Sato *et al. Hum. Cent. Comput. Inf. Sci.* (2015) 5:23
DOI 10.1186/s13673-015-0036-z

Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

Design of fusion technique-based mining engine for smart business

Atsushi Sato[1], Runhe Huang[1] and Neil Y Yen[2*]

*Correspondence:
neil219@gmail.com
[2] School of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu, Fukushima, Japan
Full list of author information is available at the end of the article

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 data mining engine based on the concept of the KID model is designed and developed to be responsible for the universal collection of data and mining valuable information which is primarily from real world, cyber world, and social world. Second, we go further into the fusion process of the collected data and meaningful information extracted and interpreted by algorithms or fused algorithms in the data mining engine (e.g., the consumer purchase data shared by real-world company) and turn them into valuable knowledge about the situation of customers and business situations based on the concept of knowledge, information, and data. A three-layer analysis and mining procedure is designed to enhance the mining engine through conventional RFM (Recency, 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.

Keywords: Data fusion, Multi-layered fusion, Mining engine, Planning-based prediction, Smart business, Cyber-I, Hyper world

Background

Smart business, by definition, indicates the ability to achieve goals which are set according to the development tendency of business [1]. 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 [2, 3].

Transforming data into meaningful and useful information [4] that support the implementation of smart business is a long journey. Although rapid development in information communication technology [5] 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—Hyper World [6, 7]—in which data is supposed to be from diverse channels [8] and in unstructured formats [9]. 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 [10] 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 [11], which enables users to find matching search trends. Oracle Data Mining Engine (DME) [12] is the infrastructure that offers a set of in-database data mining functionality to its JDM (Java Data Mining) clients via a DME connection object. Amazon, the retail giant has been focusing on product recommendation engine, but recently, released Amazon Kinesis [13], 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 [10] 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 [10] 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.

This study, to the above-mentioned reason, attempts to make obtained knowledge from data apply to the development of future smart business. So three major tasks are considered. 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 (Recency, 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.

Rest organization of this paper includes: "Related work" details the previous studies that relate to this study; "Universal design of fusion technique-based smart business framework" addresses the design of the fusion technique-based smart business framework; "A case study on a retail business" gives a case study demonstrating the feasibility and preliminary results with the support of proposed framework; and "Conclusions" then concludes this paper and indicates potential extension of this work.

Related work

Summary of the past works that relate to this paper are studied. Three issues are particularly focused. First we discuss the existing fusion techniques in the data level. We then indicate methods that support the mining and analysis process at this stage of fusion. The importance to employ the fusion technique for the development of smart business paradigm is then given. A short summary to the above-mentioned issues will be given as the end of this section.

Fusion techniques in data level

Fusion techniques in data level have been studied in both disciplines of information system and statistics. Fusion techniques in information system is focused on the algorithms and process models deriving meaningful information effectively and efficiently for decision support from acquirable multiple data or multi-source information. The techniques are applied for tracking object, event detection, context recognition, medical diagnosis and information management in military, academic, commercial, and industrial fields [14, 15]. In contrast, fusion techniques in statistics concentrate on understanding of data analysis in observational study. The frameworks for integrating multi-source data as single-source data including missing values [16, 17] are based on missing data problems and multiple imputation proposed by Rubin [18], which is the framework for the analysis of incomplete data in observational study. Besides, algorithms and evaluation criteria for feature selection [19] are studied in order to extract appropriate subset from massive input variables.

Data fusion-based mining

In many cases of multi datasets analysis, a new dataset is generated by gluing their common variable such as customer-ID, time, geographical coordinates and response variable, and useless variables are cut from the dataset. Therefore, selecting appropriate variables from a dataset to explain a response variable is a fundamental data fusion

technique in mining process. Stepwise regression [20] extends simple regression analysis to treat multi explanatory variables. It adds/removes a variable one by one into/from the regression model on the basis of the criterion such as prediction criteria, likelihood, information criteria and Bayesian posterior probabilities. Principal Component Analysis [21] is another approach to reduce the number of variables, which compresses high dimensional data to a lower one by transforming to uncorrelated variables. Classification techniques [22] such as Naive Bayes Classifiers and decision trees learn correlations between a response variable (called class label) and other variables based on empirical probability. As a result, a handful of effective predictors can be clarified.

