Information Engineering Express
International Institute of Applied Informatics
2015, Vol.1, No. 2, 51 - 59

A New Approach to Web Mining: A Search Engine Offering Result of No Assumption

Yuri Suzuki [*], Makoto Ito [†],
Norio Ishii [‡], Takashi Hara [§]

Abstract

Recently, rapid development in information and communications technology has led to explosive growth in the amount of available information, including substantial volume of data on the Web. One step toward managing and making use of this explosion of information is to enhance search technologies so that they can easily retrieve the necessary data. To this end, numerous studies are already underway, focused on Web navigation, Web mining, and related fields; however, due to the massive amount of information available on the Web, the precision of Web mining is not yet very high. In this study, we propose a method for providing multifaceted search results and effective Web mining by not only using keywords but also leveraging relationships that include information contained within a resource.

Keywords: serendipity, web mining, web navigation, Wikipedia

1 Introduction

Recently, rapid development in information and communications technology has led to explosive growth in the amount of available information, including substantial volume of data on the Web. These developments have become evident since the beginning of the 21st century, with the International Data Corporation predicting that we will reach 40 zettabytes of data by 2020 [1]. Furthermore, with the expansion of services, such as online shopping, blogging, and social networking services (SNS), as well as the increase in the number and variety of smartphones, the content we generate continues to diversify with the Internet users who utilize such content making up an increasingly greater proportion of the overall population [2]. More and more opportunities to use websites to target the increasing number of users for enterprise information and advertising has provided a momentum for the explosive generation of information [3].

[*] Chubu University, Aichi, Japan
[†] Meitetsu Information System Co., Ltd., Aichi, Japan
[‡] Aichi Kiwami College of Nursing, Aichi, Japan
[§] Snapshot Inc., Aichi, Japan

One step toward managing and making use of this explosion of information is to enhance search technologies so that they can easily retrieve the necessary data. To this end, numerous studies are already underway, focused on Web navigation, Web mining, and related fields; however, due to the massive amount of information available on the Web, the precision of Web mining is not yet very high [4]. In current Web searches, emphasis is still placed on the use of search keywords entered by users to determine valid resources. In this type of search, only a resource's relation to the keywords is evaluated, rendering the multifaceted evaluation of resources is impossible.

In this study, we propose a method for providing multifaceted search results and effective Web mining by not only using keywords but also leveraging relationships that include information contained within a resource. As the first stage of our research, we focus on Wikipedia, in which target resources are written as per predetermined rules, and we investigate its utilization.

2 Proposal and Advantages

In current Web searches, emphasis is placed on search keywords entered by users, which are then used to determine which resources on the Web are valid (Figure 1). In our study, the term resource refers to elements that fulfill a search objective or are necessary elements.

As a specific example, if a user entered "Freud" as a keyword, he or she would then select resources from candidates that include "Lucian Freud," "Sigmund Freud," and the movie "Freud." However, because this type of search only evaluates a resource's relation to given keywords, it makes multifaceted evaluation of the content of Web resources impossible.

Therefore, in this study, we aim to provide multifaceted search results and an effective Web mining system that uses the relationships of contents taken up on the Web and in the physical world and derives the relationships of contents in the knowledge layer (Figure 2). More specifically, if a user enters a person's name, such as "Lucian Freud," the system would analyze resources to derive contents related to the person, including their "birthplace" and "academic history." Searching resources related to those contents would allow the system to provide multifaceted search results. For example, the system could automatically provide results unrelated to the keywords specified by the user, such as resources related to painter George Grosz, who shares a birthplace with Lucian Freud, or a resource related to designer Alexander McQueen, who went to the same school as Freud.

Discovering contents based on these types of relationships provides potential advantages and opens the door to new approaches in many various fields. For example, in work related to planning exhibits in art museums, the system could provide knowledge that could be used to inspire new plans based on new approaches, such as displaying works by artists from the same birthplace. Furthermore, by applying this to investigative learning in the education field, the system could allow multifaceted learning through new approaches by providing contents related in ways unexpected to the learner.

