*Proceedings of Engineering and Technology Innovation, vol. 7, 2017, pp. 20 - 24*

# Preliminary Study on a System for Visualization of Big Data in SMEs

Yasuo Uchida[1,*], Miao Xinyun[1], Seigo Matsuno[1], Yasushi Iha[2], Makoto Sakamoto[3]

[1]Department of Business Administration, National Institute of Technology, Ube College, Ube, Japan

[2]Department of Media Information Engineering, National Institute of Technology, Okinawa College, Nago, Japan

[3]Department of Computer Science and Systems Engineering, University of Miyazaki, Miyazaki, Japan

Received 19 July 2017; received in revised form 30 July 2017; accepted 03 August 2017

## Abstract

The 2012 White Paper on Information and Communications in Japan issued by the Ministry of Internal Affairs and Communications of Japan advocates use of big data under its "Special Theme: ICT-induced and accelerated Disaster Recovery and Japan's Re-birth." However, the analysis in the Japan Users Association of Information Systems' white paper on its 2014 IT trend survey for companies reports that less than 10% of companies utilize big data, and it would appear that progress in its use is centered on large firms. Under such conditions, use of big data is becoming a challenge for the purpose of ensuring the survival and success of SMEs as well. As a result, R&D and technological support for SMEs are becoming pressing issues. However, at present there has been almost no academic research concerning policies and future directions for use of big data at SMEs. Accordingly, this study conducted the modelization of the procedure for visualization of big data for SMEs. Specifically, we organized the procedure as a tutorial, from obtaining data of Japanese hot-spring areas using web scraping, to visualizing them using the visualization software Cytoscape

**Keywords:** big data, visualization, SMEs, Cytoscape

## 1. Introduction

This study is intended to research and develop a system for visualization of big data suited to SMEs, as a tactical information tool to support SMEs' strategies for success under conditions of increasingly intense global competition. That is, it aims to probe a framework that is easy to adopt and superior in terms of operability for the collection, storage, analysis, and use of big data. At the same time, it also aims to elucidate empirically the ideal form of a strategic information infrastructure for SMEs and challenges in its operation and administration.

In this study, we carried out a preparatory consideration of visualization of big data by SMEs. Specifically, we organized the procedure as a tutorial, from obtaining data of Japanese hot-spring areas using web scraping, to visualizing them using the visualization software Cytoscape.

## 2. Trends in Use of Big Data at SMEs in Japan

At present, there are very few examples of successful use of big data by Japanese SMEs. In addition, how big data is used at SMEs depends on individual planning by each company. Accordingly, this paper will begin by summarizing measures taken and research trends related to the use of big data at Japanese SMEs. It also will examine a number of examples of early adopters.

* Corresponding author. E-mail address: uchida@ube-k.ac.jp

Tel.: +81-836-35-7567; Fax: +81-836-35-7567

For example, the report "Enriched Living and Economy from Connected IT: The Value and Reliability of Big Data"[1] from the Research Group on IT Infrastructure for Living and the Economy of the Information-technology Promotion Agency, Japan (IPA) (IT Infrastructure for Living and the Economy of the Information-technology Promotion Agency, Japan) both explains in simple terms what big data means for managers of companies aiming to provide new services using big data and identifies results such as expansion of business opportunities through summarizing examples of early adopters of big data, advantages and issues in service realization, and efforts to resolve these.

In addition, the 2014 White Paper on Small and Medium Enterprises in Japan [2] from the Small and Medium Enterprise Agency mentions use of data on corporate transactions (big data) as a "key" to revitalization of regional economies.

Looking at the activities of SMEs in the field, in November 2014 the Osaka Chamber of Commerce and Industry published the results of a survey intended to ascertain matters such as expectations, needs, and issues involved in use of big data by second-tier companies and SMEs [3]. While the results of this survey show that approximately 81% of companies are interested in "information (data)" as "useful for management purposes," respondents also identify the following as the top three "issues in use" of data:

- "Difficulty of understanding the cost-effectiveness of use of information (data)" (64.9%)
- "Lack of human resources to analyze information (data)" (56.9%)
- "Lack of understanding of methods of using information (data)" (34.0%)

Accordingly, we decided to proceed with research and development focusing on these three points. First of all, we identified as a necessary condition the ability to use personal computers having specifications like those used in ordinary administrative-level operations instead of high-priced computers, to keep costs down as much as possible. We also decided to use, in principle, software such as open-source software that can be used free of charge as tools needed for analysis and visualization. Another prerequisite we identified was that the data analysis must be of a degree capable of being conducted by employees who have the skill levels needed to analyze data using spreadsheet software (such as Microsoft Excel), since it is difficult for SMEs to secure staff that have specialized data analysis skills. Furthermore, we decided to provide hints on use of data by describing specific examples of methods of their use.