Advances in data fusion for smart business

Collecting, managing, analyzing and visualizing data are conducted for decision making support on business. Data collected from distributed information systems are treated in data management and warehousing techniques, and appropriate datasets/information are generated and provided to analysts or inquirers. The data is divided into three types, structured data such as transaction data and customer information, web-based unstructured data, IoT-based continuous data [23]. Knowledge Discovery in Databases (KDD) and Data Mining techniques for business intelligence focus on one of the data types. Online Analytical Processing (OLAP) [23] begins analytical process dynamically once gathering structured data into a database, this allows a user to quickly obtain the answer of his query. Text mining and web mining techniques [24] enable to gather, organize and visualize information of company, industry, product and customer from unstructured data on the web. IoT applications [5] monitor the status of users, things and environment through continuous data processing so as to support personal life or business enterprise. Data analytics platform for enterprise are provided by various companies. Google in 2011 released Data Mining Engine called Correlate [11], which enables users to find matching search trends. Google Analytics [25] finds patterns of user activities on a website and a report is shown for improving the contents of the website. Oracle Data Mining Engine (DME) [12] is the infrastructure that offers a set of in-database data mining functionality to its JDM (Java Data Mining) clients via a DME connection object. IBM Cognos Business Intelligence [26] provides optimized access to structured data and OLAP and an integrated business view of any number of heterogeneous data sources. It can also handle continuous data in monitoring of business activity. Amazon, the retail giant has been focusing on product recommendation engine, but recently, released Amazon Kinesis [13], 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. The process of dataset generation in conventional data analytics is mainly accomplished by integrating multi datasets and selecting related variables. However, objectives and selected algorithms of an analytics are correlated to datasets which are generated for the algorithms. Considering the correlation in the dataset generation can induce partial analytic algorithms and fuse the processed information and the law dataset. Companies obtain meaningful datasets/information for their business through the data fusion [27].

Summary

Data fusion techniques in information systems and statistics aim to gain accurate and credible information from multi datasets. Some data analytics platform enable to access to processed information through OLAP before analysis process. The fusion technique in our work describes datasets with feature labels and defines algorithms by their inputs and outputs. The correlations between datasets and algorithms enhance to search and generate appropriate datasets with corresponding algorithms.

Universal design of fusion technique-based smart business framework

Key to successful implementation of a smart business paradigm relies on many aspects [28, 29]. 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]

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, the ultimate goal of our KID model [30], together with a set of fusion techniques (information and knowledge fusion in the process of reasoning and problem solving with constraints to a specific business and goals of a company) 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

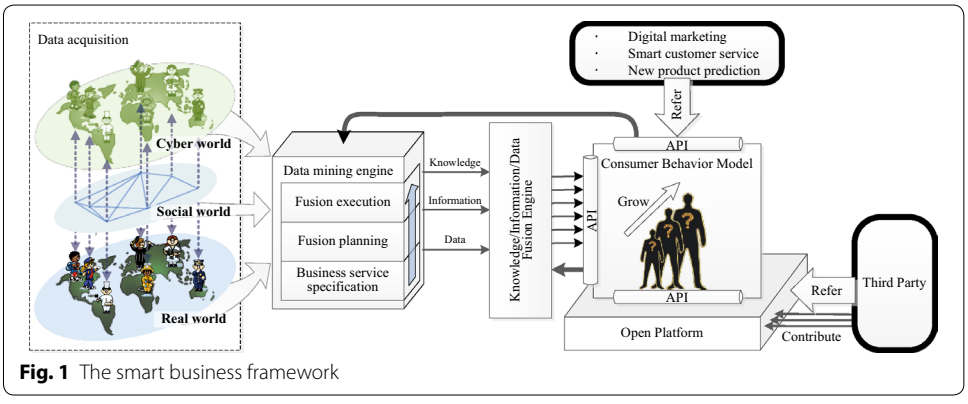

Fig. 1 The smart business framework

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[b] 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 algorithm fusion process, and select appropriate[c] mining algorithms as a sequence of fusion plan 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 [31], even knowledge, information, and raw data itself, may be fused in the case of necessary. This approach is based the concepts of knowledge, information, and data, and their interrelations in our previous work [30];
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. The data-information-knowledge-data cycle enables the customer model more and more approximate to its real person; and
5. *Open platform* is a 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 [32], 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 [33, 34], 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.

