ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(S9), 235–239, May 2015* DOI: 10.17485/ijst/2015/v8iS9/68316

The XML based Electronic Document Image Retrieval System

Cheol-Joo Chae, Kiseok Choi and Kwang-Nam Choi*

Department of R and D System Development, Korea Institute of Science and Technology Information, 245 Daehangno, Yuseong-gu, Daejeon, Korea; knchoi@kisti.re.kr

Abstract

The amount of multimedia information has been increasing by the rapid growth of the Internet technology. Recently the need for the efficient image retrieval technology has been raised since an electronic document includes a significant amount of image information. In this paper, we designed and developed the information retrieval system for images and tables included in an electronic document by converting an electronic document to XML. The method proposed in this paper represents the information of images and tables in the XML format. Therefore the system could be implemented conveniently using the commercial system supporting XML and increased the retrieval accuracy for images and tables.

Keywords: Electronic Document Image, Image Extract, Image Retrieve, Table Retrieve, XML

1. Introduction

The share of images in an electronic document is increasing based on the growth of digital image related technology, recently. Therefore, the research interest in the use and search of images in an electronic document is growing too. Currently search engines provides image search services on the web, however the accuracy and efficiency are not sufficient[1,2]. The XML format becomes a standard format to exchange information and describe the data structure. Moreover, studies about the data-technology standard for the purpose of information storage and exchange are being conducted in a range of fields. Therefore, this paper proposed the image retrieval method by converting an electronic document to the XML format. We attempted to increase the accuracy of the image retrieval technique by using the XML format to extract and store an image, and search the location information of an image. Additionally we built the database for image search designed for a web-based image retrieval system. The structure of this article is, in the first chapter, the previous studies on the image search are reviewed; awe explained the design and the experimental result in chapter 3; and the conclusion is in chapter 4.

2. Related Works

The previous image retrieval technology can be categorized into three types. The first type is simple retrieval technology based on keywords. This approach extracts the topic or related keywords of an image, stores in the database, and manages the data to be searched. Normally the keyword-extraction method is not possible to be done automatically; it is required to represents manually to build the image-content information. Recently a system consists of index, which are extracting keywords from web documents is also employed. This method requires the technology based on expert knowledge of relevant sectoral images. Due to the different technologies, it is hard to interpret the meaning of the image, or assure the accuracy of the keywords. Moreover, unnecessary contents or keywords can decrease the searching efficiency when extracting automatically from the web documents such as HTML. The second technology is the

*Author for correspondence

content-based image search technology based on the specific vectors of an image. The data included in the image such as color, sharp, spatial and textures are the representative feature vectors. There are studies on the content-based image search technology that stores and manages the feature vectors in the database, and searches by comparing with the vector-based data. Normally the feature vectors of an image are automatically extracted and stored to support the content-based image search following the preprocessing process considering the system performance and search time. Although automatic image extracting can be adequate, because of the large volume of the extracted data and the significant cost for preprocessing process, it is not a proper approach for searching a large volume of images. Moreover for searching based on the automatically extracted feature vectors, the accuracy can decrease since the search intention or the identifiability and meaningfulness included in the image cannot be considered.

The third approach mixes above two methodologies to increase the efficiency. It extracts the image vector data automatically after the pre-processing process of the image, and represents the meta-information manually such as keywords, meaning information, or visual information which is not able to extract automatically. However this system also cannot provide the result as appropriately modified forms by interpreting various search methods and meaningfulness. It still may cause a problem during the integration or exchange with other systems. Therefore the data modification technology that can search data using various methods in various environments is required. The introduction of the multimedia standard technology for the data exchange between different systems is also required. We applied the XML technology to structuralize the content information of an image, XQL (XML Query Language) to query the data from the XML document, and the XSL (eXtensible Stylesheets Language) technology to convert the documents[3–7].

