Implementation of a Personalized Beauty Web Magazine using Mind Mining

Yujin Song, Yongjang Cho, and Seungmin Rho

Abstract—The number of websites and applications that provide beauty data to consumers is drastically increasing. Consumers usually select and purchase products based on product information and customer reviews on the Internet, as well as listening to the opinions of friends. Consumers, however, often find it difficult to find the perfect product because there is a vast amount of data available online, which makes the process of selecting products more difficult. To address this problem, we have analyzed vast amounts of big data provided by both consumers and the beauty industry using mind mining, data mining, text mining, and natural language techniques to identify preferences. We suggest a way of implementing a personalized beauty web magazine by identifying preferences from analyzed data.

Research Keywords—Opinion Mining, Semantic Web, Natural Language Process, Structured Data, Beauty

1 INTRODUCTION

The beauty industry is becoming increasingly active in providing cosmetic products based upon personal preferences. There has recently been a drastic increase in the amount of cosmetic reviews and related data being used by applications and websites to deliver beauty information to consumers. Consumers can easily obtain beauty-related information through the Internet using a smartphone; however, it takes considerable time for consumers to find what they want because of the vast amount of intangible data that is widely distributed. There are now two ways to search for a desired product on a beauty site. One can search a portal site, which usually results in many irrelevant hits, or one can go to a beauty website and read the reviews and ratings of selected products. Both of these methods involve scrolling through unnecessary information and performing repetitive tasks, causing consumers to waste time and making product selection more difficult. To address this problem, we recommend developing a website that enables consumers to locate products more efficiently.

In this paper, we recommend a method of developing a practical, personalized beauty web magazine using data mining, text mining, and dynamic web technology.

- *Yujin Song is with the Media Software Department, Sungkyul University, Anyang-si, Korea. E-mail: uzini_@naver.com.*
- *Yongjang Cho is with the Media Software Department, Sungkyul University, Anyang-si, Korea. E-mail: dydwkd48670@gmail.com.*
- *Seungmin Rho (corresponding author) is with Media Software Department, Sungkyul University, Anyang-si, Korea. E-mail:smrho@sungkyul.ac.kr.*

2 ANALYSIS OF MINED DATA

2.1 HTML Parsing Using Jsoup

Jsoup is an open-source Java library of methods that extracts and manipulates data stored in Hypertext Markup Language (HTML) on websites. Jsoup parses [1] the HTML and outputs the HTML in the form of a URL, file, or string. In this study, we used Jsoup to parse HTML and collect data wherever the Open API was not supported.

2.2 Data Processing Using Solr

Solr is an open source enterprise search platform provided by Apache that enables real-time indexing and faceted searching [2]. After completing a variety of settings, it generates a file as a result. It returns the result of the indexed file and applies it to the web using data from the resulting data.

3 BEAUTY DATA MINING SIMULATION

3.1 Beauty Data Site Selection and Parsing

Table 1 indicates the amounts of data that may be collected based on data type and the amount of

data available at a site. Crawling then takes place, and parsing of the HTML data from the selected beauty site using Jsoup [1]. Product names, skin types, ages, ratings, reviews, and image sources are gathered and saved in a comma separated values (CSV) file. Table 2 represents a portion of the data we collected and stored in a MySQL database.

Table 1. Data site research

Site name	Amount of Data	Collectability
BeautyNet	1.2 million	O
AprilSkin	1,080	X
LohaCell	14,500	O
TheFaceShop	14,000	X

Table 2. Parsed beauty data list

Product	Type	Age	Grade	Contents	Image
LibBalm	Oily	20	4	저렴하다	Src
Remover	Normal	30	5	좋다	Src
Eyecream	Dry	40	4	가볍다	src

3.2 Korean Data Processing

Korean uses a separate setup procedure when using analyzers such as Solr Lucene because Korean differs from other types of text. To process Korean, we had to employ morphological removal, margin treatment, and synonym processing by using the Arirang open API [3].

3.3 Indexing and Searching

Once the system was configured for Korean data processing, we indexed the collected beauty data and created a result file. We only searched for and retrieved indexed data suitable for use in consumer queries. The search was conducted using the Solr query syntax and a standard syntax analyzer [2]. After the query results were analyzed, they were stored in an XML-formatted document. Information was then extracted from the XML document and displayed on a web page.

Table 3 is an example of an analysis of Korean beauty data based on Arirang and Solr. Complete refinement of the data was not possible, but in our initial investigation, we only needed to parse the text into separate sentences, phrases, and words.

