International Journal of Software Engineering and Its Applications
Vol. 9, No. 4 (2015), pp. 171-180
http://dx.doi.org/10.14257/ijseia.2015.9.4.18

An Efficient Modeling and Dimensioning Approach of Internet Business Area based on OLAP

Kil Hong Joo[1] and Nam Hun Park[2*]

[1]Dept. Of Computer Education, Gyeongin National University of Education, San 6-8 Seoksudong Manangu Anyangsi, Gyeonggi, Korea, 430-040
[2]Dept. Of Computer Science, Anyang University, 102 Samsungli, Buleunmyun, Ganghwagun, Incheon, Korea, 417-833
*[1]khjoo@ginue.ac.kr, [*2]nmhnpark@anyang.ac.kr*

Abstract

Domain registration organizations have long recognized the importance of their customers. For domain registration service providers to match the domain registration with registrants' propensity and company service requirements is still a challenging issue. This research seeks to suggest an analytical method that can identify in which market domain registration organizations should secure potential group of customers who could contribute to higher domain sales while supporting their decision making process by linking their own possessed domain registration information with external attribute data on in which business areas the registered domains are utilized online. Through experiments, several business areas were found to have consistent trends and changes with those in overall market internet business areas and internet business area changes as desired to find in this research.

Keywords: *Data mining, OLAP, Data cube*

1. Introduction

Domain registration organizations have long recognized the importance of their customers. However, since the product, domain largely relies on the purpose of use of a company a registrant belongs to in its nature, other than individual purpose of use, domain marketing hardly targets individuals. Domain registrants are not any specific group of people and their companies also provide far diversified services. In this situation, it is extremely difficult for domain registration service providers to match the domain registration with registrants' propensity and company service requirements in performing CRM [1]. Moreover, the information possessed by domain registration organizations is limited only to domain registration-related aspects without those on how the registered domain is used in what kind of services online or how to classify information for management, etc. It is true that such organizations, even though they manage key information – address sources – for online business, they have not utilized it more effectively. To overcome the situation, they would need key information that encompasses aspects about a registrant, enterprise and online business altogether. By analyzing such information, domain registration organizations can anticipate online business trend earlier than any other and perform CRM (customer relationship management) for customers in the business areas expected to change [1,2,3].

This research seeks to suggest an analytical method that can identify in which market domain registration organizations should secure potential group of customers who could contribute to higher domain sales while supporting their decision making process by

[*] Corresponding Author

ISSN: 1738-9984 IJSEIA

linking their own possessed domain registration information with external attribute data on in which business areas the registered domains are utilized online.

The analytical method herein is to be presented via OLAP technique starting from the decision support system (DSS)[4,5]. An analyst searches data cube by using OLAP process and finds exceptional data areas. This method is called a "hypothesis-driven" search method[6]. As the number of dimensions becomes larger, data cube search range also grows larger, limiting the method in searching for exception data. As for an alternative, a "discovery-driven" method was suggested for data cube search [7]. The method uses a pre-calculated measurement representing data exceptions in all dimensions as a guide for its data cube search. Any cell having a huge difference from its estimated value in data cube is regarded as an exception. This method increases the possibility that users find out abnormal patterns (exceptions) at a random set level. The method works more favorably the more the number of dimensions and the larger the hierarchy structures of each dimension [7,8].

This research is structured as follows: In Chapter 2, related studies are described. Chapter 3, to find out online business area changes, explains a data cube modeling method and data warehouse generation method and the criteria for exception search while presenting OLAP-based system design where related techniques were applied. Chapter 4 compares general statistical data and interpretation with the results of the discovery-driven method-using OLAP system and verifies them. Chapter 5 presents this research conclusion and follow-up study plans.

