Knowledge Digest Engine for Personal Bigdata Analysis

Youngrae Kim, Jinyoung Moon, Hyung-Jik Lee, and Chang-Seok Bae

Electronics and Telecommunication Research Institute, Daejeon, Korea
218 Gajeongno, Yuseong-Gu, Daejeon 305-700 Korea
{youngrae,jymoon,leehj,csbae}@etri.re.kr

Abstract. The bigdata analysis has an issue of high knowledge creation. In this paper first, we define personal big data, and using personal bigdata created by user activity we try to create high knowledge about the user. We have created personal bigdata analytic engine and knowledge digest engine for high knowledge creation and personalized service. The engine is used to collect, process and analyize personal big data. And In the process we refine, associate, and fuse data for analysis. In this paper, we show the process of analyzing personal big data, and detailed structure of analyzing engine for persoanl big data. High knowledge about the user will lead to better personalized services, and better adaptive services.

Keywords: User Activity Analysis, Personal Bigdata, personalized services, adaptive services.

1 Introduction

Personal bigdata are created every day. From GPS, acceleration sensors, to electroencephalography sensors are easily reachable to users who are interested in their personal bigdata [1, 2]. Currently these data are used as it is or used with basic pattern recognition or basic statistical analysis. Such data as GPS are used in navigation system as GPS sensor value [3], and in some application, GPS data are used to find a favorite places that user likes to go using statistical approach [4]. These data are collected using multiple devices and sources shown on Table 1.

These collected data are unstructured and they cannot be made to knowledge without data processing. The purpose of personal bigdata analysis is to inference the specific user characteristics that are not normally shown with basic statistical approach. In this paper, we show the steps of data processing for personal bigdata analysis and the steps of analysis engine we have structured to fit the personal big data analysis.

James J. (Jong Hyuk) Park et al. (eds.), *Human Centric Technology and Service in Smart Space,*
LNEE 182, pp. 261–267, DOI: 10.1007/978-94-007-5085-2_34,

Table 1. Devices/Source for collection of personal bigdata

Devices / Source	Collected Data
Zephyr HxM BT	Heart rate (HR), Speed, Distance
Withings Weight	Weight, BMI, Fat
Withings BP	Blood pressure, HR
Nonin SpO2	Oxygen level, HR
Jawbone UP	Movement, Sleep Pattern
Fitbit ultra	Walking, Distance, Calorie, Time
Motoactv	Walking, Distance, Calorie, Time
Nike+	Walking, Distance, Calorie, Time
Bodymedia FIT	HR, Temperature
Smartphone	GPS, Time, 3-axis Accelerometer, etc
Social network	Written contents, associated friends, time, location, etc.
Purchase	Purchase, price, amount, place
Web logs	Keywords, Time, Platform
Schedule	Associated Friends, time, location
NeuroSky	EEG
Emotiv EPOC	EEG

2 Background

2.1 Personal Bigdata

Bigdata is known for three Vs; Volume, Variety and Velocity [5]. Currently these bigdata analyses are mostly done in the fields of social network analysis [6, 7] as a analysis of many people. We define personal bigdata as a data created by the user's activity that has the attribute of bigdata. The personal big data records have volume, in sense that these data are recorded over a lifetime. Also even though we will not be written in this paper, the video and audio also can be used as a personal data, which is known to be huge in volume. The devices give variety of personal data, which are structured, unstructured, semi structured data. These data are used to service personalized services in real time, and these data are created in streams, which gives velocity attribute of bigdata.

2.2 Personalized Services

Current commercial personalized services use stereotypes [8, 9]. Such internet services group people into gender, ethnic, age, and interest groups to provide different services for different people. In research papers, for personalized services, used real "personal" data for analysis to provide real personalized services [1]. But, these services are domain-specific, and use limited information of user.

In this paper, we show how to analyze personal bigdata user for personalized services.

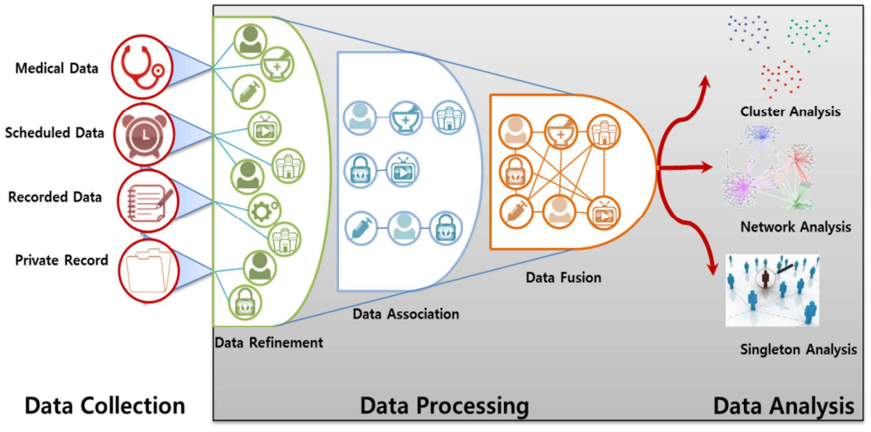

Fig. 1. The process of personal big data analysis

3 Data Processing of Personal Bigdata

Figure 1 shows the data processing of personal big data. The process consists of three part data collection, data processing, and data analysis.

