services or an unexpected event by taking care consumer's emotional feeling and provide transparent services [34] to make them feeling satisfactory and happy.

## 5 The scenario

This section describes two business scenarios with the focus on provision of the personalized advertisements and services. This system runs the scenario simulations to demonstrate what and how personalized advertisements and services are concluded and delivered along with the

growth of the model. The scenarios as given in Fig. 6 are as follows: a woman, Rosemary was a housewife who registered in an electronic commerce and Internet based company which consists of the several businesses: online retail, travel agent, banking system, et al. like "Rakuten". After she registered, the system generates the corresponding base model.

A. Consumer Purchasing Behavior Modeling

One day, Rosemary bought the red wines made in the specific region of France on the site. After that she bought them regularly since she likes them. Starting with Rosemary's history action behavior input, the data mining engine works along the following processes to conclude like "she seems to like red wines made in France".

1) *Data level fusion:* A filtering and clustering algorithm is applied to her purchasing history data and generates a data package indicating her favorites.
2) *Algorithm level fusion:* A simple statistical analysis technique is used for extracting her frequent shopping pattern as information like "often to buy red wines". A Principle component analysis is exploited to discover the knowledge about her shopping tendency like "the probability of buying red wine made in France is high".

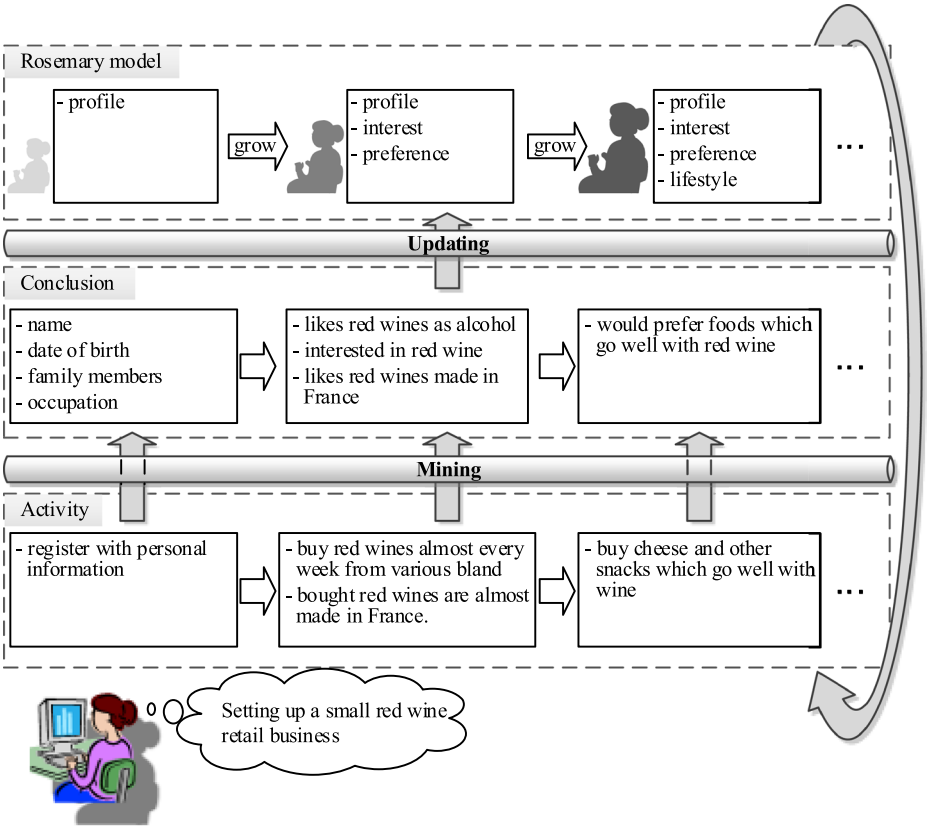


**Fig. 6** The growth of Rosemary model as her accumulated activity data

3) *Result level fusion:* A fusion mechanism is employed to integrate data, information, and/or knowledge from various data mining techniques so as to make a conclusion like "she seems like red wine made in France".

The outcomes (conclusions) resulted from the data mining engine are interpreted as the consumer characteristics to be added to the consumer purchasing behavior model for spiraling enhancement of the consumer behavior model.