In a typical web search, the information being searched for can be categorized as either a "thing" or "matter." Searches for things include, for example, a search for a work by a specific artist with a particular title. These are searches in which the user is looking for a specific, physical entity that he or she has seen or heard about. Examples would include searches

for "Édouard Manet" or his work "The Fifer." These types of searches are the target of existing search engines and are not treated in this research.

Searches for matters are those in which the user has a vague interest in some matter (such as a specific artist or work). This research proposes a system that will allow an approach to search for further information on the web related to the matter being searched based on an analysis of knowledge on the web. To provide an example, when checking a specific artist, one could reference the school from which he or she graduated using public sites. Organizing these matters would allow other artists who also graduated from the same school to be found based on the associations with that school. Hence, a search for Lucian Freud would display Alexander McQueen.

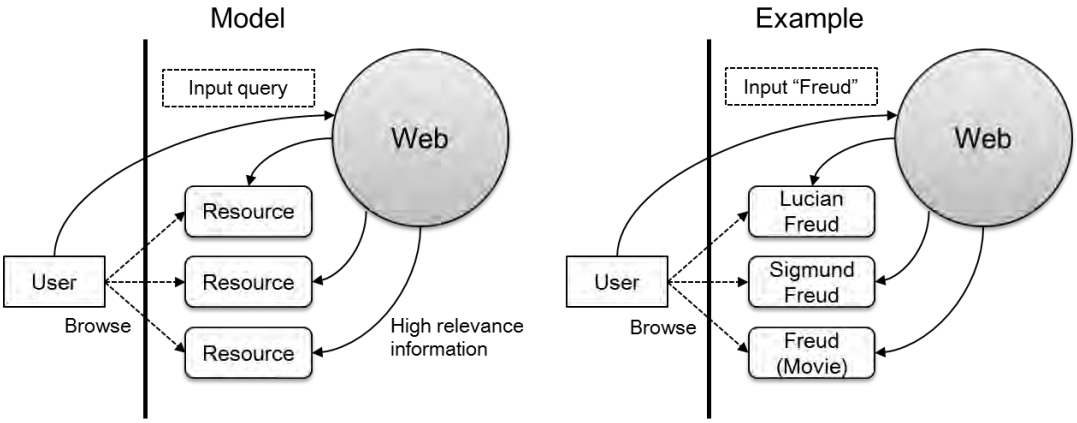

Figure 1: Example of a current keyword-based search system

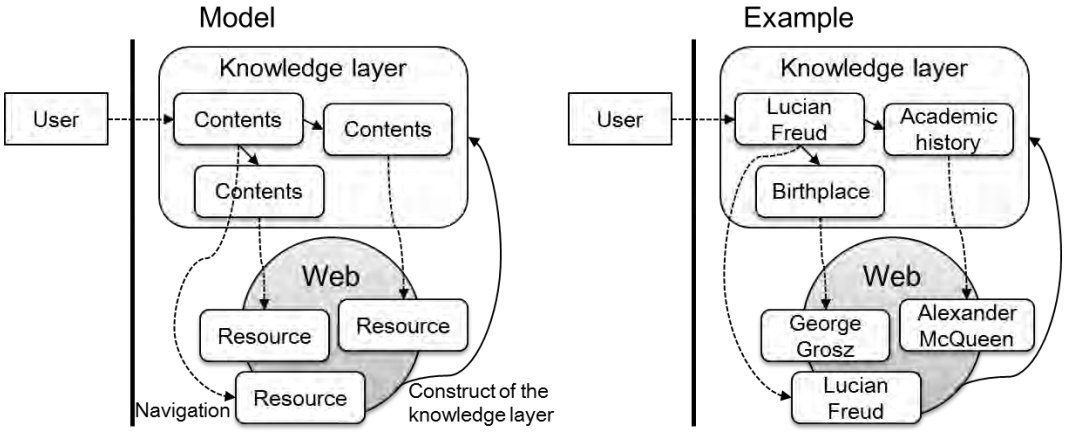

Figure 2: Our proposed system

3 Constructed System

In our study, we aimed to construct a base for the abovementioned Web mining system. This required a target for analysis that was organized and structured in such a way to make analysis possible. We also had to avoid problems with copyrights related to the target for analysis. Wikipedia was the optimal solution to fulfill these conditions. Therefore, in this study, we introduce a method that analyzes a data dump from Wikipedia, which is publically provided free of cost.