2. Modeling of Internet Business Area Dimension and Database

To identify and analyze changes in the domain-based internet business area, the relation OLAP was employed in this research. ROLAP system needs data based on domain data warehouse which have been accumulated for years. And to connect the changes in internet business areas in its analysis, the system needs separated external data accumulation. The external data, here, should have domain keys and attribute data on the internet business use area. Such data should be stored in an automated system. The accumulated internet business attribute data are linked to the basic domain data warehouse to provide multidimensional data structures for ROLAP. The basic domain data and internet business area information are linked by the domain values of each information set as a connection key. Corresponding data are collected to find out an exceptional or abnormal pattern by analyzing the relationship among such information. In this thesis, data cube was used for multidimensional data modeling and recognition. Figure 1 represents 4D data cube consisted of product, region, internet business and point of registration. If a dimension is given, cuboid lattice can be formed and each lattice represents data at a summary level. Figure 2 exhibits cuboid lattice for data cube having region, internet business area and registration point as its dimensions. In Figure 2, the base cuboid is the 3D cuboid consisted of region, internet business area and registration point and the 2D cuboid consisted of internet business area and registration point is the summary of region. The 0D cuboid has the highest summary level. In Figure 2, the 0D cuboid is the summary of all of the 3 dimensions and marked as all.

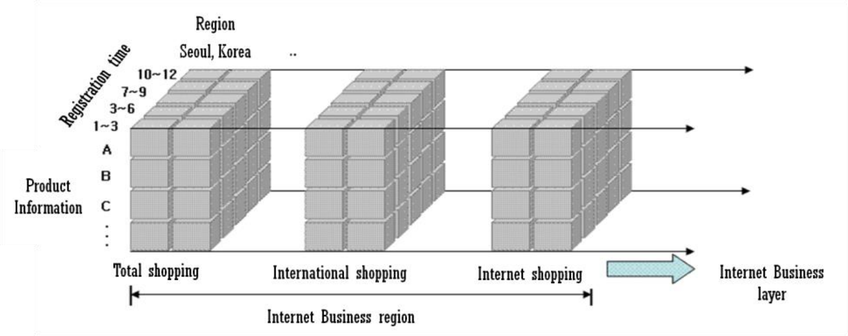


Figure 1. Data Cube of Internet Business Region based on Domain

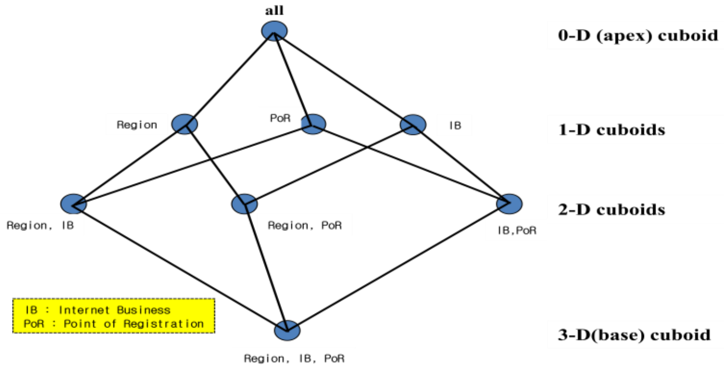


Figure 2. Cuboid Lattice

To examine internet business area changes, a separate database needs to be built based solely on better organized and more necessary data among the initial data. Therefore, it is fundamental to better organize product data on newly-generated domains having a registration point and region data possessed by the domain registering customers (enterprises). It is also vital to form accurate connection with internet business area data generated separately from outside. Therefore, to meet such conditions, this research defined the mutual relationship between product data and data on registration point, region and internet business area in a star schema structure as shown in Figure 3. Here, each data represents one single dimension in OLAP. Actually when analyzing changes in domain-based internet business areas, such data help provide the results from diverse changes in internet business areas according to registration point, region and product type. For multidimensional analysis of internet business area changes, each dimensional tables are based to form fact tables of domain-based internet business areas. And based on the fact tables, integrated tables are formed for subsequent OLAP analysis.