## 3. Visualization of Big Data

### *3.1. Steps from data collection through visualization*

The data subject to visualization can be broken down into two main categories. The first consists of data in the possession of the company itself. In this case, the company has ascertained the content of the data sufficiently and it is easy for it to process the data on its own. The other category consists of data that is present on the Internet. In this case, it is difficult to understand the structure of the data and they are not easy to obtain. However, sometimes SMEs will want to obtain and utilize these data. Accordingly, this study will consider the steps used when obtaining and processing data present on the Internet. Since the main objective of this study is to illustrate a data processing model, we limited the purposes of visualization itself to the following content:

- Subject data to be collected: Data on hot-springs resorts in Japan, published on the Internet
- Purpose of visualization: To visualize the locations and water qualities of hot-springs resorts
- Steps in visualization: Obtaining data through Web scraping [4], conducting a number of preprocessing steps, and then using Cytoscape [5] to import the data as network information and visualize it in the form of graphs.

### 3.2. Data acquisition and processing

When obtaining data through Web scraping, the permission of the data provider must be obtained in advance. There is a need to consider how to avoid burdening the servers and network when actually obtaining the data. Besides, the end-user license agreement must be complied with for the data obtained. Although we used the Python language [6] as a software environment for obtaining and processing data, we arranged the model as one consisting of steps that could be used even by non-specialists, with consideration for ease of use.

(1) Analysis of Web pages

There was a need to analyze the data structure of Web pages and identify the data obtained. This can be done using the View source" feature of a Web browser (Fig. 1).

```
294 <section class=  >
295 <div  class="result">
296   <img src="http://www.spa.or.jp/img/search/result.png" alt="検索結果" />
297    2690件の温泉地が該当しています。
298 </div>
299     <ul class="result_item">
300       <li>
301       <!--温泉画像-->
302       <a href="./detail_p/?F_ID=240900&pg=0">
303       <img class="photo" src="http://www.spa.or.jp/img/search/noimage_search.png" width="250"/>
304       </a>
305       <div class="article">
306       <div class="location">三重県 名張市</div>
307       <div class="name"><a href="./detail_p/?F_ID=240900&pg=0">赤目温泉</a></div>
308 <!--泉質-->
309       <div class="spa_quality">
310         <img src="http://www.spa.or.jp/img/search/spa_quality.png"/>
311         放射能泉                                                    </div>
312
313       <div style="sales_point">滝川の渓谷沿いに位置する温泉地。赤目48滝の入り口にあり、香落渓の観光基地に好適。</div>
314       </div>
315       <div class="clear"></div>
316       </li>
317     </ul>
318     <ul class="result_item">
319       <li>
320       <!--温泉画像-->
321       <a href="./detail_p/?F_ID=240730&pg=0">
322       <img class="photo" src="http://www.spa.or.jp/img/search/noimage_search.png" width="250"/>
323       </a>
324       <div class="article">
325       <div class="location">三重県 度会郡大紀町</div>
326       <div class="name"><a href="./detail_p/?F_ID=240730&pg=0">阿曽温泉</a></div>
```

Fig. 1 Example of displaying a Web page's source

(2) We used a Python program to obtain the desired data from within Web pages through Web scraping. In this study, we obtained only data on the names and water quality of hot-springs resorts.

(3) We used the Python program to look up the latitude and longitude of the hot-springs resorts in Google Maps [7].

(4) We processed the above data using Excel and other tools and saved it as network data. An example of the format of the data is provided below. In this case, we used latitude as the Y-axis value on the graph after inverting positive and negative signs, since display coordinates and axial directions on the monitor are reversed.

Sample network data format:
Prefecture name, hot-springs resort name, water quality, X coordinate (longitude), Y coordinate (latitude)

### 3.3. Visualization using Cytoscape

Cytoscape is a tool for visualization of networks (through a graph structure). For this reason, the subject of processing needs to have a network structure. Accordingly, we decided to analyze the locations of hot-springs resorts and their local prefecture capitols, as an example of a network. Fig. 2 shows an example of visualization of information on hot-springs resorts in Yamaguchi Prefecture resulting from loading network data to Cytoscape and color-coding the information by water quality.

In the center of the graph is the Yamaguchi Prefecture capitol. From this graph, the reader can identify the mutual positioning from the latitudes and longitudes on the map and the water quality from the color coding of the hot-springs resort names.