3.1 Wikipedia Usage

Wikipedia is characterized as a large-scale, wiki-based web encyclopedia that allows any user to modify article content through a web browser. The ease of editing encourages internet users to add content, and the resultant site encompasses massive amounts of content ranging from general concepts to new ideas in several fields such as culture, history, mathematics, science, society, and technology. As of November 2011, Wikipedia had reached 3.8 million English-language articles and 780,000 Japanese-language articles. Given that the world's largest encyclopedia, Britannica, contains approximately 65,000 articles over 60 volumes, Wikipedia encompasses 46 times the number of articles.

In addition to encompassing several topics, Wikipedia has numerous features such as term uniqueness through concept differentiation using URLs, allowing narrow categorical searches using structures with category links, and semi-structured data in an easily analyzed data format [4].

Using metadata as valuable information is crucial for clarifying the Wikipedia structure. Thus, terms with shared recognition of meanings are required. For this reason, it is necessary to establish bibliographical data, such as Wikipedia shared recognition in Dublin Core, web and text authors, titles, and creation date as vocabulary to be described in metadata. In this study, we construct metadata for obtaining target data from the massive amount of data on Wikipedia. Appending metadata to individual pieces of information allows searches that accurately reflect the character of the data in question. In this study, we form basic metadata as per standards designed to be compatible with any metadata description method.

There are three main methods for analyzing Wikipedia data: collecting and analyzing HTML, analyzing RSS transmitted as the most recently modified content, and analyzing public dump data with no charge. In this study, we use the dump data analysis method. Publicly available files are given for each language in the following form: jawiki-latest-pages-meta-current.xml. This is an XML format file that allows users to specify custom tags. Because XML features standardized descriptions using tags, the structure is easy to comprehend and even massive amounts of data can be easily analyzed once the structure is understood.

3.2 System Functions

The constructed system is composed of three functions (Figure 3). The first function converts resources provided by Wikipedia to a database for use in this system. The second function constructs the knowledge hierarchy. Using resources obtained by parsing Wikipedia, this converts resources that compose the knowledge hierarchy into a database. The third function is an output function that allows displaying the results of mining resources converted to a database based on the knowledge hierarchy.

3.2.1 Database creation

n the first and second functions, database systems are introduced for the resource database conversion and knowledge hierarchy database conversion. In this study, we use mongoDB [5]. This is a non-RDBMS, open-source document-oriented database categorized as Not only SQL, which is non-RDBMS. Its features include eliminating the need to define fixed schema such as in RDBMS, the ability to store array data, and the storage of data in a format similar to the lightweight JavaScript Object Notation (JSON) descriptive language. In addition, it can utilize Wik-

ipedia XML data nearly as-is, and MapReduce (described later), which is used in the knowledge hierarchy construction function, can only be processed in mongoDB. Moreover, we use Ruby as the scripting language and Rails (Ruby on Rails) for the web application framework [6], [7]. Ruby provides drivers for mongoDB (Ruby-mongo-driver) [8]. We use Mongoid as an O/R mapper to map objects to a relational database [9].

