Building a Restaurant Menu Presentation Database and Visualization

Yeong Hyeon Gu, Seong Joon Yoo*, Dongil Han, Sung Wook Baek, Byung-Joo Shin, and Yun Hwan Kim

***Abstract*—**For restaurants' successful advancement to local markets and settlement, it is important to develop localized menus according to food culture region and menu and to keep refining the menus. However, there are difficulties in collecting data for developing localized menus according to food culture region and menu and to improve the menus. In order to resolve such difficulties, this study built a menu presentation DB according to food culture region and menu. For this, we collected and classified menu images using a wrapper-based Web crawler, extracted and matched menu name by restaurant, classified menu images according to food culture, region, and menu item, built a menu presentation DB, and visualized the results in order for users to analyze easily. Through a user experiment, the implemented system was found to be able to search accurately and 5-12 times faster and, therefore, to solve problems of interest successfully

***Index Terms*—**menu presentation, restaurant database, similar image search. visualization

I. Introduction

Since the late 2000s, the globalization of Korean food has been promoted in diverse ways by the government, academic circles, and businesses, but around 20 of domestic food service companies that advanced to overseas markets in December 2008 are believed to have closed or withdrawn their businesses. According to a survey, moreover, only about 5% of Korean restaurants operating in foreign countries are making profits, and the other 95% are doing business poorly or have experienced failure [1, 2].

With active efforts to open overseas markets by domestic food service companies, some visible achievements have been made in the globalization of Korean food, but there are still obstacles to the successful advancement of Korean menus to foreign markets due to differences in food culture among countries and regions. Factors contributing to failure in overseas advancement include lack of preliminary surveys, difficulty in developing localized menus and supplying food materials, etc. Although it is required to survey local food culture and menus continuously, systematic preliminary surveys of overseas markets are not being conducted because they cost a lot of money, labor, and time.

Accordingly, we need a new system for surveying menus in consideration of local food culture in foreign countries. Such a system is expected to enable the development of menus in harmony with local food culture. The system needs to include a menu presentation DB according to food culture region and menu and provide it to food service companies. In addition, big data may be used to survey and analyze restaurant menus made by local customers and to grasp local consumers' preferences and menu-related trends.

For Korean restaurants' successful advancement to local markets and settlement, it is crucial to develop and refine localized menus according to food culture region and menu, but faced with difficulties in localization, they have so far developed menus only after a market survey for target areas or used menus in Korea without any change.

In order to solve these problems, this study built a menu presentation DB according to food culture region and menu, which is usable in the development and refinement of localized menus. For this, we collected and classified menu images using a wrapper-based Web crawler, extracted and matched menu name by restaurant, classified menu images according to food culture, region, and menu item, built a menu presentation DB, and visualized the results in order for users to analyze easily.

This paper is structured as follows. First, Chapter 2 analyzed related works, and Chapter 3 described restaurant review collection techniques. Chapter 4 introduced the menu image collection technology, Chapter 5 explained similarity-based menu image search using collected data, and lastly Chapter 6 drew conclusions.

II. Related Works

An ordinary Web crawler automatically circulates through Web servers, analyzes the contents of web pages, and extracts URLs included in the contents. Then, it navigates the URLs one by one and collects Web documents. A large volume of Web documents collected in this way is used in a search engine.

On the other hand, some crawlers collect documents of a specific theme. Focused Crawler [3] and Topical Crawler [4, 5, 6, 7] are mounted with a document classifier or rules about documents to be collected in themselves. If the classifier's performance is low and the rules are not comprehensive, however, data desired by the users cannot be collected.

For this reason, research has been made on wrappers that can analyze the structure of a specific site and collect data directly from the site. A crawler using wrappers creates rules for extracting necessary information, and collects contents from a site automatically. However, information collected in

Manuscript received December 5, 2014.

Yeong Hyeon Gu is with the Department of computer engineering, Sejong University, Seoul, 143-747, Korea (e-mail: yhgu@sju.ac.kr).

Seong Joon Yoo is with the Department of computer engineering, Sejong University, Seoul, 143-747, Korea (e-mail: sjyoo@sejong.ac.kr). *Corresponding author.

Dongil Han is with the Department of computer engineering, Sejong University, Seoul, 143-747, Korea (e-mail: dihan@sejong.ac.kr).

Sung Wook Baek is with the Department of digital contents, Sejong University, Seoul, 143-747, Korea (e-mail: sbaik@sejong.ac.kr).