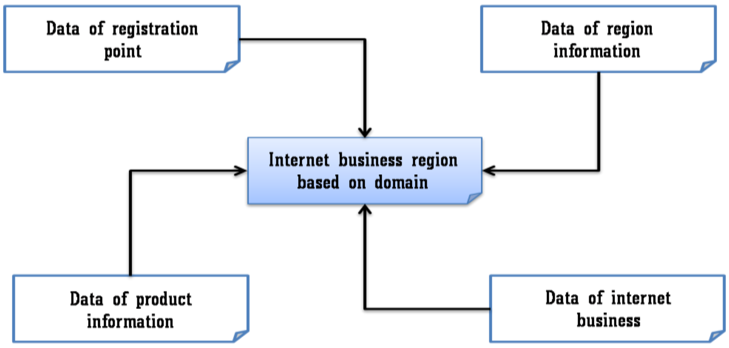

Data of registration point
Data of region information
Internet business region based on domain
Data of product information
Data of internet business

Figure 3. Star schema to structure OLAP

Figure 4 shows the structure of integrating each dimensional tables consisted of region table, product table, registration point table and internet business area information table through fact tables. The basic key values of the dimensional tables are referred to as external keys of fact tables. The baseline dimension for the analysis should have a well-structured hierarchy of internal temporal levels or substructural hierarchy. Data such as domain registration point should be stored in a manner that facilitates value extraction from area change with time and exception occurrence in relation to internet business area. Data on region should also have a well-structured hierarchy.

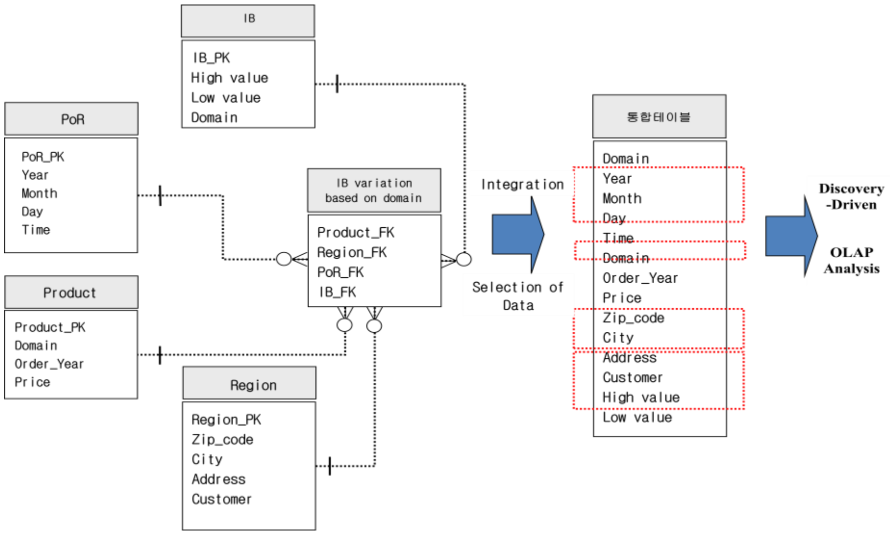
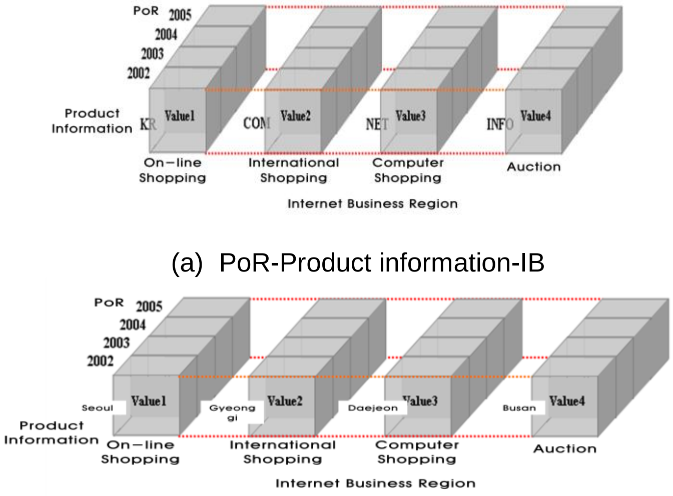