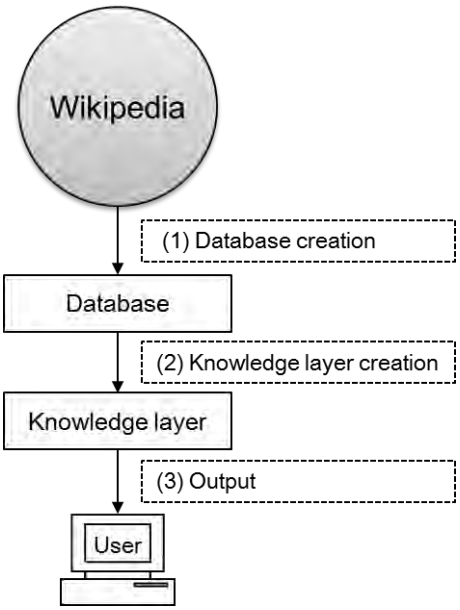

Figure 3: Functions of our Wikipedia-based multifaceted search system

3.2.2 Knowledge layer creation

The second function, which constructs the knowledge hierarchy, performs clustering using MapReduce. MapReduce is a software framework introduced by Google in 2004 to support distributed computing on large-scale datasets [10]. The knowledge hierarchy construction function parses Wikipedia to obtain resources, and then, based on the data, it converts resources that compose the knowledge hierarchy into a database. To implement this function, the system extracts article templates with the Map process and groups articles with the same template formats and values with the Reduce process.

The Map process performs the following three processes. First, a keyword is entered. Then, it searches Wikipedia resources and confirms the template in use. Finally, it generates JSON data from Wikipedia XML data. This contains the following: {id: {template name and key elements and their values}, Wikipedia pageId}. Here, the key element refers to template property data, which is used to perform new searches based on this content.

The Reduce process performs the following two processes. First, it rearranges the data generated in the abovementioned Map process using id: {template name and key elements and their values}. Next, it aggregates pageIds. These processes generate data for new searches.

3.2.3 Output

To verify that the content of the constructed knowledge hierarchy is appropriate, we implement an output function that displays the results of mining resources converted to a database based on the knowledge hierarchy. This function uses the arbor.js library, which allows the display of graphs in HTML5 [11]. In addition to graph display, a function to output text is implemented.

3.3 Libraries and Licenses

Table 1 shows licenses related to the software libraries used in the system.

Table 1: Libraries and licenses used in the system

Library	Version	URL	License
arbor.js	0.91	arborjs.org	MIT license
MongoDB	2.4.8	www.mongodb.org	GNU AGPL v3.0
Ruby	1.9.3	www.ruby-lang.org/ja	2-clause BSDL Ruby's license
Ruby on Rails	4.0.2	rubyonrails.org	MIT license
Ruby-mongo-driver	1.9.2	api.mongodb.org/ruby/current	Apache license v2.0
Mongoid	4.0.0	mongoid.org	MIT license

4 Trial Results and Discussion

In this study, we report some of the mining results from our constructed system.

Figure 5 shows results related to the keyword "The Fifer." The results display other works by the artist Édouard Manet, including "A Bar at the Folies-Bergère" and "Portrait of Émile Zola," another work composed in the same year, "The Origin of the World", and other works at the Musée d'Orsay, including "The Church at Auvers" and "Portrait of Dr. Gachet." In addition, "Breezing Up" is shown as a work of the same width. While groupings, such as works by the same artist or works at the same museum, are predictable, categorizations using grouping such as "works of the same width" provide unexpected data. This information would be new to a user who has personally never had the chance to see the work, and could not be obtained just by viewing images obtained from image data on the internet.

Figure 6 shows the results related to the keyword "loquat," which is a plant that belongs to the Rosoideae and Maloideae subfamilies; results show that "rose" and "medlars" belong to the same classification. In addition, because the scientific name of the loquat was determined by Carl Thunberg, other plants and organisms named by Thunberg are shown, including the "trumpet lily" and a mammal of the bovidae family called the "blue duiker." In this example, the information related to the plant's classification is expected, but the system also presents information unrelated to the plant through the relationship with the scientist responsible for its name, which we call a new discovery. As another example, in investigative learning, a learner's approach could expand from an investigation into the "loquat" itself to an investigation into the one who named it through this newly discovered relationship.