Byung-Joo Shin is with the Department of computer engineering, Sejong University, Seoul, 143-747, Korea (e-mail: shinbyungjoo@gmail.com).

Hwan Kim is with the Department of computer engineering, Sejong University, Seoul, 143-747, Korea (e-mail: kimyh@sju.ac.kr).

http://dx.doi.org/10.15242/IAE.IAE1214012

this way contains a lot of unnecessary data and this impairs the accuracy of information. In this study, thus, the information is classified once more using rules.

Ordinary image search engines such as Google[8] index documents containing images, so the results of image search include many images irrelevant to the theme. In addition, accurate analysis is difficult because data about restaurants, countries, or areas are insufficient or inaccurate. What is more, it is impossible to find similar menu presentations in a specific area.

III. Collection Of Restaurant Reviews

In order to analyze menu presentations, we need to build a DB on restaurants (business name, telephone number, address, business hour, closing days, etc.). The basic DB was built with information collected as in Table 2 from overseas restaurant portal sites (hereinafter 'portal sites') as in Table 1.

TABLE I: Countries and Sites for Data Collection

Region	Country	Site Address
America	U.S	www.yelp.com [9]
	Canada	www.yelp.com
Europe	France	www.yelp.com
	Spain	www.yelp.com
Asia	Japan	www.tablelog.com [10]
	Singapore	www.openrice.com [11]
	Hong Kong	www.openrice.com

TABLE II: The Information of Portal Sites

No	Information	No	Information
1	ID	17	Good for Kids
2	Name	18	Take-out
3	Rating	19	Smoking
4	Review#	20	Coat Check
5	Category1	21	Outdoor Seating
6	Category2	22	Noise Level
7	Category3	23	Noise Level
8	Address	24	Good For Dancing
9	Neighborhoods 1	25	Best Nights
10	Neighborhoods 2	26	Ambience
11	Neighborhoods 3	27	Good for Groups
12	Price Range	28	Happy Hour
13	Business Info	29	Has TV
14	Nearest Transit Station	30	Accepts
15	Alcohol	31	Credit Cards
16	Hours	32	Parking

In order to collect bibliographic information provided by the portal sites, this study used a wrapper-based Web crawler. The purpose of using the wrapper-based Web crawler was accurate and fast collection of bibliographic information from each portal site. Here, a wrapper means a rule or program for extracting only desired information from an information source. An ordinary Web crawler traces URL links and collects all the contents of HTML documents, but a wrapper-based Web crawler extracts only specific types of data.

For each site, the structure of web pages expressing bibliographic information to be extracted is analyzed and a wrapper DB is constructed. Then, the DB is applied to the wrapper-based Web crawler and data are collected. Figure 1 shows the result of analyzing web page structure common to portal sites.

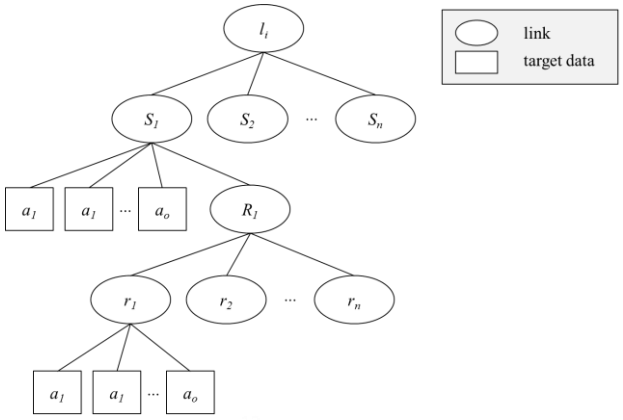

Fig. 1. The structure of restaurant portal web pages (based on information to be extracted)

l_i is an element of set L, and means an URL for accessing the restaurant list page of each portal site.

$$L = \{l_1, l_2, ..., l_m\}$$

s_j is an element of set S, and means an URL for accessing the page of details on each restaurant. The number of restaurants varies among the sites.

$$S = \{s_1, s_2, ..., s_n\}$$

a_k is an element of set A, and information to be extracted ultimately. The information includes 'Name,' 'Address,' 'Hours,' etc.

$$A = \{a_1, a_2, ..., a_o\}$$

The materialized data of set A can be expressed as follows.

$$l_1 \cdot s_1 \cdot A = \{\text{'}Gary\ Danko\text{'}, \text{'(415) 749-2060'}, \text{'}CA\text{'}, ...\}$$

R_j is another set included in s_j, and means a set of reviews posted by users. The set consists of p reviews.