3.1 Data Collection

The data are collected using the sensors introduced in the Table 1, or data created by the user. Such data as email, schedules, and personal notes also can be used as a personal bigdata as well as a private record and medical data can be used. Due to the lifetime difference of data, some data such as profile is collected only once, while some data such as social data are collected in batch, or once a week, and some data such as EEG is collected every millisecond.

3.2 Data Processing

3.2.1 Data Refinement

Personal bigdata are sometimes unstructured, and not complete. And since these data came from different source of origin the value it represents may need normalization. Therefore in the data refinement process, natural language processing or database mapping is used to create values that can be used in analysis from the personal bigdata. Missing value imputation is used to fill in the missing values using regression or mean value. Sometimes missing value imputation is not used depending on the data type, such data as heartbeat is has a lifetime of 2~3 seconds while heartbeat pattern of daily life can be found using logs of heart beats, the missing value does not need to be imputed every time for a personalized services. Noise filtering algorithms are used to remove the background noise. Lastly, for data refinement the selection of useful and

meaningful data is needed. These data are used as it is for a personalized service, and are used in the association.

3.2.2 Data Association

Using the refined data, we associate different kinds of data. We check for correlation, association using statistical and data mining approach. These links represent knowledge of a user. For example, using GPS data and medical data, we may find the favorite hospital the user goes for a specific disease. Using purchase data and time data, we may find the frequency of buying specific goods. Linking separate data allow us to find more specific user activity and allow us to represent more accurate data. The linked data itself also can be used in personalized services.

3.2.3 Data Fusion

For data fusion, we use the associated data and link them together to form a new network. We used network construction algorithm to form different kind of networks for analysis. Different scenarios are created according to the network.

3.3 Data Analysis

We use clustering analysis, network analysis and singleton analysis for data analysis. For clustering analysis, we separate each entity to the specific time frame then select the attributes for clustering. The timeframe that has similar values are clustered. These clusters are used to provide personalized services.

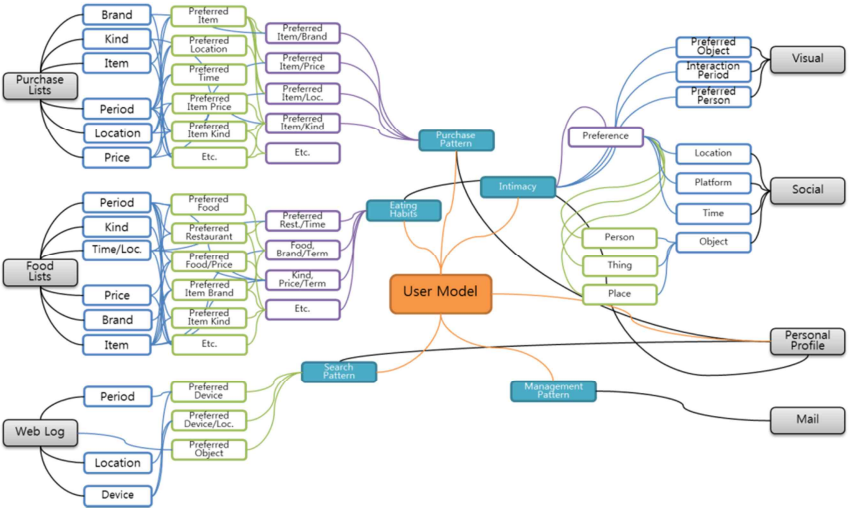

Fig. 2. Creation of user model using analysis of personal big data

The networks that are constructed in the data fusion level, the network are used to recognize specific status of the user. Also data itself can be a network, such as social network data; therefore the network analysis is applied to the specific data.

On the singleton analysis, we try to see the changes in the single value depending on other value changes to gather information of the single value. Such analysis as change in weight depending on different variables can be analyzed using singleton analysis.

These analyses are used to create user model, shown on Figure 2. Seven basic logs are used for analysis, and user model, which represents the characteristics of user, is then used for personalized service

4 Personal Bigdata Analytic and Knowledge Digest Engine

4.1 Personal Bigdata Analytic Engine

Personal bigdata analytic engine consist of four modules shown on Figure 3; preprocessing module, feature extraction module, analysis module, and low level knowledge information handle module.

