ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 10 (2020) No. 2, pp. 40-52
DOI:10.33168/JSMS.2020.0203

A Big Data Based Cosmetic Recommendation Algorithm

Jiyoung Yoon and Soonhee Joung*
Department of Consumer Studies, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Korea
*graceyoon928@ewha.ac.kr, *jsh@ewha.ac.kr (corresponding author)*

Abstract. The purpose of this study is to develop a recommendation system to help consumers who want to purchase cosmetics to choose cosmetics more easily and comfortably. For this purpose, the cosmetics classification of 'Hwahae App', is cosmetics application in Korea, was used and developed 'recommendation system based on similarity algorithm'. This study conducted a previous study on the algorithms that make up types and recommendation systems based on Big Data. Among the numerous cosmetics, the data on consumer choice attributes and types of cosmetics were collected through the production of 'crawling Bot'. Then, the frequency of word appearance between documents was confirmed for the unstructured data information and the skin type variables of consumers. Finally we designed a system that recommends the top five products that most closely resemble the desired product, through the combination of the selection attributes that consumers want most. This study has practical value to help customers and academic meaning of recommendation system using bigdata.

Keywords: Recommendation algorithm, cosmetics consumer, big data, similarity algorithm.

1. Introduction

Recently, the amount of information has increased dramatically as the IT technology industry has developed rapidly. With so much information being poured in, interest in more effective technologies is also increasing in analyzing and understanding the information we need. And we live in newly released products, newly created products and goods every day.

In fact, every time new goods are poured out, we spend monetary and temporal costs such as purchasing samples before purchasing products, asking for feedback

after use comments using advertisements, Blogs, SNS, etc (Jung, 2017). The 'big data system', which can utilize a lot of data meaningfully, is expected to be more important technology not only in modern society but also in future society (Yoon and Joung, 2019). And it is expected to be a key technology that can change human life innovatively (Yun and Youn, 2017).

Big data is not simply a technological advance or trend. It can be seen as a kind of decision-making action that effectively extracts and utilizes enough data about decision. The field of big data applications in the evolving smart era will be quite wide. In this study, the recommendation system considering consumer's taste among various big data utilization technologies is to be examined. Big Data based recommendation systems have revolutionized customer marketing methods (Yun and Youn, 2017).

In the past, the consumer marketing technique was provided to all consumers in a lump sum without being subdivided, but now the actual service considering the individual characteristics and preference of consumers is possible by providing a vast amounts of information through big data. The big data-based recommendation system, which can be useful in everyday life, can be regarded as a very innovative technology benefit (Madden, 2012). So, the recommendation system is the only way to solve any of these problems. Recommendation systems or recommended algorithms automatically process these processes, allowing consumers to recommend items and information they want, reducing the burden and time of having to click on advertisements and blogs.

The cosmetics industry of Korea is loved around the world under the nickname of ‘K Beauty’, and its size is growing rapidly (KOTRA, 2017). However, cosmetics are a personal product, and the specific products that individual consumers prefer are very different. In addition, personal disposition is greatly influenced by the value of consumers' personal tastes, such as skin type and fragrance that are important to each individual (Noh *et al.*, 2017). Unlike in the past, modern consumers are actively participating in SNS activities and evaluation activities through online site reviews. The amount of accumulated consumer information is being accumulated and becoming data, which is used to design a recommendation system based on big data for cosmetics consumers (Huang *et al.*, 2018; Madden, 2012).

So, this study aims to study the recommendation system for cosmetics, a product that many consumers use daily. The purpose of this study is to encourage more consumers to recommend better products for their tastes and to increase customer satisfaction. In addition, it is expected that individual marketing for consumers will be possible in terms of enterprise, thereby increasing the repurchase rate and further securing additional customers. In addition, this study is expected to have academic and theoretical meanings for the research of consumer recommendation system based on Big Data.

2. Big Data

Big data is a vast and short generation cycle compared to the past, and includes large-scale data including characters and video data as well as simple numbers. In other words, Big data can be used as an important resource for future competitiveness.