$$R_j = \{r_{j1}, r_{j2}, ..., r_{jp}\}$$

Review r_{jp} contains information to be extracted about the review. The information to be extracted is defined as set A', and information included in the set is 'Contents of Review,' 'User,' Date of Registration,' etc.

$$A' = \{a'_1, a'_2, ..., a'_q\}$$

The materialized data of set A' are as follows.

$$R_1 \cdot r_{11} \cdot A = \{\text{'}So\ good...\text{'}, \text{'}John\text{'}, \text{'}2013\text{-}05\text{-}24\text{'}, ...\}$$

A wrapper DB should be built based on the web page structure analyzed above. The wrapper DB has data about the starting position and end position of bibliographic

information in order to extract bibliographic information. l_i is a preset URL.

The DB should have position information not only for data to be extracted expressed in rectangles but also those expressed in ovals. It is because although information to be extracted ultimately are the rectangular nodes, their corresponding URLs are necessary in order to access the information. Accordingly, the wrapper stores position information in the structure as follows.

{Node name: starting position value^n : end position value^n}^{n}

Here the value of the starting position and that of the end position can be a string or the combination of multiple strings, and if it is the combination of multiple strings, there should be delimiters that distinguish the strings.

Figure 2 expresses s_j in HTML (Hyper Text Markup Language). That is, a web page to be analyzed for building the wrapper DB is expressed in HTML.

```
<table>
  <tr>
    <td rowspan=4>
      <img src=./images/noimage.jpg>
    </td>
  </tr>
  <tr>
    <td> Name: </td>
    <td> Maison Kay Restaurant</td>
  </tr>
  <tr>
    <td> Phone No: </td>
    <td> 123-123-1234 </td>
  </tr>
  <tr>
    <td> Address>
    <td> 613 Kennedy Ave Toronto, ON M6P Neighborhood </td>
  </tr>
</table>
```

Fig. 2. An example of s_j page in HTML

The following is an example of wrapper DB built based on Figure 2.

{$s{\cdot}a_1$: *Name* ; <*td*> : *Name* ; </*td*> ; </*td*>}
{$s{\cdot}a_2$: *Phone No* ; <*td*> : *Phone NO* ; </*td*> ; </*td*>}
{$s{\cdot}a_3$: *Address* ; <*td*> : *Address* ; </*td*> ; </*td*>}

Here, a1 has the values of the starting position and the end position for restaurant Name. In order to find string 'Maison Kay Restaurant,' the system first finds string 'Name,' and then if string <td> is found, the starting position of business name can be found. For the end position, in addition, <td> should be found twice after 'Name' is found.

Figure 3 is the system structure for building the basic DB. The wrapper-based Web crawler loads data on the structure and position of information to be collected from each portal sight through the wrapper DB. Based on the loaded wrapper information, bibliographic information is fetched from each portal site.

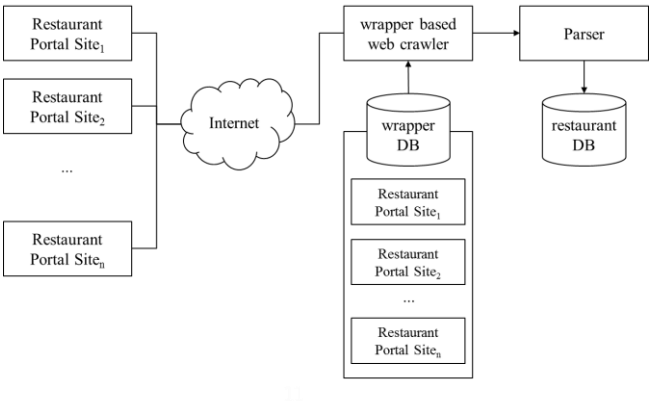

Fig. 3. System structure of the basic DB

IV. Collection Of Menu Images

As in the collection of restaurant reviews, a wrapper-based Web crawler is used also in the collection of restaurant images. First, site structure is analyzed as preprocessing to make a wrapper-based Web crawler. What to be collected are restaurant images at site 'YELP,' and according to the results of analysis, the site has a structure as in Figure 4.

```
<div class="photo-box pb-120s">
<a href="/biz_photos/gang-nam-tofu-los-angeles?filter_by_
userid
=True&amp;
select=6vSW1O1zA-W1jfa8tbWsbzg&amp;userid=aeeBwwlg
1WY
DCYPmJfoUw">
<img alt=""
class="photo-box-img"
height="120" src="http://s3-media4.ak.yelpcdn.com/bphoto/6vS
W1O1zA-1jfa8tbWsbzg/120s.jpg"
width="120">
</a>
</div>
```

Fig. 4. An example of image expression structure in a restaurant review portal site

'IMG' is a HTML tag for showing an image on a web page, and 'SRC' is an attribute for designating the URL of the image. That is, using the contents of tag 'SRC,' the system can collect images through the paths where the images are stored. This process was implemented in JAVA, and the address and tag information of images was stored in the DB. A total of 34,447 image s were collected, and other information was also collected using wrapper-based structure.