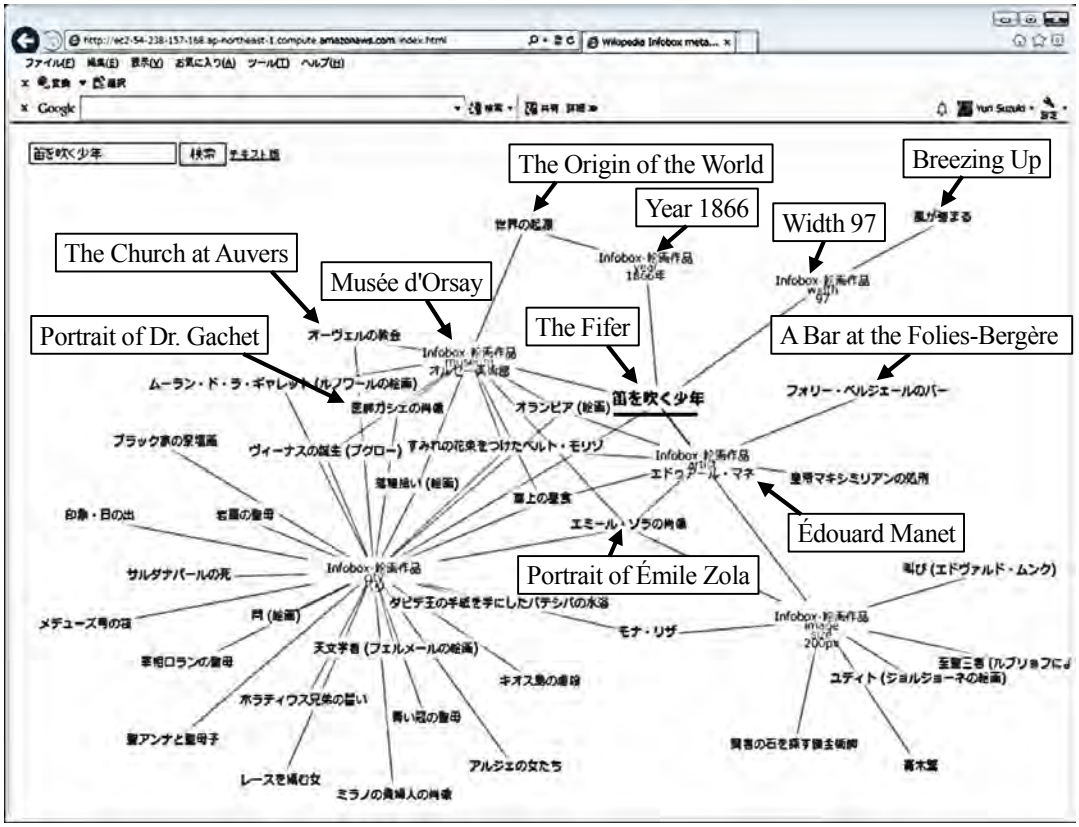

Fig. 5. Mining results based on "The Fifer"

