A Research on Constructing Consumption Pattern Analysis System through Purchase Records

Hoo-Young Lee, Koo-Rack Park* and Dong-Hyun Kim

1Department of IT Engineering, Kongju National University, 31080, Rep. of Korea; hooyoung.paul.lee@gmail.com, ecgrpark@kongju.ac.kr
2Department of IT Convergence, Woosong University, 34606, Rep. of Korea; dhkim@wsu.ac.kr

Abstract

Objectives: To understand the economic indicator through purchase records. Methods/Statistical Analysis: In this thesis, a system analyzing the economic indicator of the individuals is proposed. The purchase information such as receipt can be either directly entered into the application by the users or entered into the application through the actualized optical character recognition technique. The purchase records such as receipt of the individuals can be entered into the application to be saved in the database and the saved data can be analyzed to analyze the economic activities conducted by the individuals. Since such system can be used to confirm the amount of money spent as well as the particular field the consumption leans toward, the individuals are able to set up their future economic plan. Findings: The proposed system visualizes the detailed information of the economic activities conducted by the individuals based on its database in which diverse purchase records such as receipts that occur during the economic activities conducted by the individuals are saved. The preexisting researches were focused on using the data-mining technique to develop a system that recommends products to customers by predicting their purchase intention. However, most of these researches were conducted to support companies to promote efficient marketing and therefore were unsuitable for providing the integrated economic indicator to the individuals. Accordingly, the proposed system can be used to analyze diverse consumption patterns based on the purchase records of the users and the analyzed result can be used as an important indicator of the individuals’ economic activities. Improvements/Applications: The proposed system can be valuably used to understand the economic indicator of the individuals. In the future research, the accuracy of the optical character recognition must be studied and the big data analysis must be studied as well.

Keywords: Consumer, Consumption Plan, OCR, Purchase, Receipt

1. Introduction

In modern society, individuals have been conducting their economic activities through diverse payment systems such as credit card, debit card, mobile micro payment and cash. As for such economic activities, the development of IT technology has changed the consumption pattern of the consumers and such consumers not only usually conduct cognitive purchases through the use of logical methods based on information and alternatives but also usually conduct habitual purchases as well. In particular, the purchase behavior of such consumers is usually based on the brand preference. In addition, there are services provided through the use of mobile application program channels related to the brands to promote marketing strategies and the consumers are lured into spending more than required through such services. To limit such excessive consumption, respective credit card companies or financial companies provide the indicator of the individuals’ economic activities through collecting and analyzing the purchase records of their customers. Individuals can analyze the provided indicator to understand the size of their past spend and limit excessive consumption and such indicator can be used as the
A Research on Constructing Consumption Pattern Analysis System through Purchase Records

basic data for the future economic activities. However, although such consumers require the integrated data on diverse forms of their economic activities, such integrated data are not actually provided.

In this thesis, a system providing the visualized economic activity indicator to each individual through using the Optical Character Recognition (OCR) to extract/input/analyze the text data collected from either the receipts directly entered in by the individuals or the scanned receipt images is proposed. It is expected that the consumers through such proposed system will be able to acquire the integrated information on their overall economic life and use such acquired integrated information as the important ground data for their rational economic life.

1.1 Tesseract

Tesseract is an open source OCR engine developed by Hewlett and Packard for 10 years from 1985 to 1994. After the development, Tesseract was announced as an open source in 2005 and it has been developed by Google since 2006\(^4\). The following Figure 1 shows the operating process of Tesseract\(^5\).

![Figure 1. Operating process of Tesseract.](image)

Initially, the image threshold is used to convert the input image into the binary image. Secondly, the connected component of the binary image is analyzed to extract outline of each component and such extracted outline is saved as the binary data. Thirdly, the text line is analyzed to systematize the components and organize them into word units according to the space between characters. Finally, the components organized into words are recognized as word units and page units.

Since Tesseract does not have a specific output file format, a stage for producing the output file does not exist. The recognition level of Tesseract can be decreased due to lighting, angle and font\(^6–8\).