The information, "she used to buy red wines", is enhanced with the predefined fusion rules like rule 1 and 2 in Table 1. As a result, her model is updated as "red wine is her favorite item in alcohol" and "she is interested in red wine". At the same time, the enhanced information is also extended with another information, "she buys red wines from various bland", and the predefined fusion rule like rule 3. As a result, her model is updated as "she would perhaps be a retailer of red wine".

B. Dynamic Evolutionary Model and Personalized Services/Advertisments in EC Site

At first Rosemary model is nothing more than her profile. But various attributes of her model are filled and updated as time goes with the continuing updated activity history. The personalized services and advertisements are provided accordingly referring to her growing model.

Personalized advertisements/services are actively pushed to individual consumers according to their current favorites, interests, and preferences which are extracted from their purchasing activities and other action behaviors.

In this scenario, after updating her interests in her model, the system sent her advertisement mails about the popular wines, the wine cooler and the cheese as snacks et al. Then, Rosemary clicked the cheese information link and buy some cheese and other snacks. As the result, the system guessed that she would prefer foods which go well with wine and added her tendency to her model. This makes the system sent more appropriate advertisements or services to her and she is getting more and more involvement with this system and the system know more and more about her and continues to enhance her behavior model. Therefore, her action behavior and the system services become a good and positive circulation. Thus, her behavior model is also continuously updated or grows.

## 6 Implementation of smart business framework

A prototype implementation of the proposed framework given in Fig. 7 is conducted, which includes:

**Table 1** An example of predefined fusion rules

| Rule ID | IF | THEN |
|---|---|---|
| 1 | ?item is included in ?category and he/she used to buy ?item | ?item is his/her favorite item in ?category |
| 2 | ?item is his/her favorite item | he/she is interested in item |
| 3 | he/she often to buy ?item and he/she buy ?item from various bland | he/she would perhaps be a retailer of ?item |
| … | … | … |

A. Data Mining Engine Module

This module contains several classes like *Data*, *Dataset*, *Pattern*, *DatamingAlgorithm*, *DataminingManager*, and abstract class, *AlgorithmFusionCenter*. *Data* objects takes the acquired data from the database and *Dataset* objects generate datasets from *Data* objects by the data fusion algorithm.

*DataminingManager* class provides a variety of methods to decide the policies according to the meta-mining rules for appropriate combination of datasets, mining task requirements, and mining algorithms. *AlgorithmFusionCenter* is an abstract class which provides some abstract methods for fusing data mining algorithms by taking consideration of datasets, algorithm features, and mining requirements. The data mining results are represented in a variety of patterns interpretable and usable by *KIDFusion* objects.

B. Knowledge/Information/Data (KID) Fusion Module

This module provides mechanisms for data, information, and knowledge fusions in 6 different fusion patterns. It includes two databases and two main classes, *KIDFusionManager* and *KIDFusion*. *KIDFusionManager* holds a set of fusion policies and decides KID fusion method according to business service requirement and consumer's situation. With the selected fusion policy from *KIDFusionManager* object, *KIDFusion* object actually performs the fusion process by invoking a fusion algorithm which takes data mining results from *Pattern* objects, human features from *AttributeSet* objects, and fusion task requirements from Fusion rules in the database.