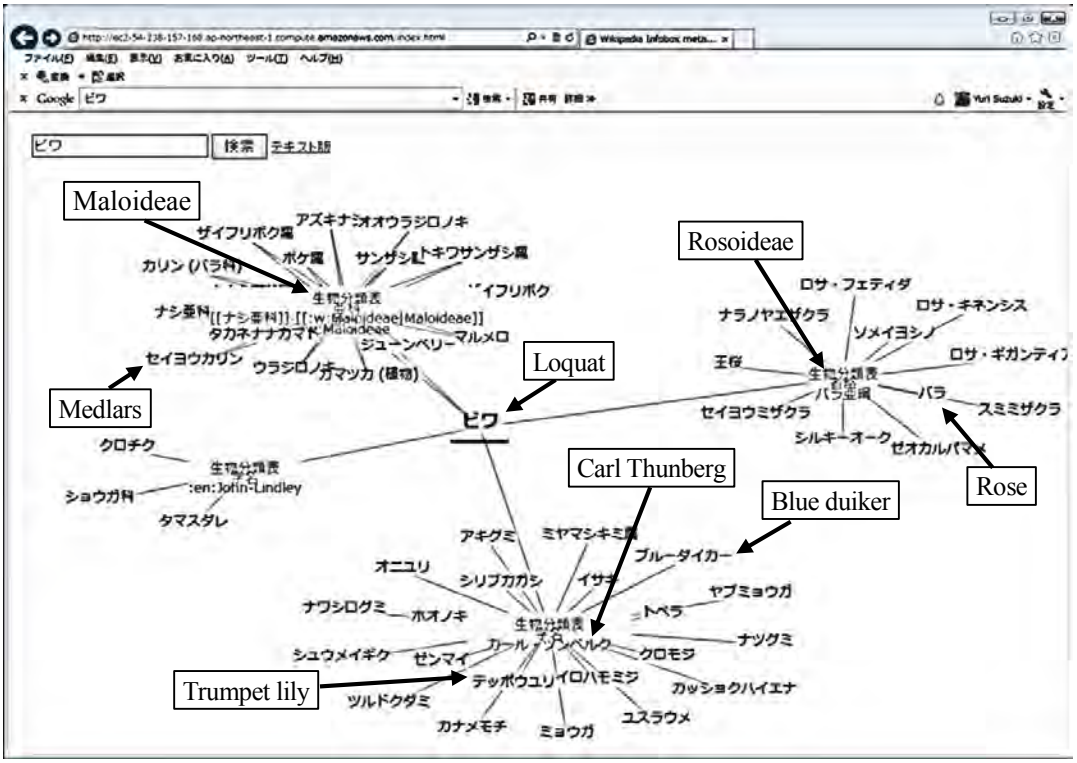

Fig. 6. Mining results based on "loquat"

Figure 7 shows the results for "Louvre Sculpture Museum;" based on these results, we observe that the "Kitano Museum of Art" and the "Kanno Museum of Art" are also museums specializing in sculpture. We also realize that many other museums were opened in 1987, including the "Meguro Museum of Art" and the "Oita City Historical Museum." Furthermore, the Louvre Sculpture Museum itself was designed by architect Kisho Kurokawa, and other museums he designed, including the "Nagoya City Art Museum" and the "Kumamoto City Museum," are shown. In this example, the field of art that is the central focus of the museum is expected, but the system also obtains results that have new meanings, such as the fact that the museum itself is a work of art by a famous architect.