3. The Design and Evaluation of Proposed System

3.1 Proposed Method and Design

We extracted images and tables as an object (image and table) or a JPEG format, and stored in the data repository to retrieve images and tables from the electronic document. The index and tag information are structured using the caption information of the objects. The chapter, section, and paragraph, and the page information are added for searching the electronic document. We also designed to store the thumbnail files separately to provide the search function after storing the image and table information. The logic to extract images and tables is designed as follows. The electronic document is converted to XML by entering the file name of the electronic document. The XML Converter extracts the objects (images and tables) after processing sentences and words with more than one line.

3.2 Evaluating the Performance of the Proposed Method

The proposed method successfully extracted 123 images from the total 126 images in ten documents but failed to extract three images. For the table extraction, it extracted 39 tables and failed to extract five. For the image extraction, the method showed the success rate of 99.8%, and it shows the success rate of 89.6% for tables. Figure 2 presents the successfully extracted images and tables. The reason for the extraction failure was the caption-information error occurred when the electronic document was created. Figure 3 presents the result of the failure. For the figure with two captions in a line, both of the captions were extracted. For the table, if there was no table line below a caption, the algorithm checked if the table existed above the caption. However, hidden lines below the caption were extracted on the page. To tackle the limitations, for an image, a line adjoined the first line of a caption was processed to be recognized as a caption. For a table, a line adjoined the first line of a caption was processed to be recognized as a continuing caption based on the coordinate information, and if there was a table border (line) between the adjoined lines, the lines before the border were considered as a continuing caption. For both images and tables, if there were more than two continuing lines, only the first two lines were saved as a caption. We structured a temporary working database and repository to build the image and table information of the electronic document using the method proposed. We then saved the image and table information after converting the electronic document to XML, and developed the search system based on the caption information on the web. Figure 5 presents the search results of the images and tables in the electronic documents on the web[8–10].

Table 1. The process of converting the electronic document to xml, and extracting images and tables

1. Enter the file name of the electronic document. 2. Create the XML document from the electronic document. 3. Read metadata from the database. 4. Front configuration - Construct the XML document from the data in the database 5. Body configuration - Review all of the lines in all of the pages. - Create the XML items for the image and table patterns. - Create the relevant XML items for the chapter and section patterns. (Consider the chapters and sections up to the third level) - Create the relevant XML items for the text patterns. - Create the XML based on the chapters and sections by analyzing the sentences and words. - Extract the images of the object based on the captions of the images and tables. 6. Save converting logs. 7. Save the created XML in the database.

Figure 1. The XML-based image and table extraction class diagram

4. Conclusion

In this paper, we proposed a method to retrieve images and tables in the electronic documents using XML. We also proposed a method to search images and tables in the electronic documents on the web by applying the proposed method. The method proposed in this paper represents the information of images and tables in the XML format, so that the system could be implemented conveniently using the commercial system supporting XML, and increased the retrieval accuracy for images and tables.