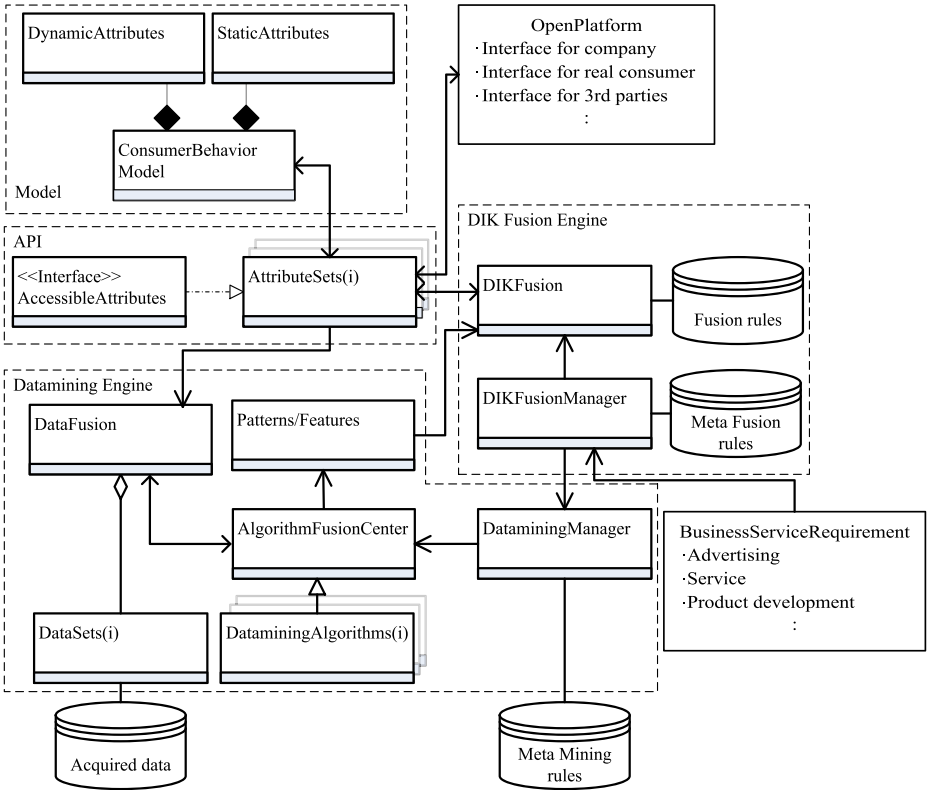


**Fig. 7** Class diagram of implementation for a smart business framework for smart business framework

C. API Module

API provides interfaces between the consumer behavior model and the KID fusion engine. It includes *AccessibleAttribute* interface and *AttributeSet* class. The former defines the specifications for the interface to the consumer behavior model. The *AttributeSet* objects are a set of attributes interpreting consumer's characteristics and behaviors for use in KID fusion engine or for contribution to the model.

D. Model Module

This module is built based on the concept of Cyber-I. It is composed of consumer's dynamic attributes and static attributes describing a consumer's features, in particular, purchasing behavior features. A Model object is the digital description of a consumer. StaticAttribute object describes a consumer's static features, relatively stable, not often changed or updated and DynamicAttribute object describes a consumer's dynamic features, frequently changed or updated.

## 7 Experiment

This section discusses the results of experiment on proposed framework. Except the conceptual modules and models, the core portion, which is the Data Mining Engine, was primarily focused. Details on results and findings are addressed:

<<Data Fusion>>

One objective factor to examine whether a higher accuracy rate can be performed by fused dataset than ordinary dataset (or raw data). This experiment was conducted with the dataset, in a visitor survey regarding their personal preference on online-purchasing, collected from the Internet (Rakuten E-Commerce website) within 2006–2012 among 800,000 users.