V. Menu Image Search

Using LIRE (Lucene Image Retrieval), an image search library of Lucene, similar images are searched. The method of searching similar images is as follows. First, for photographs stored in a Web server, the characteristics of the images are extracted using CEDD [12] (Color and Edge Directivity Descriptor), one of LIRE libraries, and the data are indexed and saved as a file. CEDD consists of processes extracting a color histogram in the HSV color space from an image block and calculating the color unit, and also extracting luminosity in the YIQ color space and calculating the texture unit, and then quantizing. As in Figure 6, when image Pi is clicked, indexed files are searched and image P_{sim} defined as follows in order of similarity is output.

$$P_{sim} = \{\{P_j, sim_1\}, \ldots, \{P_{j+n-1}, sim_n\}\}$$

Here, P_j means the path of a similar image and *sim* means a similarity score calculated in advance using the LIRE library. Table 3 is a pseudo code of the similar image search process.

TABLE III: PSEUDO CODE OF SIMILAR IMAGE SEARCH

```
INPUT:
  L = LIRE indexed files
  P = A Picture to search by similarity

OUTPUT:
  O = A set of similar picture with P
  S = A set of similarity of O

Initialize:
  O ← {}
  S ← {}

simImageSearch(L, P, O, S)
1       arrTmp ←{}
2       arrTmp ← LireSearch(L, P)
3       for result in arrTmp:
4               O ← Get picture path in result
5               S ← Get Similarity in result
6               Visualize(O + S)
```

Because the same image as the selected one is also indexed as in Figure 2, the final output is n-1 images excluding the image whose similarity is 1. For example, if image {"/images/20140615_105311.jpg"} is selected, its similar images and similarity scores like {{"/images/20140615_105315.jpg"}, {0.8779756}, …} are fetched and visualized.

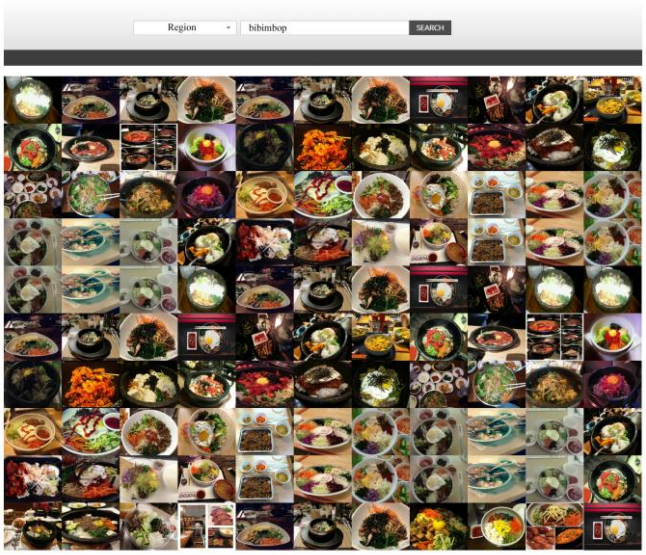

Fig. 5. A screen showing the result of tag-based menu image search

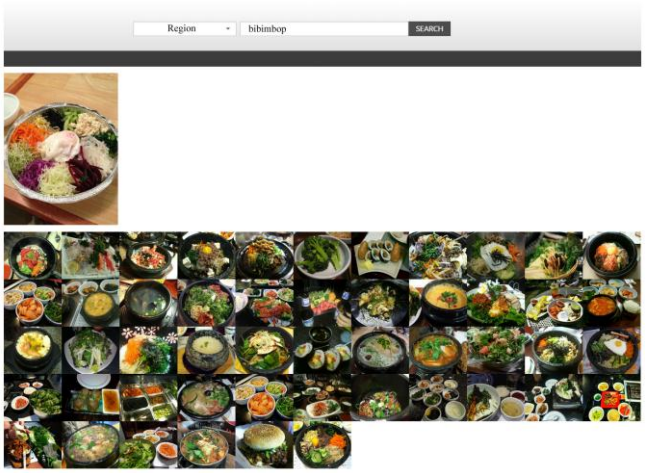

Fig. 6. A screening showing the result of similar menu image search

VI. EXPERIMENTS

In order to test how fast and accurately desired contents are found using the restaurant menu presentation image index and search system developed in this study, an experiment was conducted as follows. A questionnaire as in Table 4 was distributed to 5 users and the time to get answers to the questions with and without using the system was measured. The questionnaire also included questions about the users' opinions about the system.