Table 3. Split text

Field Value(index)	Text
여름에 포인트로 바르기 좋을 것 같아요	여름, 포인트, 바르기, 좋을, 같아요
피부에 쫀쫀하게 붙는 느낌 이예요	피부, 쫀쫀하게, 붙는, 느낌, 이예요
트러블 없고 순한 느낌 이예요	트러블, 없고, 순한, 느낌, 이예요

4 CONCLUSION

In this study, we examined the use of data mining in building a personalized web page. Figure 1 illustrates the system architecture used for our study. It also shows the process we used to p ersonalize data through visualization, web-crawling, data mining, and the analysis of unstructured data obtained from the beauty industry. Figures 2 and 3 are screenshots of the webpages created using our method.

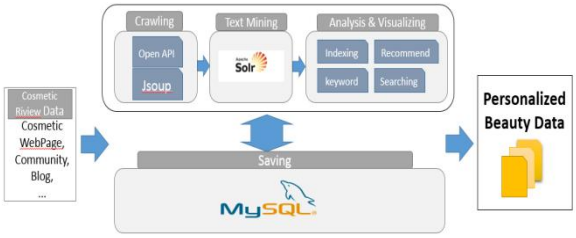

Fig. 1. System Architecture

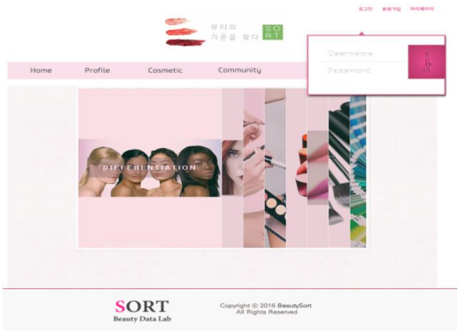

Fig. 2. Screenshot of the main page

Fig. 3. Cosmetic search page

According to the Samsung Economic Research Institute (SERI), sales in the online beauty industry in Korea have grown at a rate of 10.1% annually in Korea, rising to 11.9 trillion won. The Korean beauty industry is a high value-added industry; therefore, it is necessary to use state-of-the-art technology to provide better service to customers.

Future research will involve the development of methods for refining text, for example, elimination of morpheme and synonym processing, to enhance the value of the information provided to customers on the web.

ACKNOWLEDGMENT

This work was supported by the Korea Foundation for the Advancement of Science & Creativity(KOFAC), and funded by the Korean Government(MOE).

REFERENCES

[1] H. Schildt, Art of JAVA: Parser using recursive usage. Trans. M.Bongjae. Seoul:Information Culture History , 2004. April.
[2] R. Kuc, Building and Managing Apache Solar 4: Data Indexing. Trans. P.Jaeho. Seoul:Akon Publishing, 2014. April.
[3] Sumyung. "Internet: cafe.naver.com/korlucene", 2008. Oct.

Yujin Song is pursuing the B.S. degree in Multimedia Engineering in Sungkyul University, Korea. Her research interests are big data analysis and web application.

Yongjang Cho is pursuing the B.S. degree in Multimedia Engineering in Sungkyul University, Korea. His research interests are big data analysis and VR/AR games.

Seungmin Rho received his M.S. and Ph. D Degrees in Computer Science from Ajou University, Korea in 2003 and 2008, respectively. He visited Multimedia Systems and Networking Lab. in Univ. of Texas at Dallas from Dec. 2003 to March 2004. Before he joined the Computer Sciences Department of Ajou University, he spent two years in industry. In 2008–2009, he was a Postdoctoral Research Fellow at the Computer Music Lab of the School of Computer Science in Carnegie Mellon University. He had been working as a Research Professor at School of Electrical Engineering in Korea University during 2009–2011. In 2012, he was an assistant professor at Division of Information and Communication in Baekseok University. Now he is currently an assistant professor at Department of Media Software at Sungkyul University. His current research interests include database, big data analysis, music retrieval, multimedia systems, machine learning, knowledge management as well as computational intelligence. He has published 100 papers in refereed journals and conference proceedings in these areas. He has been involved in more than 20 conferences and workshops as various chairs and more than 30 conferences/workshops as a program committee member. He has edited a number of international journal special issues as a guest editor, such as Multimedia Systems, Information Fusion, Engineering Applications of Artificial Intelligence, New Review of Hypermedia and Multimedia, Multimedia Tools and Applications, Personal and Ubiquitous Computing, Telecommunication Systems, Ad Hoc & Sensor Wireless Networks and etc. He has received a few awards including Who's Who in America, Who's Who in Science and Engineering, and Who's Who in the World in 2007 and 2008, respectively.