Basically, Tesseract consists of C++ Library. However, Tess4 usable in JAVA environment and Tesseract-Android-Tools serving as the android library are provided to allow diverse actualizations.

1.2 Optical Character Recognition

OCR is a technique that converts/saves Korean, English and numeric fonts into editable texts through scanning printed documents/images. As for the history of OCR, it was initially introduced in 1928 when G. Taushek from Germany registered a patent for the character recognition method in which the pattern-matching technique is used to compare the prepared standard pattern characters with the input characters to select the input characters most similar to the standard pattern characters as the involved characters. OCR is applied to diverse fields.

It is widely applied to banking, medical industry, legal industry involving security character, bin, digital library, character recognition requiring no corrections/efforts by human beings, auto number plate recognition and writing recognition\(^9–12\). The character recognition methods are divided into template matching method, statistical method and structural analysis method based on how the character pattern is expressed and classified. The template matching method is used to find the most similar form by classifying the character patterns into the array forms and comparing them to the template pattern. This method was frequently used in the earlier stages since the characters used at the time were usually based on one form. However, due to occurring problems, it is not frequently used these days\(^13\).

The statistical method is used to recognize the characters through extracting particular vectors from the recognition target. In this method, the statistical probability distribution characteristics of particular vectors are calculated through learning stages and such characteristics are used to divide particular vector spaces per each classification. Such classification model is mathematically well-defined and the matters considered important in
such method are how well-defined or well-expressed the characteristics of the input pattern are and how such characteristics are extracted\(^ {14}\). The structural analysis method is used to recognize the characters through extracting the connectivity among base elements such as stroke that construct the characters based on the fundamentals of the character construction. This method is theoretically well-established and its strength is that the method itself is simple. However, its weakness is that the rules applied to particular characters can be diversified depending on the type font and therefore the recognition takes considerable amount of time\(^ {15}\).

Recently, a research on the neural networks model, one of the AI systems for recognizing the character patterns, has been attracting the interest. Such neural networks model models the structure of human brain and simply connects neuron, a standard unit, to the calculation element to result in outstanding performance. Accordingly, such neural networks model is suitable for syntropic problems requiring various calculations such as voice/character/image analysis\(^ {16}\).

### 2. Proposed Work

#### 2.1 System Diagram

The diagram of the system proposed in this thesis is as shown in Figure 2.

![Figure 2. System diagram.](image)

To conduct the experiment with the proposed system, an application exclusively operating on the user’s Smartphone was actualized. Such exclusive application features two major functions. Its initial function allows the user can directly input the purchase records to the database through the input UI or scan the purchase records to upload them on the server. Its second function displays the data which confirms the result analyzed from the server. The OCR program extracts the user’s purchase records from the image data saved on the server and then saves the result on the database. Based on such produced data, the consumption-concentrated categories and purchase amount information are extracted from the user’s purchase records and visualized into the chart data to be provided to the user.

#### 2.2 System Process

Figure 3 shows the process of the proposed system.

![Figure 3. System process.](image)

The proposed system consists of 3 major stages. In the initial stage, the purchase records are input. The purchase records are directly input or the scanned receipt files are uploaded on the server through the exclusive application and the data are input through the OCR-based input method. At this point, already prepared images can be uploaded or the camera application can be used to film the images to be uploaded. The uploaded images are saved on the pre-defined server as JPEG (Joint Photographic Experts Group) format. In addition, the metadata such as file size, file name, saved date and file directory are saved on RDBMS (Relational Database Management System).

In the initial stage, the purchase records are converted into the database. The user can either directly input the purchase records through the input UI (User Interface) or
input the text data extracted from the uploaded image to the database. In third stage, the extracted data are visualized. The text data saved in the database include diverse data such as name of the goods purchased, purchased date, used amount and name of the card used. Based on the count of such data, the count of the items purchased per category is extracted and the sum of the amount spent for purchasing goods per period is calculated. If necessary, the required data can be extracted through the use of aggregate functions such as AVG, MAX and MIN to visualize the purchase records into pie/column charts to be provided to the user.