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 [35] 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 mechanisms are 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 to discover knowledge for a company. A three-level fusion technique for different business objectives and a set of fusion-based learning algorithms for prediction are developed.

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 which is based our human-like cognitive metaphor model [30]. In fact, the data mining engine equates to a human-like intelligent entity which can think rationally with some cognitive skills and perform problem solving. With understanding of objectives of a company, it can adopt the backward

chaining approach and match a sequence of algorithms in a fusion plan with selected data resources to achieve mining necessary features and valuable information.

As shown in Fig. 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.

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.

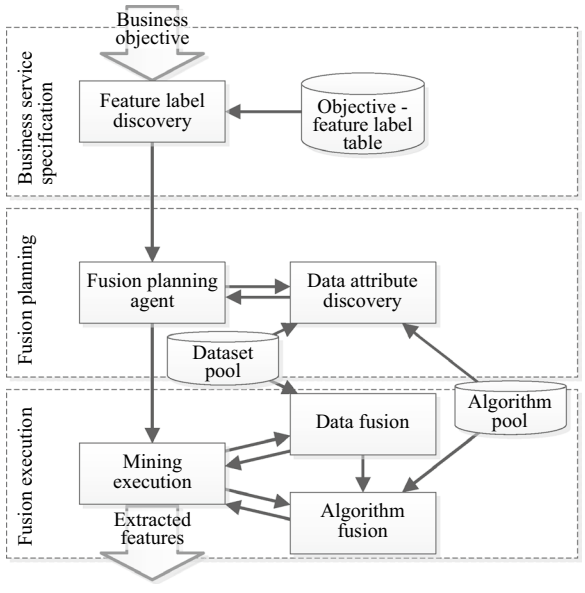

Fig. 2 Process flow of fusion technique-based data mining engine design

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 [36, 37] 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 [38]. 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.

In this paper, we concentrate on those high-royalty consumers of a real-world retail company, and attempts to support the discovery (and prediction as well) of the periodical purchase cycle of their consumers. Three steps are included in this method.

Consumer cluster generation

The consumer can be briefly categorized by their purchase records through the original RFM [Regency (R), Frequency (F), and Monetary (M) Value] model [39, 40]. A ten-scale range is also applied to indicate the high-royalty and low-royalty consumers in our scenario.

Given that notations, R_{avg}, F_{avg}, and M_{avg}, refer to the average score of regency value, frequency value, and monetary value in the scenario, we may first use the general expression to model the consumers like:

$$\bar{X} = \frac{(x_{max} - x_{min})}{L} \tag{1}$$

where $\bar{X} = \{x | x \in R, F, M\}$ denotes the average of sub-element sets in the domain R, F, or M, and x_{max} and x_{min} then identify the maximum value and minimum value of the elements in selected sets. The notation L, $L \in \mathbb{N}$, $L = \{l | l = \{1, 2, \ldots, n\}, l \neq 0\}$, indicates a regular number defining the scale to be applied on the different level of consumers. In our case, a detail scale at 10 is applied (i.e., $L = 10$).

Following this definition, the RFM-value, V_{rfm}, can be obtained through:

$$V_{rfm} = \frac{\sum \bar{X}_u}{U} \tag{2}$$

where $\sum \bar{X}_u$ denotes the sum of average values in R, F, and M of a given consumer u, and U is the sum of consumers that a company has in the records. So that we can then separate the high- and low-royalty consumers simply through:

$$if\ \bar{X}_u < V_{rfm}\ then\ u \in RFM_{low}\ else\ u \in RFM_{low}$$

Then we proceed further to discuss the implicit preference of consumers in both high-royalty and low-royalty, and we attempt to categorize consumers who possess similar purchase behaviors. The SOFM (Self-organizing feature map) [41] is applied as the basis for learning. Multiple inputs at each learning phrase are expected. We assume that the error rate (e.g., noises from input) is low and acceptable since all the inputs are from same scenario (retail store) of a company. So that we can create a map from all these n-dimension inputs and keep the topology for further usage. The following process is applied.