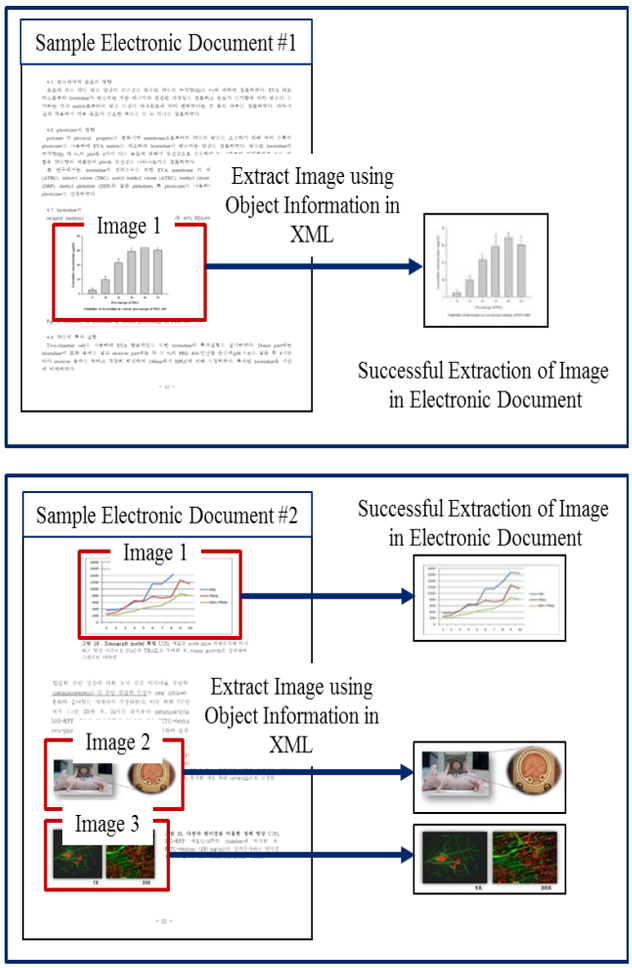

Figure 2. The examples of image extraction succeeded.