### 3. Experiment and Consideration

To actualize the system to be constructed in this thesis, one computer consisting of Intel Core i3 CPU 3.10 GHz and 4G RAM was used to construct a server for OCR and data visualization. The Ubuntu-14.04 was installed as the OS and the JAVA version used for developing the application was 1.8.0_74 and the MySQL-Server 5.6 version was used as the DB server.

#### 3.1 Mobile Application

The user can input data and confirm analyzed data through the mobile application actualized for testing as shown in Figure 4. The user can use the UI shown in Figure 5 to directly input the purchase records to the database according to the appropriate form. The input data is then saved on the database through the data verification process. Figure 6 shows another input method. The purchase record data such as receipt can be filmed into photos and converted into images to be uploaded on the server. Then the purchase records can be extracted through the OCR technique and the respective information can be saved on the database. The analyzed data are provided to the user in a chart form as shown in Figure 7.

#### 3.2 OCR Conversion Program

The following Figure 8 shows the conversion code using the OCR technique.

The program for extracting the text from the uploaded receipt images was actualized through the use of JAVA and tess4j-3.0.jar, an open source-based OCR library was used. The developed program communicates with the database of the server where the images are saved every second to confirm whether or not there are new image data. If there are new image data, it receives the location information of the involved image files to approach the involved files, extract the text data and save the data back on the database.
3.3 Construction of Database

The structure of the database table is as shown in Figure 9. The table consists of Member Table where the user information is saved, Image Table where the image meta information is saved, Receipt Table where the purchase records based on the text extracted from the input receipt data are saved and Category Table where the product classification information is saved.

3.4 Data Analysis Experiment

The proposed system can be used to confirm the particular fields the consumption is concentrated on and how much money is spent by the principal conducting the economic activities. In addition, the analysis can be diversely processed through classifying period, item and payment system used. In this thesis, to demonstrate such analysis, the spend frequency and spend amount data collected from the user for three months were analyzed per each category. To conduct such experiment, the purchase categories were classified into 14 categories including food, housing and comm. as shown in Table 1.

As for the spend frequency analysis, as shown in Figure 10, the group by function was selected and used from the aggregate function of the database to count the data of the involved category from the input data to extract the involved frequency. As a result, as shown in Figure 11 and Figure 12, Food (68.2%) indicated the highest frequency and was followed by Traffic (16.2%) and Housing (3.9%). In addition, as for the spend amount per category, Food (1,690,710 won) indicated the highest spend amount and was followed by Cloth (314,100 won) and Comm. (293,990 won). Through such analysis, it was found that the spend frequency was higher for Traffic and Housing Categories and that the spend amount was relatively higher for Cloth and Comm. Categories.

SQL for calculating the percentage of the spend item per category from the input database is shown in Figure 10.

---

**Table 1**

| Member_Table      | Image_Table        |
|-------------------|--------------------|
| member_id         | member_id          |
| member_name       | image_seq          |
| member_gender     | image_path         |
| member_address    | image_name         |
| member_phone      | image_size         |
|                   | ...                |

| Receipt_Table     | Category_Table     |
|-------------------|--------------------|
| member_id         | category_code      |
| image_seq         | category_name      |
| receipt_seq       | category_use       |
| category_code     | category_regist_date|
| receipt_data_01   | ...                |
| receipt_data_02   |                    |
The purpose of this research is to actualize a system that provides the spend amount information to the user through using the purchase records of individuals/households serving as the principal of the economic activities either directly input or extracted through OCR as the basic data for setting up the database, classifying them into pre-defined categories and analyzing the spend-concentrated items. Such analyzed results are expected to be used as the ground data for the economic activities of individuals/households and are expected to make contributions to their rational consumption life.