(1) *Initialization of all possible inputs* The inputs here refer to data, processed data (i.e., information), and knowledge collected or derived from every stage in our framework.

(2) *Calculation of nearest neural to selected pattern among dataset* A pattern is randomly selected based on the Euclidian distance on the space. The Eq. (3) is utilized

$$D(a_i, b_j) = \sqrt{\sum (a_{ik} - b_{jk})^2} \tag{3}$$

where a_i denotes the current pattern i, and b_j represents a neural on the space map. The distance formulation is utilized to obtain a nearest neural on the map, and meanwhile, update the function among related neural sets.

(3) *Update correlated neural sets* After we obtain the nearest neural to the given pattern, the distance among the pattern and other neural sets in the same group of target neural [the winner in step (2)] may change. The update rule is defined as

$$w_p(t+1) = w_p(t) + K_p(t) \cdot \left[a(t) - w_p(t)\right] \tag{4}$$

Equation (4) exists if $i \in N_C(t)$. In this learning formula, notation $w_p(t)$ indicates the weight of a neural p at time t. The notation $a(t)$ is the selected pattern and $N_C(t)$ is the size of coverage range at time t. The notation $K_p(t)$ is the kernel function for neural p at time t. Kernel function is defined:

$$K_p(t) = \alpha(t) \cdot \exp\left(-\frac{||n_j - n_i||}{2\sigma^2(t)}\right) \tag{5}$$

The kernel function $K_p(t)$ includes the learning function, $\alpha(t)$, and an adjustment function, σ. It changes in accordance with the movement of time t. In general case, the learning results meet the normal distribution [42]. During the period of learning phrase, some issues concerning the efficient and correctness of learning rate may be taken place. However, this issue can be ignored since the input is focused and from single company.

Periodical sale analysis

In addition to the implicit preferences extracted from the collected data, the next mission is to develop an universal model that prompts the prediction. In order to achieve well, and efficient, prediction, understanding the trend (or movement of peak [43] in trend) is also important after the preference of user is correctly obtained.

For this purpose, a model for periodical sale analysis in the Fusion-based Learning method is designed. This model is primarily applied to obtained the most likely action(s)

which our potential consumers may take in the next given period (e.g., four quarters in a year or different months in a season, etc.). Three phrases are included:

1. *Calculating the difference on item purchase period* The first phrase applies the concept of recency-value to calculate the transactions of every consumers. It concentrates on relationship of DF_i^t (difference on date frame) between the NT_i (current time for purchasing) and LT_i (last time for purchasing) where notations t stands for the n-th purchase of the i product by a specific consumer. This process can be formulated by:

$$\begin{aligned} DF_i^t &= NT_i - LT_i \\ DF_i &= \left\{DF_i^1, DF_i^2, \ldots, DF_i^t\right\}, t = 1, \ldots, T \end{aligned} \tag{6}$$

2. *Calculating the difference on purchase period of consumer* This phrase first calculates the average purchase time period, ST_i^t (sales time for item), for each product. It indicates the number of purchased product, k_i^t, of product i on t-th by consumer. Then the obtained sales time items are put into a set in sequence $ST_i = \left\{ST_i^1, ST_i^2, \ldots, ST_i^t\right\}$ [see Eq. (7)]. After that, the minimum sales period, $STmin_i$, and average sales period, $\overline{ST}_i$, can be obtained by Eqs. (8) and (9).

$$ST_i^t = \frac{DF_i^t}{k^t}, ST_i = \left\{ST_i^1, ST_i^2, \ldots, ST_i^t\right\} \tag{7}$$

$$STmin_i = \min\left\{ST_i^t\right\} \tag{8}$$

$$\overline{ST}_i = \frac{\sum ST_i^t}{T} \tag{9}$$

3. *Analysis of item recommender period* The last phrase is then applied to calculate the item recommender period for a specific product, which can be revealed by $PRI_i[d_{min}, d_{max}]$. It especially concentrates on the potential bias, B_i, and the standard derivation, σ_i, between $STmin_i$ and $\overline{ST}_i$.

$$B_i = \overline{ST}_i - ST\min_i \tag{10}$$

$$\sigma_i = \sqrt{\frac{1}{T}\sum\left(ST_i^t - \overline{ST}_i\right)^2} \tag{11}$$

A case study on a retail business

In order to well explain the proposed fusion technique based data mining engine, customer purchase record data obtained from a retail business corporation is applied as a case study in this Section.