Big data is user generated contents by information generated through social media, public data, location information based data, etc (Bae *et al*., 2012). Big data is aimed at extracting new insights by rapidly processing and analyzing a vast amount of data in terms of quantity as well as the size of data. Distributed processing solutions such as Hadoopand visualization of analys is results are performed in the IT field. The important point is how to interpretand utilize big data, and it is time to accelerate the efforts of the social economy using big data. Through analysis of raw data, which is a material of big data, new information and facts about products and services can be found. As a result, companies can manage various information and main contents about their companies in real time by combining big data analysis techniques or deliver new information to customers.

In fact, in the case of the Ritz-Carlton Hotel, we use data on 1 million customers world wide to identify customer needs and provide customized individual services. In the case of Walmart, it operates Walmartlabs apps using mobile and social shopping features at each branch, and everyone investigate the consumption patterns of individual consumers through Walmartlabs and reflect them in sales (Yoon and Kwon, 2012).

In this way, it is considered that information using big data is not concentrated in specific areas, but it can be valuable in that it is possible to draw opinions without distinction of the world class and provide customized information to each individual.

3. Hot Topic Detection

To measure the similarity between products, TF-IDF (Term Frequency – Inverse Document Frequency) algorithm is widely used (Kim and Park, 2013; Noh et al., 2017). Term frequency (TF) is a value that indicates how often a particular word appears in a document, and the higher this value, the more important it may be in a document. However, the importance of a word is reduced if it does not appear much in one document and frequently in another. Document frequency (DF), the reciprocal of this value is called the inverse document frequency (IDF).

That is, TF-IDF is a word that multiplied by TF and IDF, the higher the score, the less in other documents and the more frequently in those documents. There are a number of methods for calculating TF values. Depending on the length of the document, the frequency value of the word can be adjusted.

It is the most well-known way to quantify words in a document to calculate similarity or importance, and to use them when you want to weigh important and non-critical words differently. Simply put, numeric values that represent the

importance given to each word in the document.

If past data collection studies have invested a long time to detect major hot topics, recently, it has been possible to detect information on the title, source, and contents of hot topics more easily and quickly by using crawling (Gryman, 2010). The higher the similarity between important topics at the moment, the more parts of the topic of the document appear (https://ko.wikipedia.org/wiki/Tf-idf).Conclusion and Proposals.

Housing is one among the essential needs of person and Government has the prime duty to supply affordable housing for all income groups. During this regard, Government has got to take adequate measures for policy implication that specialize in the middle-income group in terms of budgeting, loans from provident funds, pension etc. If these provisions were made, several real estate loan opportunities will rise easing the financial requirements of housing for middle-income group. Public sector should undertake a greater number of site services scheme for exploitation and will work ashore policy and land price mechanism. For instance, in Sonadanga area most of the serviced plots are acquired by higher-middle income group. If certain number of small plots (2-3 katha sized) were developed and facilities were provided considering the middle-middle and lower-middle income group then they also could easily afford those plots. So, Government should develop land considering people of specific income group to supply affordable housing for all people. On the opposite hand, formal private sector in Khulna City isn't yet developed properly due to lack of reliance on the developers and lack of organizations among the individual owners just in case of co-operatives. Government can take initiative during this regard providing an organizational body for funding and monitoring the works of personal formal sector. By this, more people are going to be ready to afford housing by raising funds for initial deposit requirement and can also believe the standard and facilities provided by the developers.

4. Recommendation System

The recommendation system is one of the new marketing methods to help consumers who are in the midst of choice in a mass production society where new products are poured out (Han *et al.*, 2013). The recommendation system is a method of automatically recommending the most appropriate information by filtering the information that the consumer wants from a large amount of information. In other words, it is the system that can significantly reduce the effort of spending time and money directly to find the information that consumers want. Typical methods of such recommendation systems include 'Euclidian distance measurement method' and 'cosine similarity measurement method (Han *et al.*, 2013). This is a technique that checks the distance between products and recommends another user's purchase product that is most similar to the user who wants the information. In addition, there

are knowledge-based recommendation system, hybrid system recommendation system, and content-based recommendation system that are recommended according to consumers' preference (Gryman, 2010).