TABLE IV: QUESTIONNAIRE

No	Question	Time
1	Find 100 bibimbap menu presentation images in LA..	
2	Check if the images found in Step 1 are correct	
3	Find 10 images using a white dish among the bibimbap images in LA collected in Step 1..	

The experiment first measured the time taken by each user, after reading the questions, from starting search until getting the answers using conventional search engines and without using the system. Then with using the system developed in this study, it measured the time taken from starting search by accessing the websites visualized in this study until getting the answers.

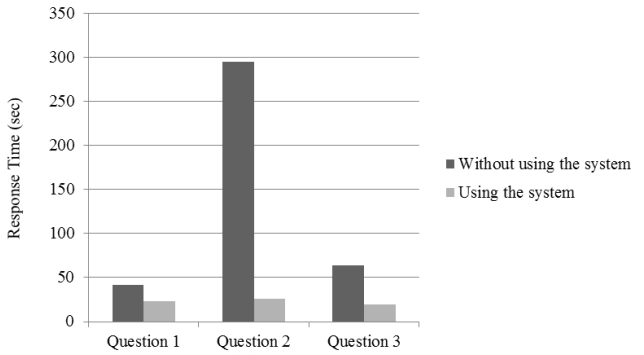

Fig. 7. Average response time

The reason for making questions as in Table 4 is to compare accuracy and speed between when using the system developed in this study and when not using it. According to the results of the experiment, accuracy was high both with and without using the system. As to speed, however, as

compared as in Figure 7, the average response time for each question was about 5-12 times longer without using the system. This proved that using the system, we can search restaurant menu presentation images in a specific area accurately and faster.

VII. CONCLUSION

This study built a menu presentation DB according to food culture region and menu in order to resolve difficulty in developing and refining localized menus according to food culture region and menu. For this, we collected and classified menu images using a wrapper-based Web crawler, extracted and matched menu name by restaurant, classified menu images according to food culture, region, and menu item, built a menu presentation DB, and visualized the results in order for users to analyze easily.

According to the results of a user experiment with the implemented system, search using the system was accurate and 5-12 times faster, and therefore the system was found to solve successfully the problems that this study intended to resolve.

Using the restaurant menu presentation search system developed in this study, food service companies planning overseas advancement may be able to design menu presentation adequately for the corresponding food culture region and menus and this will contribute to their effective settlement in the local market. Future research will collect and analyze competing menus in addition to Korean menus, and also plan to develop convenient methods for surveying and analyzing local customers' reviews of restaurant menus.

REFERENCES

[1] Korean Food Foundation, "Korean food market research", 2010.
[2] Korea Franchise Association, "Analysis of franchise failure factors", 2009.
[3] C. Bertoli, V.Crescenzi, and P. Merialdo, "Crawling programs for wrapper-based applications," *IEEE International Conference on Information Reuse and Integration*, pp. 160-165, 2008.
[4] S. Soderland. C. Cardie, and R. Mooney, "Learning information extraction rules for semi-structured and free text," *Machine Learning*, vol. 34, no. 1-3, pp. 233-272, 1999.
[5] S. Chakrabarti, M. V. D. Berg, and B. Dom, "Focused Crawling: A new approach to topic-specific web resource discovery," *Computer Networks.*, vol. 31, no. 11-16, pp. 1623-1640, 1999.
[6] C. Chang, M. Kayed, M. R. Girgis, and K. Shaalan, "A survey of web information extraction systems," *IEEE Transaction on Knowledge and Data Engineering*, vol. 18, No. 10, pp. 1411-1428, 2006.
[7] J. Yang, T. Kim, and J. Choi, "An interface agent for wrapper-based information extraction," *In proceedings of Pacific Rim International Workshop on Multi-Agents*, pp. 291-302, 2004
[8] https://www.google.co.kr/
[9] https://www.yelp.com
[10] https://www.tablelog.com
[11] https://www.openrice.com
[12] S. A. Chatzichristofis, and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval", ICVS'08 Proceedings of the 6th international conference4 on Computer vision system, pp. 312-322, 2008