Figure 4. Schema of Internet Business Based on Domain

3. Multidimensional Query and Search in Internet Business Area

Multidimensional database is designed from a user's perspective according to its ways of use. Users can analyze an issue from diverse angles through calculations such as pivoting, drill-down, drill-up, etc. By considering user characteristics, the database is capable of providing strategic alternatives where analysis strategies and appropriate feedback are offered in a multidimensional way. Users can cut part of cube through a slicing and dicing method and process data more significantly from their perspective. Also through pivoting, users can change report rows and columns and page dimensions randomly to gain their desired types of information in diverse forms. Users can also easily perform drill-down from upper-layer items to lower-layer items or drill-up from lower to upper layers according to the hierarchy structure between each dimensional items. User can gain one single cell value as a result of multidimensional query or gain 2D or 3D or more dimensional sub-cube. Unlike conventional application which simply presents formalized reports on a fixed screen, it stresses a conversation formation with users. Figure 5(a) represents multidimensional query asked regarding registration point, product information and internet business area information. Figure 5(b) represents multidimensional query asked to gain necessary information from registration point, region information and internet business area information. The discovery-driven method uses an exception indication that supports data analysis process in all set levels. This study defines the exception indication applicable to data cube for analyzing internet business area number changes as follows through the equation presented in [9].



(a) PoR-Product information-IB

(b) PoR-Region-IB

Figure 5. Multidimensional Query

Definition1. Measurement for change analysis

In data cube *C* having n dimensions, *r* dimension $d_r(1 \le r \le n)$'s value at i_r location is $y_{i_1,i_2,\cdots,i_n}$, Then the expected value $\hat{y}_{i_1,i_2,\cdots,i_n}$ is defined as follows.

$$\hat{y}_{i_1,i_2,\cdots,i_n} = f(\gamma^{G}_{(i_r|d_r \in G)} | G \subset d_1, d_2, \cdots, d_n) \qquad (1)$$

Here, the γ value is a threshold value. Users can adjust the value to gain diverse results. ■

To more clearly explain Definition 1 with a cube comprising A, B and C dimensions as an example, the expected values of *i*th cell in A dimension, *j*th cell in B dimension and *k*th cell in C dimension $\hat{y}_{ijk}$ are expressed to as a function with 7 items.

$$\hat{y}_{ijk} = f\left(r, r_i^A, r_j^B, r_k^C, r_{ij}^{AB}, r_{jk}^{BC}, r_{ik}^{AC}\right)$$

Definition 2. Exception indication for change analysis

Of the intrinsically generalized difference values, a specific cell value which is relatively larger, can be viewed as an exception. Statistically the status of being relatively larger is judged by comparing with the values of expected standard deviation $\delta_{i_1,i_2,\cdots,i_n}$ calculated as the difference. Therefore, if the standardized differences are larger than a certain threshold value γ, they are defined as a supercritical value and if smaller, as an under critical value. Exception value $s_{i_1,i_2,\cdots,i_n}$ is defined as follows:

$$s_{i_1,i_2,\cdots,i_n} = \frac{\left| y_{i_1,i_2,\cdots,i_n} - \hat{y}_{i_1,i_2,\cdots,i_n} \right|}{\delta_{i_1,i_2,\cdots,i_n}} \qquad (2)■$$

In this paper, 2.5 was used fundamentally as the value falls under 99% probability in normal distribution. So in this research experiment, exception values are checked, which are drawn according to the value changes. Also, in data cube search, the following 3 variables are used:

(1) SelfExp : a value representing which cells are exception at the current level. If it turns out to be super/undercritical value by the equation (2), it becomes an exception herein.

(2) InExp : a value representing the degree of exception of all cells within the reach of drill-down calculation in the current cell. It is defined as the max SelfExp values of all cells under the current cell.

(3) PathExp : a value representing expected degree of exception when performing drill-down in a specific path among the possible drill-down paths in the current cell. It is defined as the max SelfExp value of cells within the reach of drill-down in the current cell.