4.1. Similarity calculation

The recommended system scores similarities between consumers and consumers, items and items, and consumers and items, and operates considering the value. Similarity measurement methods underlying this recommended system include Euclidian, Manhattan street, Pearson correlation coefficient, cosine distance, Tanimoto distance, and most commonly, Euclidean distance, cosine distance, Pearson correlation coefficient, and Tanimoto distance measurement. Euclidian Street is the most intuitive and intuitive concept of distance as the easiest way to calculate similarities between consumers or items. In general, the distance between two points on two dimensions is calculated according to Pythagorean theorem. However, the formula for the Euclidian Street is [Form 1] for extending the Pythagoras theorem a little more to save the distance between the two points in the n-dimensional space.

$$\text{EuclideanD}(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Where (x, y) is two consecutive data, n means the number of data in the data set.

4.2. Euclidean distance measuring method

Euclidian distance measurement is a method of measuring similarity between consumers as a distance or calculating similarity between products as a distance (Gryman, 2010). In general, the distance between two points on two dimensions is calculated according to the Pythagorean theorema. However, to obtain the distance between two points in the n-dimensional space, the expansion of Pythagorean theorema little further is Euclidean. At this time (x, y) is continuous data, and n value means the number of data in the data set. However, to use the Euclidean measurement method, the data normalization process must be carried out first with a value between 0 and 1 (Han *et al.*, 2013).

However, the similarity calculation method through Euclidean distance does not confirm whether the vector has the same direction in that it calculates the simple distance between the two vectors. Therefore, if the Euclidean distance between the two vectors is the same when using the similarity, there is a limit that the similarity can be larged even if it has different direction.

4.3. Cosine similarity measurement method

There are many ways to find similarity in vectors. Cosine similarity is one of them. Unlike other things, the characteristic of cosine similarity is that it uses the angle of intersection.

Cosine similarity measurement method of utilizes the value measuring angle bet

ween two vectors in the space through the cosine mode. Cosine similarity measurem ent method is roughly intended for the following specific purposes. It's a commonly first, the search engine is used as an algorithm for basic ranking to obtain similariti es between search terms and documents and to show the most similarities. This is be cause the documents that are often compared to the documents that are most similar to those entered in the search engine's search window and those that the search engi ne has. Here, the objects for cosine similarity are the query language entered by the user and pairs with all the documents the search engine has. So you get a cosine sim ilarity, and you show the most similarity at the top. Second, used in Text Mining. B ecause search engine and text mining are very related, they are actually very relevan t to number one. Text mining often uses the vector space model (Vector Space Mod el) and the Term Frequency – Inverse Document Frequency (TF-IDF) which is ofte n used to obtain similarities between word sets. So it comes out in deep learning mo dels. In addition, use other analysis and repair models to obtain similarities. Someti mes it comes out, but it's not common. And cluster is used to tie data points together in a cluster model. Distance calculations in clustering are all related to computation al cost issues. So is the cosine similarity measurement method. The fundamental purpose of obtaining similarity is to make a relative comparison, s uch as whether A is similar to B or A and C are more similar.

And if the cosine similarity measurement method is the angle between the two v ectors 0° and the direction is completely the same, if the angle is 1 and 90° and ther e is no relation with each other, 0 and 180°, it appears as -1 when the direction betw een the two vectors is completely opposite. That is, the cosine value of two vectors of a single quadrant is between 0 and 1. In other words, the measured value has a va lue between -1 and 1, and the expression in the positive space between 0 and 1 is a ' cosine similarity measurement method', which means a similar degree between the t wo vector values. Similarity between two products is calculated only in the case wh ere the specific user gives rating for two products comparing similarity (Jiang et al., 2011).

At this time, the similarity of the product is calculated by applying the cosine si milarity formula based on the grade of the product. In other words, the collaborative filtering recommendation system based on the product can be continuously recomm ended according to the consumption history of the product, but since the similarity v alue is already calculated in advance (Ahn, 2010), it is often recommended to the sa me product.

5. Implementing a Recommendation System Based on Big Data Using R

The Big Data based cosmetic recommendation system introduced in this study makes recommendations by utilizing unstructured data such as packaging containers and feelings after use, as opposed to using only basic cosmetic selection attributes

such as price, skin type, moisture, and scent in a typical cosmetic recommendation system. This aims to provide consumers with a customized recommendation system that is most satisfied by considering not only similarities between cosmetic characteristics but also other feelings, and a big data-based cosmetic recommendation system will be implemented under the structure shown in [Fig 1].