Preprocessing module is module for the refinement process explained on 3.2.1. The data are preprocessed to fit the machine learning. Natural languages are transformed into weights or fuzzy logics, and different words are mapped to standard words. The missing values are imputed, and values are normalized.

Feature extraction module uses feature selection algorithms to select most useful and meaningful features from the data. These selected features are then analyzed using mining library, and statistic library. The analyzed knowledge is called low level knowledge information, and it is stored into the database.

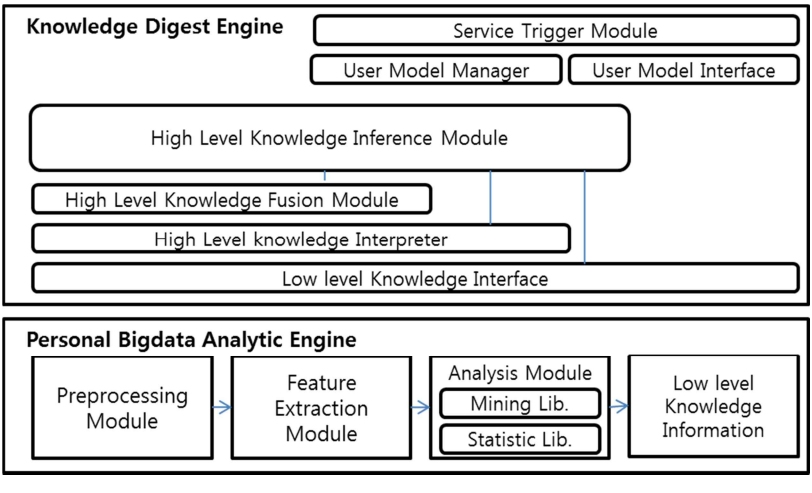

Fig. 3. Structure of Personal Bigdata Analytic Engine and Knowledge Digest Engine

4.2 Knowledge Digest Engine

Knowledge digest engine is also shown on Figure 3. As explained on 3.2.2 and 3.2.3 the data gathered from low level knowledge information and processed by the interpreter, or linking the low level knowledge information data, and then links are fused with other links to construct a network. This network is then used to create high level knowledge, which is the characteristics of the user, which will update the user model. After the user model update, the service is triggered through service trigger module.

5 Future Work

For future work, we plan to automatize the analysis and provide service accordingly. The recognition of user activity and data gathering itself needs improvements. The image analysis, natural language analysis, and auditory analysis must be improved for better real-time personalized service. The data analysis from variety of data from single source can be biased, so we plan to gather from different sources for the analysis. More data collection is needed for further verification of our engine.

6 Conclusion

In this paper, we have shown how to analyze personal bigdata and structure of personal bigdata analysis engine and knowledge digest engine. The challenge of the personal bigdata is the finding the best analysis algorithm for variety of data, and interpreting the meaning of the results. Also, the devices for gathering personal bigdata should be developed for ease of gathering information.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MEST) (NRF-M1AXA003-20110028371).

References

1. Parkka, J., Ermes, M., Korpipaa, P., Mantyjarvi, J., Peltola, J., Korhonen, I.: Activity classification using realistic data from wearable sensors. IEEE Transactions on Information Technology in Biomedicine 10, 119–128 (2006)
2. Bao, L., Intille, S.S.: Activity Recognition from User-Annotated Acceleration Data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
3. Popa, M.: Pedestrian Navigation System for Indoor and Outdoor Environments. In: Hippe, Z.S., Kulikowski, J.L., Mroczek, T. (eds.) Human – Computer Systems Interaction: Backgrounds and Applications 2, Part I, AISC, vol. 98, pp. 487–502. Springer, Heidelberg (2012)
4. Kobsa, A.: Generic User Modeling Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 136–154. Springer, Heidelberg (2007)

5. Russom, P.: TDWI Best practices report: Big Data analytics q4 2011. The data Warehousing Institute (2011)
6. Ahn, J.-w., Taieb-Maimon, M., Sopan, A., Plaisant, C., Shneiderman, B.: Temporal Visualization of Social Network Dynamics: Prototypes for Nation of Neighbors. In: Salerno, J., Yang, S.J., Nau, D., Chai, S.-K. (eds.) SBP 2011. LNCS, vol. 6589, pp. 309–316. Springer, Heidelberg (2011)
7. Marres, N., Weltevrede, E.: Scraping the Social? Issues in real-time social research. Journal of Cultural Economy (subm.), 1–52, (Article): Goldsmiths Research Online
8. Procci, K., Bohnsack, J., Bowers, C.: Patterns of Gaming Preferences and Serious Game Effectiveness. In: Shumaker, R. (ed.) Virtual and Mixed Reality, Part II. LNCS, vol. 6774, pp. 37–43. Springer, Heidelberg (2011)
9. Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (2003)