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

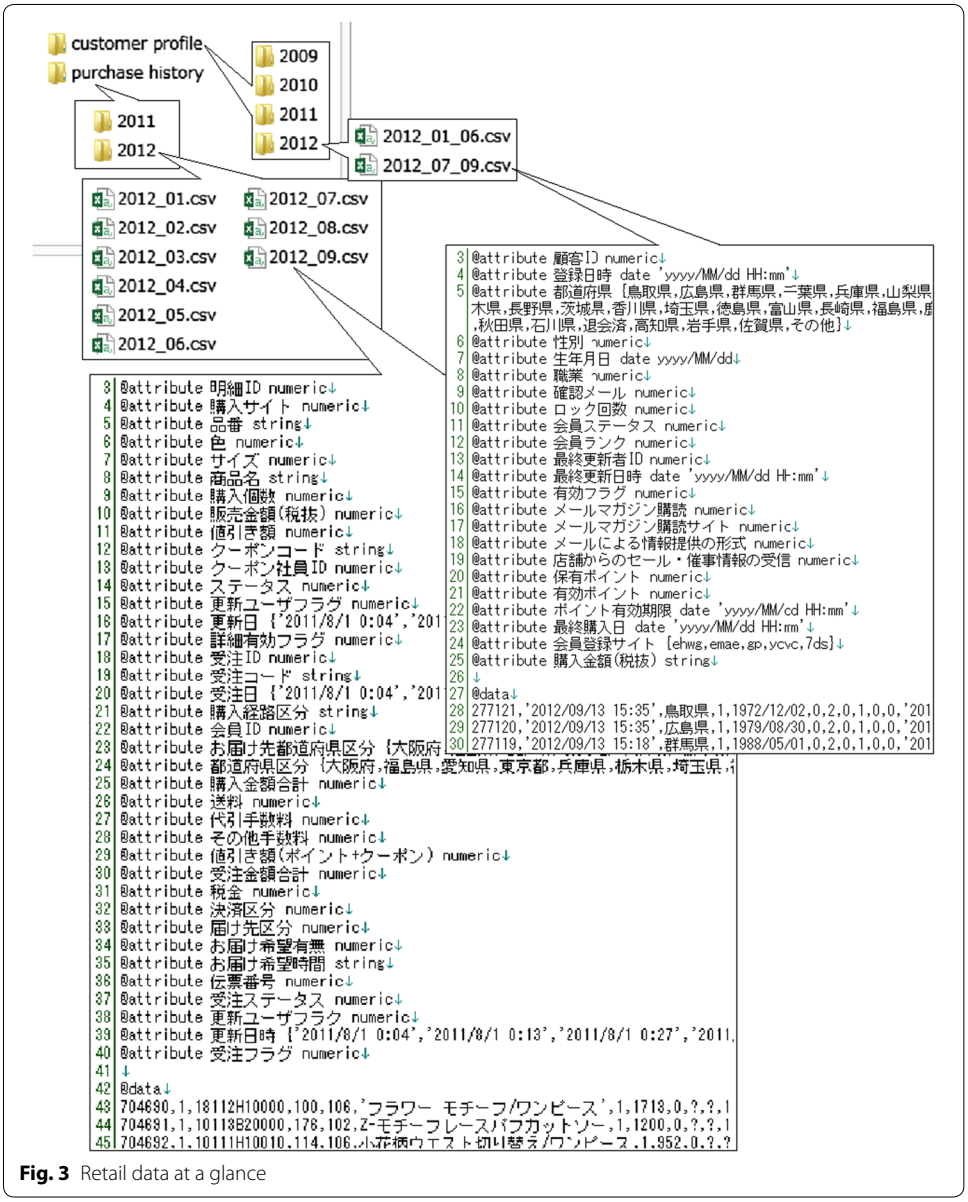

Fig. 3 Retail data at a glance

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.