4. Conclusion

The purpose of this study is to actualize a system that visualizes detailed records of the individuals’ economic activities based on the database in which the diverse purchase records (receipts) occurring in their economic activities is saved. The preexisting researches were either focused on analyzing the purchase data specified in receipts to develop the CRM (Customer Relationship Management) system or using the data-mining technique to develop a system that recommends suitable products to customers by predicting their purchase intention. However, such researchers were conducted to support companies to promote efficient marketing and therefore were unsuitable for providing the integrated economic indicator to the individuals. This proposed system can be used to analyze diverse consumption patterns based on the purchase records of the users and the analyzed result can be used as an important indicator of the individuals’ economic activities. However, the issue involved in this research is that the text recognition rate must be improved. The text recognition accuracy varied depending on the quality of the receipt images filmed through Smartphone cameras.

The future researches must be conducted to improve the accuracy of the OCR system based on the current system proposed. In addition, a method for analyzing the

| No. | Category | Code |
|-----|----------|------|
| 1   | Food     | C001 |
| 2   | Housing  | C002 |
| 3   | Comm.    | C003 |
| 4   | Cloth    | C004 |
| 5   | Beauty   | C005 |
| 6   | Health   | C006 |
| 7   | Culture  | C007 |
| 8   | Education| C008 |
| 9   | Infant Care | C009 |
| 10  | Traffic  | C010 |
| 11  | Vehicle  | C011 |
| 12  | Dues     | C012 |
| 13  | Tax      | C013 |
| 14  | Unclassified | C014 |

Figure 10. Category data extract SQL.

Figure 11. Category purchase pattern chart.

Figure 12. Category spend chart.

Table 1. Category classification
purchase records through the big data system must be studied as well.

5. References

1. Seo GD, Lee JE. A study on the effect of consumer lifestyle on brand attitude, brand attachment influence upon brand loyalty. Journal of Digital Convergence. 2016; 14(4):185–92.

2. Vijay AV. A study on consumer brand preference with reference to DTH in rural area. Indian Journal of Science and Technology. 2016 Jul; 9(27):1–5.

3. Yang BH. A link between consumer empathy and brand attachment on branded mobile Apps: The moderating effect of ideal self-congruence. Indian Journal of Science and Technology. 2016 Jul; 9(25):1–9.

4. Tesseract. 2016. Available from: https://github.com/tesseract-ocr/tesseract

5. The Tesseract OCR Engine. 2007. Available from: http://static.googleusercontent.com/media/research.google.com/en/pubs/archive/33418.pdf

6. Go EB, Ha YJ, Choi SR, Lee KH, Park YH. An implementation of an android mobile system for extracting and retrieving texts from images. Journal of Digital Contents Society. 2011; 12(1):57–67.

7. Choi SR, Go EB, Ha YJ, Park YH. An implementation of a system for extracting and retrieving texts on android platform using OCR. Proceeding of Korea Multimedia Society; 2010. p. 469–73.

8. Kim JW, Kim SH, Lee CW. Construction of educational contents using on-line character recognition. Proceeding of Korea Multimedia Society; 2009. p. 309–12.

9. Singh A, Bacchuwar K, Bhasin A. A survey of OCR Applications. International Journal of Machine Learning and Computing. 2012; 2(3):314–8.

10. Mithe R, Indalkar S, Divekar N. Optical Character Recognition. International Journal of Recent Technology and Engineering. 2013; 2(1):72–5.

11. Shrivastava V, Sharma N. Artificial Neural Network based Optical Character Recognition. Signal and Image Processing: An International Journal. 2012; 3(5):73–80.

12. Kwon SK, Jo SH. Community-based travel information system using augmented reality. Journal of Korea Multimedia Society. 2015; 18(2):97–105.

13. Kanti VM. Statistics and Images. Oxford: Cartax Publishing Company; 1995.

14. Cho WH. Introduction and development of a new recognition of the Character Recognition System. Proceedings of the Spring Conference Korea Statistical Society; 1998. p. 68–74.

15. Cho YY. Artificial intelligence system. Seoul: Hongrung Publishing Company; 2003.

16. McClell JL, Rumelhart DE. Learning internal presentation by error propagation. Parallel Distributed Processing. 1986; 1:318–62.

17. Lee SJ, Kim JH. Prediction of purchase with a receipt of the customer data and SVM. Proceeding of Korea Intelligent Information System Society; 2006 Nov. p. 179–88.