For the purpose of this study, the entire contents of the web page of the reconciliation application were scrambled and the collection target that was intended to be stored within the web page was crawled to data.

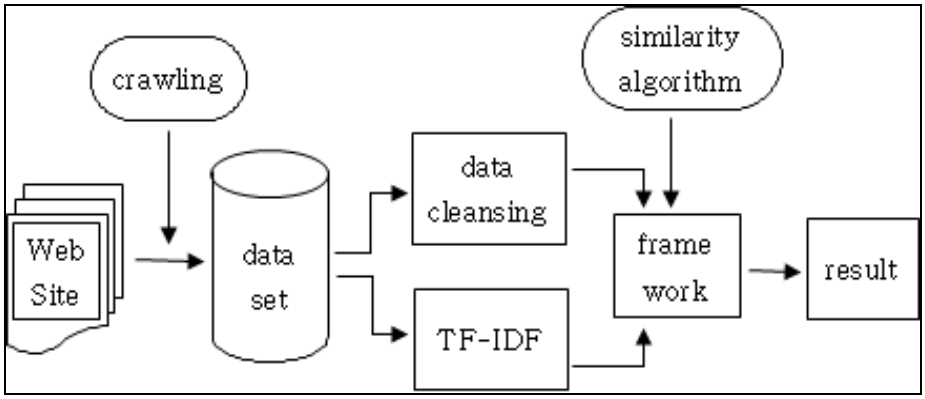

Fig. 1: Big data-based recommendation system concepts

5.1. Data collection

The crawling is a technique that reads web pages and parses them in HTML or CSS as a way to bring data from sites where no Open API is provided, and extracts the necessary data in text form.

On 'hwahae App', the operating body that operates the web does not provide Open APIs, so we used a methodology called crawling to collect information about cosmetics, and we brought cosmetics information through the production of the crawling bot [Fig 2]. The information of 140,000 cosmetics was collected using the crawling bot, and the collected information was finally entered into a variable named ‘cosmetics total’ saved in file (Table 1).

And the data collected in this study were analyzed by using text mining and hot topic-based trend analysis. In other words, the actual consumer data was used to identify individual consumer propensity. The main analysis subjects were the number of clicks on cosmetics, product rating, review of use, and time remaining in the cosmetics. For collecting and processing big data, it is intended to provide a customized recommendation system that provides practical satisfaction to consumers by considering not only the similarity of the characteristics of specific cosmetics but also the contents preferred by existing cosmetics.

```
cosmetics_total <- NULL
try (
  for (i in 157204: 162849){
   ie <- read_html(url)

   ko_name <- html_nedes ( ie, css = ' . name_ko' )
   head(ko_name)
   ko_name_1 <- str_ replace_all (ko_name [ ] %>%html_text( ), '/n', '')
   names(ko_name_2) <- c
                    .
                    .
   cosmetics_ productor<- html_nodes(ie,css = '.cosmetics_info dd')
   head(cosmetics_productor)
   cosmetics_productor_1<- str_replace_all(cosmetics_productor[1] %>% html_text( ), '/n', '')
   cosmetics_productor_2
   cosmetics_productor_2$cosmetics_productor_1
   cosmetics_productor_2$cosmetics_productor_1
   names(cosmetics_productor_2) <- c

   cosmetics_flavor <- html_nodes(ie,css = '.texture_info img')
   head(cosmetics_flavor)

   cosmetics_flavor_1 <- as. charcter(cosmetics_flavor)
   condition <- dd ( "flavor", cosmetics_flavor_1)

   if (! is.nu11(conition) && length(condition) > =1) {
     cosmetics_flavor_2 <- cosmetics_flavor_1 ["flavor", cosmetics_flavor_1)]
     if (str_detect(cosmetics_flavor_2, "img_1.jpg")){
     cosmetics_flavor_3 <- as. dafa. frame(1)
     } else if (str_detect(cosmetics_flavor_2, "img_2.jpg")){
```