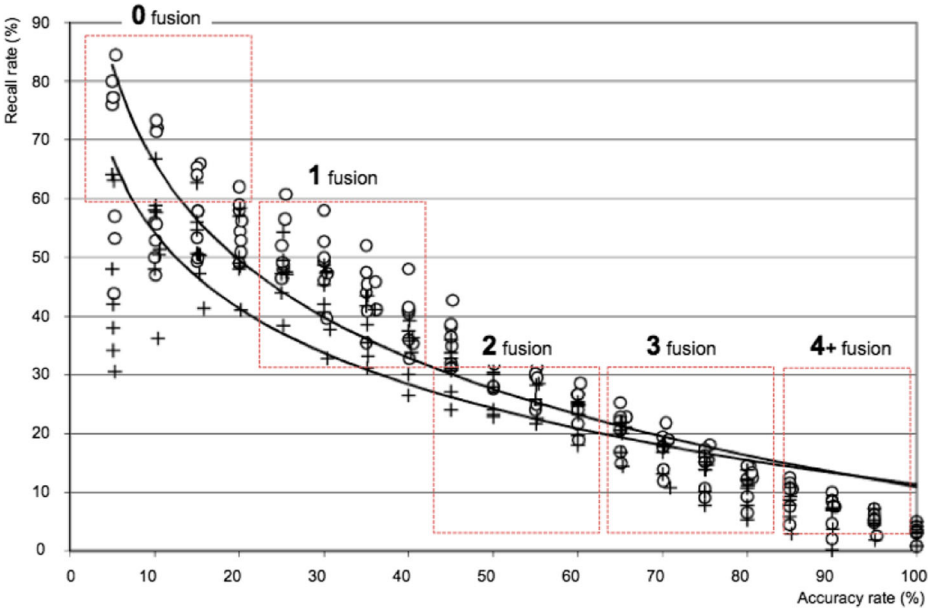


**Fig. 8** Accuracy-recall curve on fused dataset

**Table 2** Average and standard deviation on the accuracy of implemented algorithms (raw data)

| Algorithms | Accuracy (%) |
|---|---|
| Neural Network (NN) | 72.0±3.1 |
| Logistic Regression (LR) | 71.8±3.0 |
| Support Vector Machine (SVM) | 71.6±3.1 |
| Naïve Bayes (NB) | 70.6±3.2 |
| Decision Tree (DT) | 70.0±2.8 |
| K-Nearest Neighbor (KNN) | 66.1±2.5 |

In Fig. 8, the x-axis corresponds to the accuracy of selected features of dataset, and the y-axis corresponds to the recall rate of obtained feature data. Two lines represent the raw

dataset (lower line) and data with pre-processing (upper line) that exclude obvious noises (i.e., false positive in raw dataset). Higher accuracy is obviously appearing if more times of data fusion are performed, but on the other hand, the costs on computation may increase. Applying our method on present dataset, we may obtain that the fusion time at 2, depending on the attributes of data, may be the optimal solution in the data fusion process. That means, we can obtain necessary features from the dataset with reasonable computation costs (e.g., time and size of data table).

<<Algorithm Fusion>>

With the same dataset, the objective in this case is to provide the accuracy of existing algorithms, which refer to those classifiers for data processing, and the its appropriate usage situation. An overview of the average and standard deviation on the accuracies for implemented algorithms are shown in Tables 2 and 3. Table 4 then shows the average accuracy of algorithm fusion on fused dataset. The accuracy is defined as the percentage of correct classifications on the valid dataset. The differences for most algorithms are quite small given by the deviation in general cases.

Between these three tables, some interesting findings can be obtained. The first is about the applied weight functions to generate probability score for targeted dataset. It was found that the first four algorithms, Neural Network, Decision Tree, Naïve Bayes, and Decision Table, share some common attributes of data, no matter in raw or fused data, while conducting the computation. Second is that the model formula is valid over the entire attribute space but it results only local optimum. The decision tree, table and stump and nearest neighbor algorithms may divide the attribute space up into small areas, and that means, some external (but important for specific situations) attributes may be excluded during the computation. The third is about the performance of applied algorithms and their fusions. It is found that the performance of algorithm fusion on pre-processed dataset (or fused dataset in the above tables)

**Table 3** Average and standard deviation on the accuracy of implemented algorithms (fused dataset)

| Algorithms | Accuracy (%) |
|---|---|
| Logistic Regression (LR) | 73.7±3.3 |
| Neural Network (NN) | 73.4±3.4 |
| Decision Tree (DT) | 72.5±3.5 |
| Naïve Bayes (NB) | 72.2±3.6 |
| Support Vector Machine (SVM) | 71.7±3.4 |
| K-Nearest Neighbor (KNN) | 68.8±3.8 |

**Table 4** Average accuracy on algorithm fusion (fused dataset)

| | NN | LR | DT | NB | SVM | KNN |
|---|---|---|---|---|---|---|
| NN | | 76.90 | 76.45 | 76.36 | 76.07 | 74.41 |
| LR | 77.01 | | 76.63 | 76.40 | 76.10 | 74.28 |
| DT | 76.50 | 76.61 | | 75.49 | 74.34 | 73.70 |
| NB | 76.32 | 76.42 | 75.44 | | 73.86 | 72.80 |
| SVM | 76.26 | 76.03 | 74.37 | 73.87 | | 72.33 |
| KNN | 74.40 | 74.25 | 73.65 | 72.82 | 72.33 | |

reached a higher score than algorithms were separately applied although a significant improvement was not revealed (say +0.6~3.4 %).

## 8 Conclusion and remarks

A human-centered approach to understand the needs and preferences is the key to create more opportunities in various realms. For business, it becomes even important taking this matter into consideration to attract potential consumers and keep them for the determination of business profits. To collect consumer needs and behavior features make an e-commerce based incorporations able to make appropriate and effective personalized commercial advertisements to their needs. The overall statistical data of consumers make it possible for an e-commerce based incorporation to investigate and develop future potential products on demand.