Apply backward chaining fusion technique

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

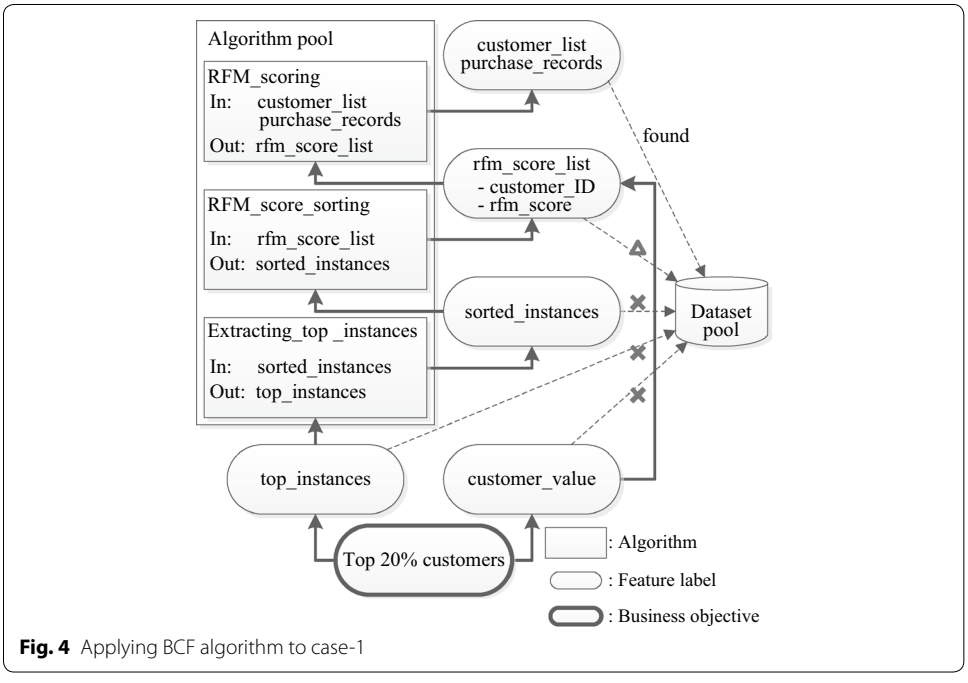

Fig. 4 Applying BCF algorithm to case-1

Analysis and remarks

The company shared about 45,000 transaction records for the experiments. The purpose of this experiment is mainly to evaluate whether the proposed mining algorithms work especially on the selection of the correct record for the mining and/or learning phrases.

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 four types of fusion techniques may be applied.

In addition, an experiment was conducted to obtain the performance of our proposed fusion-based learning method for the collected dataset. The TREC and the metrics [44] for estimating the performance are shown as follows, where P represents the precision value and R indicates the recall value. The nIAP, non-Interpolated Precision, was used to calculate the precision of a trained result within a supervised dataset.

$$P\prime = \#\text{ of relevant items retrieved}/\#\text{ of items retrieved}$$
$$R\prime = \#\text{ of relevant items retrieved}/\#\text{ of items relevant}$$

The corresponding results to queries above can be shown in Table 1.

According to the results (three rounds of experiments), it is worth mentioning that the precision and recall value reach, in average, 82 and 91%, respectively, which reveal that our fusion-based training method can prompt the analysis of the implicit patterns from the collected data in a relatively correct way. And this result also shows that the training dataset, randomly-picked from the raw data from our collaborator, have certain creditability for being the basis of make further predicions.

Conclusions

Development of smart business has become an essential factor for the traditional business to stay competitive to the ever-growing world. In addition to the global trend, the key to the success of next-generation business is the data. Data itself means nothing but a bulk of records, which is also very big and with many noises, and often cost a lot for the owners (i.e., enterprises, government, etc.) to maintain. But however, the data can also lead to the value of business if the meaning (or phenomenon) of data can be well extracted and learned.