4. Structure of Internet Business Area Changing System

The change identification system of internet business area is consisted of the linkages between basic domain information and internet business information. Therefore, the research refines basic domain information and internet-based information to generate data warehouse for each and links it to build a final data warehouse applicable to OLAP system analysis. Each generated data warehouse is consisted under the assumption of automation. It monitors the operated automation process by regular manager control and receives maintenance for a possible exceptional event. In general, customers or enterprises providing online services for a specific internet business open their homepages in the internet after registering domains with a domain registration organization. Such registered online homepages include their internet business categories in online contents portal services such as portal sites or Ranky.com for online advertisements like search keywords. In the OLAP environment, the domain-based business area change analysis system as suggested in this research, conducts preliminary processing as to the errors of domain information and customer information registered with the domain registration system then generates basic domain information (region, product and registration point) basically to accumulate data for OLAP analysis system use. From the generated basic domain information, domain information is extracted every certain interval of registration point. Internet business area data are extracted from diverse portal sites and Rankey.com which provide internet business directory service by using the screen scrapping technique. Then those are saved in the databases inside OLAP system. The accumulated information is each table to be used as a dimension. After forming an integrated table, OLAP-based multidimensional analysis is performed in the manner explained in Figure 6.

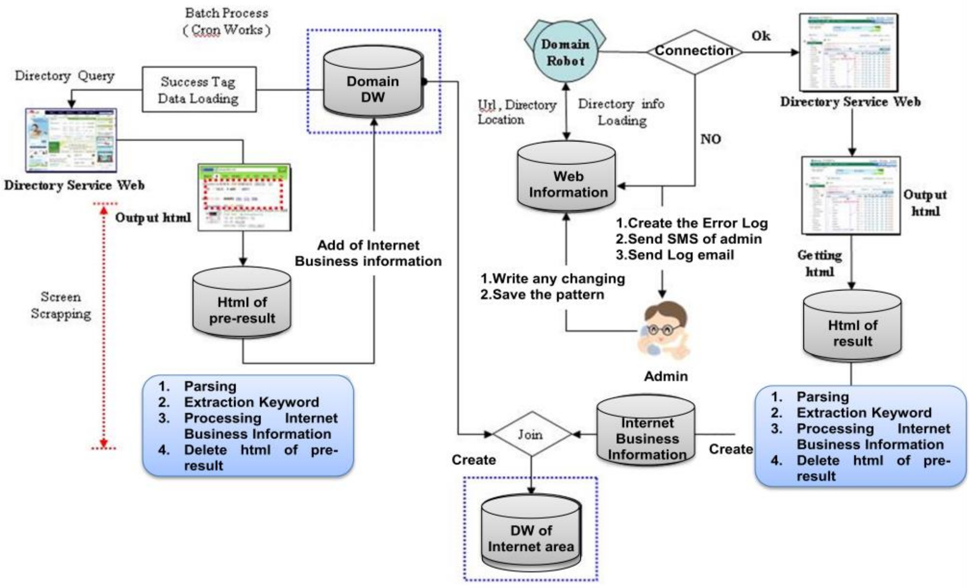


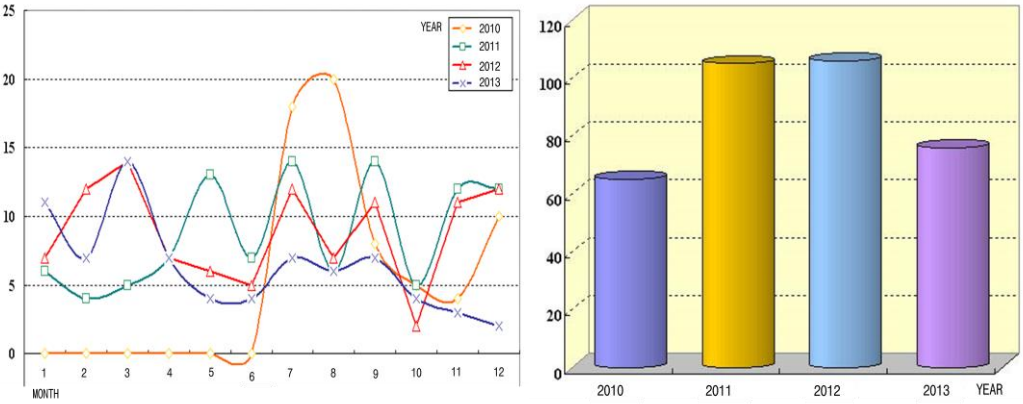
Figure 6. Flow Chart of Internet DW

In general, customers or enterprises providing online services for a specific internet business open their homepages in the internet after registering domains with a domain registration organization. Such registered online homepages include their internet business categories in online contents portal services such as portal sites or Ranky.com for online advertisements like search keywords. In the OLAP environment, the domain-based business area change analysis system as suggested in this research, conducts preliminary processing as to the errors of domain information and customer information registered with the domain registration system then generates basic domain information (region, product and registration point) basically to accumulate data for OLAP analysis system use. From the generated basic domain information, domain information is extracted every certain interval of registration point. Internet business area data are extracted from diverse portal sites and Rankey.com which provide internet business directory service by using the screen scrapping technique. Then those are saved in the databases inside OLAP system. The accumulated information is each table to be used as a dimension. After forming an integrated table, OLAP-based multidimensional analysis is performed.