Fig. 2: Crawling bot code

Table 1: Cosmetics_total data loading

library(xlsx)
cosmetic <- read.xlsx("cosmetic_total.xlsx", sheetIndex = 1)

5.2. Data Preprocessing

In this study, the 'stringsr package' and 'dplyr package' provided by R were utilized for data pretreatment. The package is one of the useful data preprocessing, but are its package. In particular, in the case of 'dplyr', the processing speed is a bit faster compared to another pre-processing package provided by R. Big data were collected through crawling and html information was removed, but it was necessary to organize the indolecs to create a recommendation algorithm. In addition, the 20 representative cosmetics for each type of representative attributes of cosmetics were composed of text formats in sentence form. To calculate the score of each type of cosmetics and which flavor has texture to make a recommendation algorithm, we went through the preprocessing process of TF-IDF (Term Frequency–Inverse Document Frequency) method used in text mining (Table 2).

Table 2: Productor data preprocessing

s <- 'A0'
cosmetic1$productor <- gsub(s, "", cosmetic1$productor)
cosmetic1$productor <- gsub("\(", "", cosmetic1$productor)
cosmetic1$productor <- str_trim(cosmetic1$productor)

[Table 2] is an instruction for removing blanks after removing Unicode characters and parentheses from the productor(producer) variables. In the case of production(productor) variables, the contents of production are contained, but the strings are not separated, so it is necessary to separate production to obtain the production countries that are necessary variables. After removing unnecessary strings, the strings were separated based on the string, which distinguishes the production country from the production area. And when a string is separated into a list format, the as.data.frame function is used to make it into a data frame format, and the rbind function and do.call function are used to bind each data frame into a row unit data frame. This was created by combining the variables called temp and the data of Cosmetics1 to create a production variable with information about the country of production of cosmetics.

5.3. Preprocessing using TF-IDF

Therefore, pretreatment was carried out using the TF-IDF method, which is used for text mining and information retrieval. TF-IDF is a method of assigning numerical values through the frequency and weight of words contained in a document to calculate similarity between documents.

Table 3: Weight_Tf-Idf function

noun_tokenizer <- function(doc)extractNoun(doc)
tdmat <- TermDocumentMatrix(docs.corp,
control = list(tokenize = noun_tokenizer,
weighting = function(x) weightTfIdf(x, TRUE),
wordLengths = c(1,Inf)))

6. Implementation of the Recommendation System Using Similarity Algorithm

Table 4: Normalization using cosine similarity and application of algorithms

library(recommenderlab)
cosine_cosmetic <- as(cosine_ cosmetic, "realRatingMatrix")
cosine_ cosmetic <- similarity(cosine_ cosmetic, method = 'cosine')

Table 5: Recommended cosmetic output algorithm for similarity

<- c("LANCOME EYECREAM", "CHANEL EYECREAM”)
qwe = function(x){
wine_N <- as.data.frame(x)
colnames(cosmetic _N) <- " cosmetic _name"
cosmetic _N$ cosmetic _name <- as.character(cosmetic _N$ cosmetic _name)
cosmetic_B <- c()
for(i in 1:length(cosmetic _N$ cosmetic_name)){
cosmetic_A <- head(filter(cosine_ cosmetic3,
variable =cosmetic_N$cosmetic_name[i]) %>%
select(names, variable, value) %>%
arrange(variable, desc(value)),5)
cosmetic_B <- cbind(cosmetic_B, cosmetic_A)}}

Based on the data on the selectivity of cosmetics data, the company wanted to a system implement m that recommends another cosmetics product that is most similar to th e ones entered by consumers. Various methods exist for measuring similarity, such as Euclidan, Manhattan and Cosine, but the similarity is more suitable for measuring the similarity in the properties of cosmetics made up of different dimensions (attributes) in that it is possible to find similar characteristics than Euclidean, which calculates distance in consideration of only the quantitative values between each item. To apply the cosine similarity matrix, 'recommenderlab package' was utilized and the corresponding data was converted into realRating Matrix form to apply cosine similarity algorithm [Table 4]. In order to apply the algorithms of the package, the 'cosine similarity algorithm' was applied by converting the data into real RatingMatrix form. Based on the results of cosine similarity, data of cosmetics prepared by consumers were entered, and an algorithm

[Table 5] was finally implemented to print out five different cosmetics with similar selection attributes for cosmetics.