Lots of approaches have been discussed and applied in the past studies, but most of them consider the algorithms (or models) that are with main focuses on specific cases and the performance enhancement as well. This research looks at this issue from a

Table 1 Results to precision-recall evaluation

	R_1	R_2	R_3
Total transactions	44,890		
Precision	82.85%	84.02%	80.81%
Recall	91.06%	90.41%	92.79%

different point of view. It especially concentrates on the fusion part of the whole process, e.g., mining, learning, etc. The design of fusion technique-based mining engine is the essential 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 core of DME is explained. Also a three-layer analysis and mining procedure is designed to enhance the mining engine through conventional RFM (Recency, Frequency, and Monetary Value) model with the supports of a set of fusion techniques. In the end, we implement a planning-based prediction model for a real-world company for expansion of the business interests. 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.

Many works remain in the future. One of them is to formulate the data mining engine and the KID fusion engine from a cognitive viewpoint so as to achieve a real generic data-information-knowledge-data framework. Our previous work on the KID model will lead significant outcomes in the follow-up development of expected generic framework.

Endnotes

[a]The term "right-to-the-needs" indicate that the service is provided to the right person, at right time, and with right contexts.

[b]Users here indicate either the directors or consumers of a company. Special requests (e.g., year-end marketing plan, promotion, etc.) may be submitted to the engine for instant outputs by directors, while survey of feedback from consumers may remind the changes (or alternative solutions) to company's roadmap.

[c]The term "appropriate" indicates those mining algorithms with efficiency and high-performance ability at the time that specific types of data is given for the fusion.

Authors' contributions

AS proposes the mathematical model of the objectives driven and backward chaining inference (a fusion planning of required algorithms and necessary datasets) based mining strategy as the core mechanism in the data mining engine, and applies the approach on a retail business scenario. RH proposes the framework of smart business, and is responsible for the overall supervision of this research work. NY revises the structure of the paper, and is responsible for the design of fusion-based learning method. All authors read and approved the final manuscript.

Author details

[1] Graduate School of Computer and Information Sciences, Hosei University, Koganei-shi, Tokyo, Japan. [2] School of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu, Fukushima, Japan.

Acknowledgements

The work is partially supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (No. 25330270 and No. 26330350).

Compliance with ethical guidelines

Competing interests

The authors declare that they have no competing interests.

Received: 23 January 2015 Accepted: 13 May 2015
Published online: 31 July 2015

References

1. Watson HJ, Wixom BH (2007) The current state of business intelligence. Computer 40(9):96–99
2. Cody WF, Kreulen JT, Krishna V, Spangler WS (2002) The Integration of business intelligence and knowledge management. IBM Syst J 41(4):697–713