5. Experimental Results

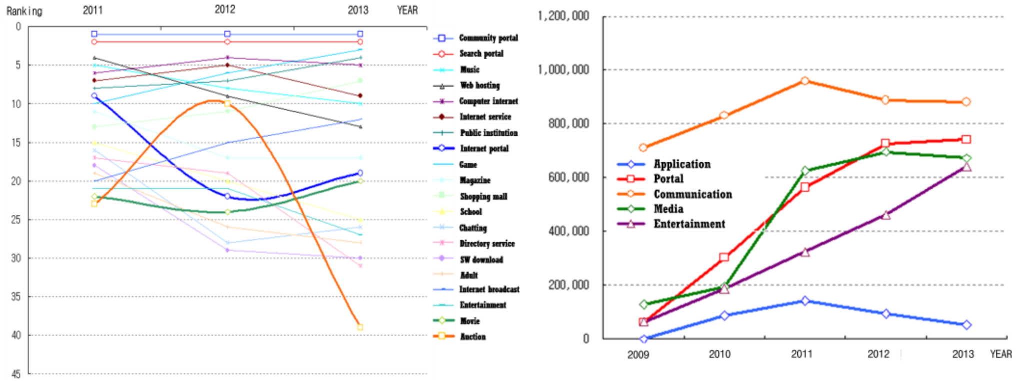
In this research, 1,719 sets of domain-based internet business data, which were linked and integrated by the key of domain in the internet business area data provided by domain data warehouse of company ***I***, the certified national KR domain registration organization, and www.Rankey.com.The data were used in t his research experiment via OLAP site where data cube discovery-driven search algorithm was applied. During the experiment, the threshold value was reduced by 0.1 from 2.5 representing 99% probability in normal distribution to 2.0 to measure expected values and exception values. Data cube was consisted of 3 dimensions < registration year, region, internet business area>, <registration year, registration month, internet business area > and<product, registration year, internet business area>. To verify the method suggested herein, internet statistically analysis was performed and its results [9] were compared. 3D data cube search was performed by using the domain internet business area data employed herein. As for registration point, since fewer than normal amount of data were collected in 2013, the exception cells in 2013 were decided not to be considered herein for they cause large gaps with the 2012 data and earlier.

The over-expressed cases viewed from the 2011 Seoul, shopping (living_general) cell were compared with general statistical analysis. As a result, it was found that Figure 7(a) represents per year/per month registration numbers based on the shopping (living_general) area. But exceptional patterns are hardly found in the per year/month graphs. As for Figure 7(b), however, if we look at the per year registration graph, registration increased upwards from 2010 to 2011 and stagnated in 2011 and 2012 with similar registration numbers. Here, the year 2010 when the stagnation started can be viewed as an year of exception.



(a) per year/per month (b) per year/per month

Figure 7. Registration numbers based on the shopping (living_general) Area



(a) Ranking of internet business area (b) Trends of sales

Figure 8. Trends of Internet Business Area

Figure 8 exhibits the trends of internet business area. The market sales trend and use visit ranks of some internet business area are compared with the interpretation of Figure 7. As a result, the porter service area market sales which had surged until 2011 as in Figure 8(b), stagnated until 2012. This phenomenon is consistent with the yearly registration trend in Figure 7(b). Also the auction business area in Figure 8(a), which turned to decrease in 2013, is a sub-area of portal service area with consistent trend of registration in portal (living_general) business area in Figure 7(b).