This study discusses an emerging framework for the understanding (and creation as well) of business intelligence. Considering the dramatic growth of data, a theoretical design of the framework that contains five major portions – Data Mining Engine, KID Fusion Engine, Growing Consumer Behavior Model, Open Platform, and Computational Psychological Model – to collect contexts, such as activity logs, behavior, etc., from end users (consumers in this study) was proposed. This framework is featured by its core approach on data processing as it highlighted the importance of potential fusion on data, algorithm, and result.

Given a concrete usage scenario in e-commerce, this framework is expected to draw attentions from users, keep them, and furthermore predict their intentions through the analysis of collected contexts in the future. The framework was also implemented on an open platform that allows add-ons services from third-party developers so that the user contexts can be collected comprehensively and objectively. Although there are still limitations and some works remain, it is certain that construction and development of a system to provide smart business services is a worthwhile attempt to meet a profound society needs.

**Acknowledgments** The work is partially supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (No. 25330270).

## References

1. Ankur N, Abhinav S and Naga KKP, (2012) "High Performance Offline and Online Distributed Collaborative Filtering," Data Mining (ICDM), 2012 I.E. 12th International Conference, Brussels, pp. 549–558

2. Cao B, Shen D, Sun J, Yang Q and Chen Z (2007) "Feature selection in a kernel space," in International Conference on Machine Learning, Oregon, pp. 121–128
3. Chandramouli B, Goldstein J and Duan S (2012) "Temporal Analytics on Big Data for Web Advertising," Data Engineering (ICDE) 2012 I.E. 28th International Conference, Wasington, pp. 90–101
4. Cook RD, Weisberg S (1982) Criticism and influence analysis in regression. Sociol Methodol 13: 313–361
5. Ewaryst T, Adrian K (2009) Internet-technical development and applications. Springer, Tesco, p 255
6. Følstad A, Hornbæk K and Ulleberg P (2013) "Social design feedback: evaluations with users in online ad-hoc groups," Human-centric Computing and Information Sciences, vol. 3, no. 18
7. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20(15):2479–2481
8. Heckerman D (1997) Bayesian networks for data mining. Data Min Knowl Disc 1(1):79–119
9. Ho TK, Hull JJ, Srihari SN (1994) Decision in multiple classifier systems. IEEE Trans Pattern Anal Mach Intell 16:66–75
10. Holmes G, Donkin A and Witten IH (1994) "WEKA: A machine learning workbench," Proceedings of 2nd Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, pp. 357–361
11. Ibrahim N, Mohammad M and Alagar V (2013) "Publishing and discovering context-dependent services," Human-centric Computing and Information Sciences, vol. 3, no. 1
12. Jolliffe IT (1986) Principal component analysis. Springer, Berlin, p 487
13. Kittler JV, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20:226–239
14. Kuncheva LI, Bezdek JC, Duin RPW (2001) Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recogn 34:299–314
15. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51:181–207
16. Li W, Gong W, Liang Y and W. Chen (2005) "Feature selection based on KPCA, SVM and GSFS for face recognition," in International Conference on Advances in Pattern Recognition, Bath, pp. 344–350
17. Linden G, Smith B, York J (2003) Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Comput 7(1):76–80
18. Ma J, Wen J, Huang R, Huang B (2011) Cyber-individual meets brain informatics. IEEE Int Syst Spec Issue Brain Inform 26(5):30–37
19. McCrae RR, John OP (1992) An introduction to the five-factor model and its applications, Special issue: the five-factor model: issues and applications. J Pers 60:175–215
20. Mehrabian A (1996) Analysis of the big-five personality factors in terms of the PAD temperament model. Aust J Psychol : Melb 48(2):86–92
21. Milenova BL and Campos MM (2005) "Mining high-dimensional data for information fusion: a database-centric approach," Information Fusion, 2005 8th International Conference, Philadelphia, vol. 1, pp.638-645
22. Nakada H, Ogawa H, and Kudoh H, (2012) "Stream processing with BigData: SSS-MapReduce," Cloud Computing Technology and Science (CloudCom), 2012 I.E. 4th International Conference, Taipei, pp. 618–621
23. Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
24. Rakesh A and Ramakrishnan S (1994) "Fast Algorithms for Mining Association Rules in Large Databases," Proceedings of 20th International Conference on Very Large Data Bases, Morgan Kaufmann, pp. 487–499
25. Smirnov A, Pashkin M, Chilov N, Levashova T (2003) KSNet-Approach to Knowledge Fusion from Distributed Sources. Comput Inform 22(2):105–142
26. Thurau HT, Klee A (1998) The impact of customer satisfaction and relationship quality on customer retention: a critical reassessment and model development. Psychol Mark N J 14(8):737–764
27. Wang X, Ni Z and Cao H, (2007) "Research on Association Rules Mining Based-On Ontology in E-Commerce," Wireless Communications, Networking and Mobile Computing (WiCom 2007), Shanghai, pp. 3549–3552
28. Wang, and Tang T (2005) "A New Data Mining Method based on Fusion Clustering Algorithm," Neural Networks and Brain, 2005. ICNN&B '05. International Conference, Beijing, vol. 2, pp.706-711
29. Xiao YY and Aiming W (2010) "Genetic Algorithm Based Bayesian Network for Customers, Behavior Analysis," Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP-2010), 2010 I.E. 6th International Conference, Darmstadt, pp. 406–409
30. Yan J, Liu N, Wang G, Zhang W, Jiang Y, and Chen Z (2009) "How much can Behavioral Targeting Help Online Advertising?," the 18th international conference on World wide web (WWW '09), Madrid, pp. 261–270