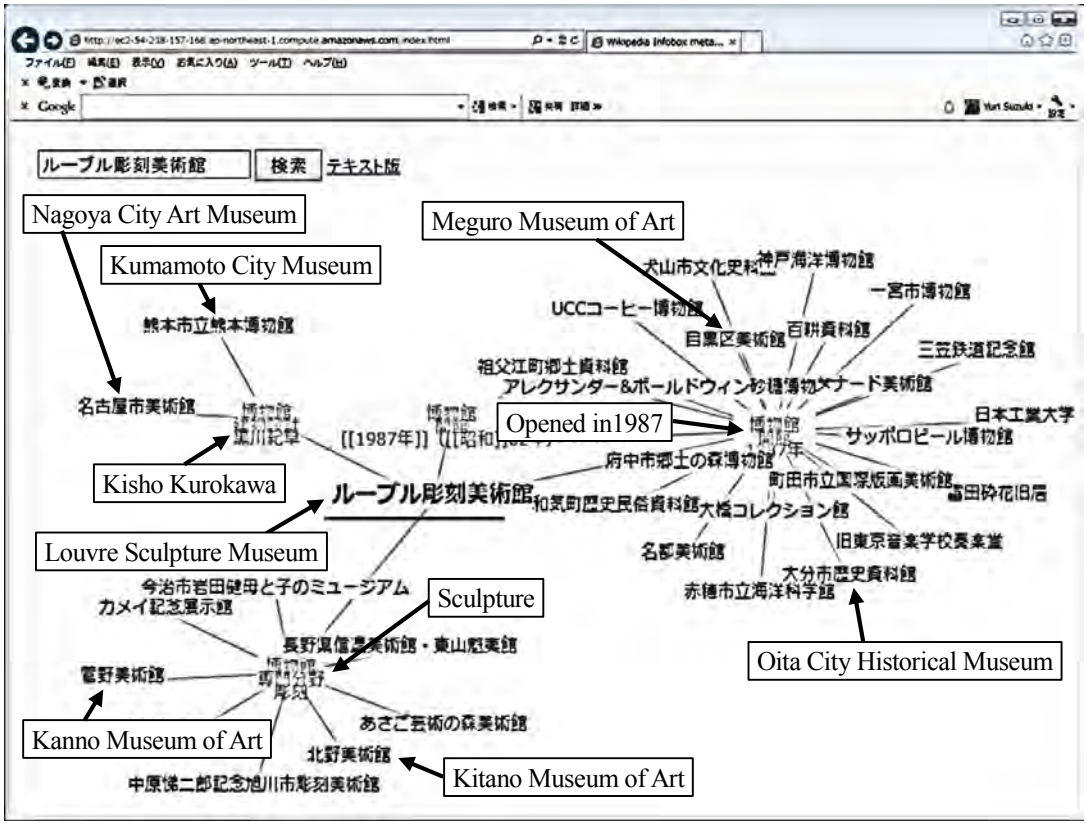

Fig. 7. Mining results based on "Louvre Sculpture Museum"

5 Conclusion

As was summarized above, we have successfully proposed a new approach for mining and shown its capabilities via a constructed prototype system. In future studies, we plan to perform an evaluation experiment of our system.

In addition, assessments are required to determine whether mining results provide information that is useful to the user. We plan to conduct assessment experiments of the proposed foundation system using cognitive science approaches, including the assessments of the constructed system's usability. After making improvements to the foundation system based on analysis and consideration of the experiment results, we intend to provide web navigation as the second step of this study.

In addition, the data provided through Web navigation in this study is expected to be applicable as a fundamental data for big data analysis. Current systems that analyze SNS data and blogs stop at the quantitative evaluation of the results of morphological analysis of text; however, the system proposed here could allow for data analysis other than simple quantitative evaluation. For example, assume that the nouns "Lucien Freud," "Taro Okamoto," and "Salvador Dali" frequently appear in a certain SNS. If these were linked to the proposed system, the property "surrealism" would be derived. This would offer the possibility of performing data analysis in regard to the question "is there a surrealism boom happening on that SNS?" that is not limited to simple quantitative assessment. In future studies, we will perform research exploring these possibilities.

References

[1] International Data Corporation, www.idc.com.

[2] Ministry of Internal Affairs and Communications, Japan, "Information and Communications in Japan," www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2012/2012-index.html.

[3] M. Kitsuregawa, "Info-plosion: Retrospection and Outlook," The Journal of the Institute of Electronics, Information and Communication Engineers, vol. 94, no. 8, 2011, pp. 662-666.

[4] K. Nakayama, M. Ito, M. Erdmann, M. Shirakawa, T. Michishita, T. Hara, and S. Nishio, "Wikipedia Mining: A Survey on Wikipedia Researches," IPSJ Transactions on Database, vol. 2, no. 4, 2009, pp. 49-60.

[5] mongoDB，http://www.mongodb.org.

[6] Ruby，https://www.ruby-lang.org/ja.

[7] Ruby on Rails，http://rubyonrails.org.

[8] Ruby-mongo-driver，http://api.mongodb.org/ruby/current.

[9] Mongoid，http://mongoid.org.

[10] MapReduce，http://research.google.com/archive/mapreduce.html.

[11] arbor.js，http://arborjs.org.