6. Conclusion

This research suggested an OLAP system design applying data cube measure and discovery-driven search method to identify changes and expectations of domain-based internet business areas. Data cube measure and indication values were presented herein as those that could resolve the problems of statistical method. Their appropriateness was presented after comparing with the general statistical analysis. The threshold values of the suggested measure and exception indication can be adjusted according to a user’s assessment. So the system is capable of identifying changes in various areas based on experiments using different threshold values. Based on the significance and background of data in each data cube dimension, more appropriate company marketing strategies would be found beyond just simple exception identification. Furthermore, it was found herein that the system could anticipate internet business area changes. The data used in the actual experiment may not be sufficient enough to anticipate diverse internet business area changes.

And as they do not have many effective numbers to be defined as a dimension, those data failed to make a good use of data cube. However, several business areas were found to have consistent trends and changes with those in overall market internet business areas. It was also found possible to anticipate internet business area changes as desired to find in this research. The more the number of data, the more reliable the analysis becomes. So more significant examination would be possible if enterprises accumulate more information in the future along with sufficient amount of domain-related internet business attribute data. Data analyzed in this manner would help OLAP-using firms predict customer loss and internet business area changes earlier than other firms so that they can engage in marketing efforts in the corresponding business area to attract customers early. Also, such OLAP is expected to be applicable to other similar services to the domain registration such as hosting and homepage production, etc.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science Technology(NRF-2014R1A1A1008524).

References

[1] Chen, Qimei, Chen, Hong-Mei. Exploring the success factors of eCRM strategies in practice. The Journal of Database Marketing & Customer Strategy Management, vol. 11, no. 4, pp. 333-343 (2004).

[2] Balaji Padmanabhan, Alexander Tuzhilin. On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities, vol. 49 no. 10, pp. 1327-1343 (2003).

[3] Firoozeh Fouladivanda, Mahsa Mohseni, Mani Shehni Karam Zadeh, Aidin Barbat. A Study on the Relation between Electronic Customer Relationship Management (ECRM) and Customers Loyalty in the International Market, Life science journal, vol. 10, no. 12, pp. 353-359 (2013).

[4] Gaurav Gupta, Himanshu Aggarwal. Overview of CRM using Data Mining, International Journal for Multi Disciplinary Engineering and Business Management, vol. 1, no. 1, pp. 15-20 (2013).

[5] Haluk Demirkan, Dursun Delen. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems, vol. 55, no. 1, pp. 412-421 (2013).

[6] Einoshin Suzuki. Data Mining Methods for Discovering Interesting Exceptions from an nsupervised Table, Journal of Universal Computer Science, vol. 12, no. 6, pp. 627-653 (2006).

[7] Arnaud Giacometti, Patrick Marcel, Elsa Negre, Arnaud Soulet. Query recommendations for OLAP discovery driven analysis, Proceeding of the ACM 12th international workshop on data warehousing and OLAP, pp. 81-88 (2009).

[8] J.Gray, S.Chaudhuri, A.Bosworth, A.Layman, D.Reichart, M.Venkatrao. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Data Mining and Knowledge Discovery vol. 1, pp. 29-53 (1997).

[9] S.Sarawagi, R.Agrawal, N.Megiddo. Discovery-driven Exploration of OLAP Data Cubes, Research Report RJ 10102(91918), IBM Almaden Research Center (1998)

Authors

Kil Hong Joo, received the M.S. and Ph.D. degree in Computer Science from Yonsei University, Seoul, Korea, in 2000 and 2004. He is currently a professor of Department of Computer Education at Gyeongin National University of Education, Korea. His current interests include mining big data, data analysis and smart learning.

Nam Hun Park, received the B.S., M.S. and Ph.D. degree in Computer Science from Yonsei University, Seoul, Korea, in 2000, 2002 and 2007. He was a post-Ph.D. at the Department of Computer Science, Worcester Polytech Institute, Worcester, MA. He is currently a professor of Department of Computer Science at Anyang University, Korea. His current interests include mining data streams.