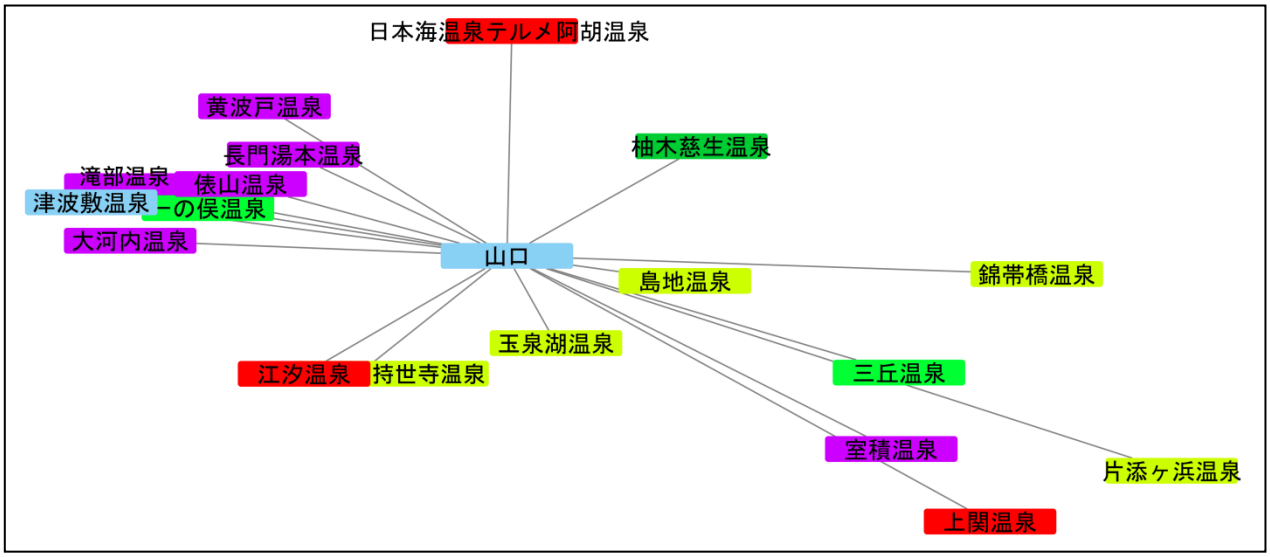


Fig. 2 Visualization of hot-springs resorts in Yamaguchi Prefecture

## 4. Discussion

In this study, ultimately we visualized structured data. First of all, the original data source of Web page source text (HTML) is semistructured data [8]. We followed the method of Web scraping to obtain data from the content of Web pages. Next, when the data obtained as in this study are composed of multiple files, there is a need for steps such as data abstraction and combination. This process requires use of data-processing tools and programming languages. In addition, occasionally it is impossible to apply general-purpose tools to conversion of semistructured to structured data, and in such cases one must rely on programming languages. Also, in the case of a locale such as Japan that employs multibyte characters, sometimes text code conversion is required [9].

## 5. Conclusion

This study employed a preparatory consideration of a system for visualization of Big Data at SMEs, elucidating a number of requirements. That is, it showed that in processes such as data collection and data processing there are many cases in which it is difficult to process the data using general-purpose tools alone. Topics for future study include development of independent tools to supplement general-purpose tools as well as development of general-purpose models for the steps involved in visualization and preparation of tutorials suitable for use by SMEs.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 15K03639.

## References

[1] IT Infrastructure for Living and the Economy of the Information-technology Promotion Agency, Japan, "Enriched living and economy from connected IT: the value and reliability of big data," http://www.ipa.go.jp/files/000001884.pdf.

[2] Small and Medium Enterprise Agency, "2014 White paper on small and medium enterprises in Japan," http://www.chusho.meti.go.jp/pamflet/hakusyo/H26/PDF/h26_pdf_mokuji.html.

[3] The Osaka Chamber of Commerce and Industry, "Results of survey on use of big data," Press Release, 2014.

[4] L. Richardson, "Beautiful Soup," Available via https://www.crummy.com/software/BeautifulSoup/. Cited 26 January 2016

[5] Cytoscape Consortium, "Cytoscape," http://www.cytoscape.org/.
[6] Python Software Foundation, "Python," https://www.python.org/.
[7] Python Software Foundation, "Pygeocoder," https://pypi.python.org/pypi/pygeocoder.
[8] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom, "Querying semi structured heterogeneous information," Journal of Systems Integration, vol. 7, no. 3, pp. 381-407, 1997.
[9] "The unicode consortium," http://unicode.org/.