7. Conclusion

Due to the development of internet technologies and various content, modern consumers can get information about their products online at any time regardless of time and place. In addition, information retrieval and purchase are simultaneously made. However, too much information can overwhelm consumers and hinder rational purchases (Gryman, 2010). The main background of this study is to develop a cosmetic recommendation system of similarity algorithms using big data, which can reduce the purchase time of cosmetics and make satisfactory purchases by recommending the best cosmetics to consumers with various conditions and tastes among various cosmetics sold in the market.

Although there are various types of recommendation algorithms, this paper deals with recommendation systems using item-based similarity algorithms.
The system uses a cosine similarity algorithm to calculate the value of the optional attribute that characterizes each cosmetic product, measuring the degree to which it most closely resembles the consumer's desired condition, and recommends the top five products with high values. In addition, the properties of various cosmetic information consisting of unstructured data were processed by the TF-IDF method, which measures the similarity through the frequency of appearance of words between documents and similar cosmetics were considered. This suggests that a recommendation system that considers the overall selection attributes of cosmetics can respect consumers' tastes more than a recommendation system that focuses on one category of various cosmetic characteristics. However, in the case of cosmetics may vary depending on skin type, income, gender, event gifts or consumer use. And it is necessary to take into account the individualized characteristics of the consumer, in that the tendency to choose according to the individualized characteristics of the consumer can change.

Therefore, in the future, based on the recommendation system developed above, studies on cosmetic recommendation algorithms considering the personalization characteristics of consumers should be carried out and based on this, it is expected that advanced studies will be made on the overall recommendation algorithms considering various consumer characteristics.

References

Ahn, Y. (2010). Link communication reveal multiscale complexity in network, *Nature*, 8, 761-764.

Bae, S.H. Kim, D.H. Kwock, I.H. and Song, Y.K. (2012) Business Future Map Created by Big Data, *Hans Media*, 1-272.

Gryman, G. (2010). Tapping into power of big data. *Technology Forecase*, 3, 4-13.

Han, E. J. Lee, S. G and Kim, T. G. (2013). Wine market segmentation and its determinants by wine buying motivations. *International Journal of Tourism and Hospitality Research,* 27(1), 67-79.

Huang, S. Chen, Y. Chen, H. Chen, L. and Fan, Y. (2018). Personalized item of interest recommendation on storage constrained smartphone based on word embedding quantization. *Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining*, 610-621.

Jiang, Y. Liu, J. Tang, M. and Liu, X. (2011). An effective web service recommendation method based on personalized collaborative filtering. *Proceedings of International Conference on In Web Services*, 211-218.

Jung, J (2017). Key to the era of the 4th industrial revolution, software quality. *Issue Report of NIPA*, 7.

Kim, Y. A. and Park, G. W. (2013). Topic-driven social fank: Personalized search result ranking by identifying similar, credible users in a social network. *International Journal of Knowledge-Based Systems*, 54, 230-242.

KOTRA. (2017). 4th industrial revolution, manufacturing innovation and smart factory. *Oversea Market News*.

Madden, S. (2012). Databases to big data. *Internet computing –Ieee computer society*, 16(3), 4-12.

Noh, Y. Lim, J. Bok, K and Yoo, J. (2017). Hot topic prediction scheme using modified TF-IDF in social network environments. *Journal of KIISE Transactions on Computing Practices*, 23(4), 217-225.

Yoon, M.Y. and Kwon,J. E. (2012). World Evolving into Big Data, *National Information Society Agency*, 1-160.

Yoon, J. Y. and Joung, S. H. (2019). A consumer recommendation system based on big data. *International Journal of Smart Business and Technology*, 7(2), 25-30.

Yun, S. and Youn, S. (2017). Recommendation system using big data processing technique. *Journal of the Korea Institute of Information and Communication Engineering*, 21(6), 1183-1190. https://ko.wikipedia.org/wiki/TF-IDF。