31. Yang S, Huang R and Ma J (2012) "A Computational Personality-based and Event-driven Emotions Model in PAD Space," Sino-foreign-interchange Workshop on Intelligence Science & Intelligent Data Engineering, LNCS, Springer, in press
32. Yang J, Yang JY (2002) Generalized K-L transform based combined feature extrac-tion. Pattern Recogn 35: 295–297
33. Yang J, Yang JY, Zhang D, Lu J (2003) Feature fusion: parallel strategy vs. serial strategy. Pattern Recogn 36:1369–1381
34. Yaoxue Z, Yuezhi Z (2011) Separating computation and storage with storage virtualization. Comput Commun 34(13):1539–1548
35. Zhong N, Ma J, Huang R, Liu J, Yao Y, Zhang Y, Chen J (2013) Research challenges and persepctives on Wisdom Web of Things (W2T). J Supercomput 64(3):862–882

**Dr. Runhe Huang** is a professor in the Faculty of Computer and Information Sciences at Hosei University, Japan. She received a Ph.D. in Computer Science and Mathematics from University of the West of England in 1993. Before joining Hosei University, she worked at NUDT for 6 years and University of Aizu for 7 years in the field of Computer Science and Engineering. Her researches include multi-agents systems, computational intelligence, ubiquitous intelligence computing, cloud computing, Hyper-world modeling. She is member of IEEE and ACM. Contact her at rhuang@hosei.ac.jp.

**Mr. Atsushi Sato** received his Bachelor degree and Master degree at Department of Computer & Information Sciences, Hosei University, Japan. He is currently a PhD candidate under Dr. Runhe Huang's supervision at Graduate School of Computer & Information Sciences, Hosei University. His research interests are big data mining and KID fusion engine.

**Mr. Toshihiro Tamura** received his Bachelor degree and Master degree at Department of Computer & Information Sciences, Hosei University, Japan. His research interests are big data mining, social computing, and Cyber world computing.

**Dr. Jianhua Ma** is a professor of the Department of Digital Media in Faculty of Computer and Information Sciences at Hosei University, Japan. Previously, he had 15 years' working experience at NUDT, Xidian University and University of Aizu (Japan). His research interests include multimedia, networks, ubiquitous computing, social computing, and cyber intelligence. He has published over 200 papers, and edited over 20 books/proceedings and over 20 journal special issues. He is a co-founder of IEEE Int'l Conf. on Ubiquitous Intelligence and Computing (UIC), IEEE Conf. on Cyber, Physical and Social Computing (CPSCom), and IEEE Conf. on Internet of Things (iThings).

**Dr. Neil Y. Yen** is an Associate Professor at the University of Aizu, Japan. Dr. Yen received doctorates in Human Sciences (major in Human Informatics) at Waseda University, Japan, and in Engineering (major in Computer Science) at Tamkang University, Taiwan in March and June 2012 respectively. Dr. Yen has actively involved himself in the international activities and devoted himself to discover advanced and interesting research directions. Dr. Yen has been engaged in the interdisciplinary realms of research, and his research interests are now primarily in the scope of big data science, computational intelligence, and human-centered computing.