3. Sun N, Morris JG, Xu J, Zhu X, Xie M (2014) iCARE: A framework for big data-based banking customer analytics. IBM J Res Dev 58(5/6):4:1–4:9
4. Parsons S (1996) Current approaches to handling imperfect information in data and knowledge bases. IEEE Trans Knowl Data Eng 8(3):353–372
5. Xu LD, He W, Li S (2014) Internet of things in industries: a survey. IEEE Trans Industr Inf 10(4):2233–2243
6. Kunii TL, Ma J, Huang R (1996) Hyperworld Modeling. In: Proceedings of the International Conference on Visual Information Systems, Australia, February, pp 1–8
7. Chen J, Ma J, Zhong N, Yao Y, Liu J, Huang R et al (2014) Waas: wisdom as a service. IEEE Intell Syst 29(6):40–47
8. Chen H, Chiang R, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188
9. Blumberg R, Shaku A (2003) The problem with unstructured data. DM Rev 13:42–49
10. Ma J, Wen J, Huang R, Huang B (2011) Cyber-individual meets brain informatics. IEEE Intell Syst Spec Issue Brain Inform 26(5):30–37
11. Mohebbi M, Vanderkam D, Kodysh J, Schonberger R, Choi, H, Kumar S (2011) Google Correlate Whitepaper. https://www.google.com/trends/correlate/whitepaper.pdf. June 2011, Google
12. Taylor KL (2011) Data mining: application Developer's Guide. http://docs.oracle.com/cd/E11882_01/datamine.112/e12218.pdf. July 2011, Oracle
13. Varia J, Mathew S (2014) Overview of Amazon Web Services. http://media.amazonwebservices.com/AWS_Overview.pdf. Jan 2014, Amazon
14. Liggins M, Hall DL, Llinas J (eds) (2008) Handbook for multisensor data fusion: theory and practice, 2nd edn. CRC Press, New York
15. Sheth AP, Larson JA (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput Surv 22(3):183–236
16. Kamakura WA, Wedel M (1997) Statistical data fusion for cross-tabulation. J Mark Res 34:485–498
17. Susanne R (2004) Data fusion: identification problems, validity, and multiple imputation. Aust J Stat 33(1):153–171
18. Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
19. Guyon I (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
20. Alan M (2002) Subset selection in regression. CRC Press, London
21. Jolliffe I (2002) Principal Component Analysis. Encyclopedia of Statistics in Behavioral Science, 2nd edn, vol 30. John Wiley & Sons, Ltd, p 487
22. Kotsiantis SB, Zaharakis ID, Pintelas PE (2007) Machine learning: a review of classification and combining techniques. Artif Intell Rev 26(3):159–190
23. Codd EF, Codd SB, Salley CT (1993) Providing OLAP (on-line Analytical Processing) to user-analysts: an IT mandate. Codd Date 32
24. Miner G, Elder J, Hill T, Nisbet R, Delen D, Fast A (2012) Practical text mining and statistical analysis for non-structured text data applications, 1st edn. Academic Press, Waltham
25. Google Analytics Official Website. http://www.google.com/analytics/ce/mws/
26. Negash S (2004) Business intelligence. Commun Assoc Inf Sys 13(1):54
27. Sato A, Tamura T, Huang R, Ma J, Yen NY (2013) Smart Business Services via Consumer Purchasing Behaviour Modeling," IEEE International Conference on Cyber, Physical and Social Computing (CPSCom'13), Beijing, China, August 20–23, 2013
28. Heck EV, Vervest P (2007) Smart business networks: how the network wins. Commun ACM 50(6):28–37
29. Lee J (2003) Smart products and service systems for e-business transformation. Int J Technol Manag 26(1):45–52
30. Sato A, Huang R (2015) A Generic Formulated KID Model for Pragmatic Processing of Data, Information and Knowledge. The 12th IEEE International Conference on Advanced and Trusted Computing (ATC2015), Beijing, China, August 10–14, 2015, IEEE publisher
31. Zins C (2007) Conceptual approaches for defining data, information, and knowledge. J Am Soc Inform Sci Technol 58(4):479–493
32. Anderson JC, Narus JA (1998) Business marketing: understand what customers value. Harvard Bus Rev 76:53–67
33. Moutinho L (1987) Consumer behaviour in tourism. Eur J Mark 21(10):5–44
34. Wang YS, Lin HH, Luarn P (2006) Predicting consumer intention to use mobile service. Inform Syst J 16(2):157–179
35. Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inform Commun Soc 15(5):662–679
36. Xiao B, Benbasat I (2007) E-commerce product recommendation agents: use, characteristics, and impact. Mis Q 31(1):137–209
37. Duhan DF, Johnson SD, Wilcox JB, Harrell GD (1997) Influences on consumer use of word-of-mouth recommendation sources. J Acad Mark Sci 25(4):283–295
38. Delone WH, Mclean ER (2004) Measuring e-commerce success: applying the DeLone & McLean information systems success model. Int J Electron Commer 9(1):31–47
39. Hosseini SMS, Maleki A, Gholamian MR (2010) Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Syst Appl 37(7):5259–5264
40. Kohavi R, Mason L, Parekh R, Zheng Z (2004) Lessons and challenges from mining retail e-commerce data. Mach Learn 57(1–2):83–113
41. Bauer HU, Pawelzik KR (1992) Quantifying the neighborhood preservation of self-organizing feature maps. IEEE Trans Neural Netw 3(4):570–579
42. Hoeffding W (1948) A class of statistics with asymptotically normal distribution. Ann Math Stat 19:293–325
43. Liu DR, Shih YY (2005) Integrating AHP and data mining for product recommendation based on customer lifetime value. Inform Manag 42(3):387–400
44. Jarvelin K, Kekalainen J (2000) IR Evaluation Methods for Retrieving Highly Relevant Documents. In: Proc. 23rd ACM Int'l Conf. Information Retrieval