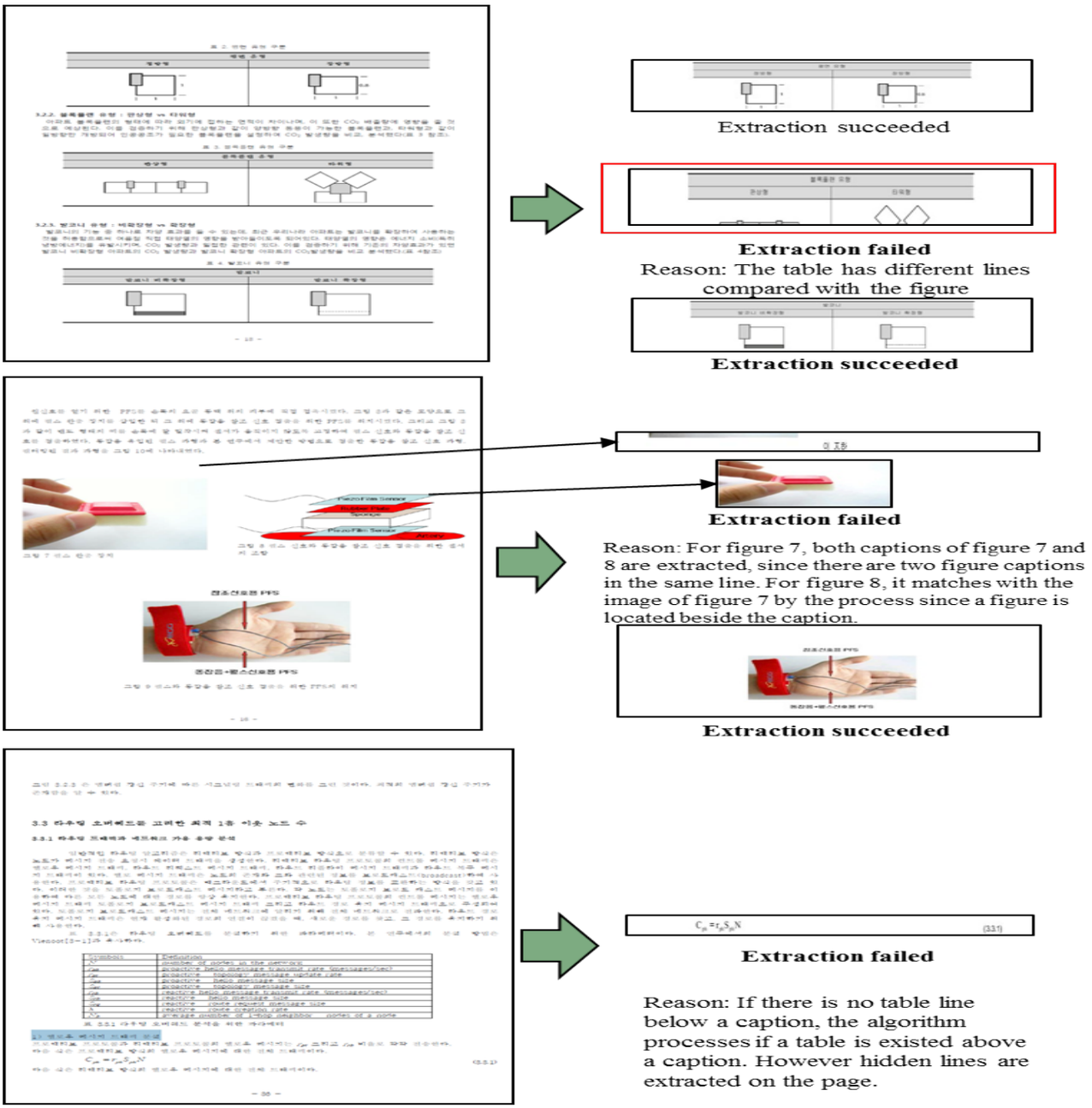

Figure 3. The examples of image extraction failed.

In the future, we will study how to extract the images and tables included in the electronic documents automatically.

5. Acknowledgment

This research was supported by Maximize the Value of National Science and Technology by Strengthen Sharing/Collaboration of National R and D information funded by the Korea Institute of Science and Technology Information (KISTI).

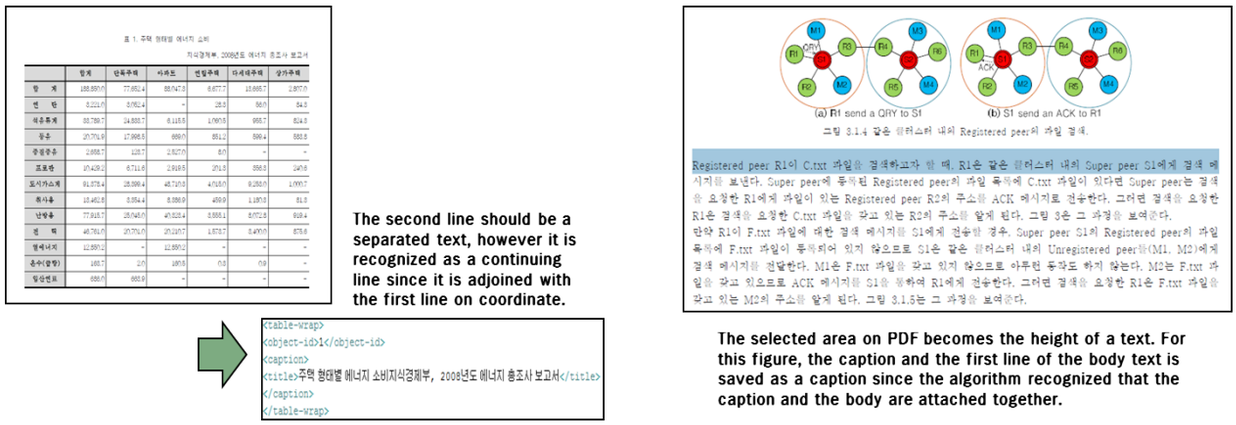

Figure 4. The example of the solution.

Figure 5. The search result of the image and tables in the XML-based electronic documents.

6. References

1. Kim J. Implementation of XML Document Structure Retrieval Engine. Gradurate School, Honam University; 2001.
2. Nam SH. A study on metadata for an image retrieval system. Graduate School, Yonsei University; 2001.
3. Jagadish HV. A retrieval technique for similar shapes. Proceedings of the 1991 ACM SIGMOD international conference on Management of data; 1991. p. 208–17.
4. Hong S, Nah Y. An Intelligent Image Retrieval System Using XML. Journal of Korea Multimedia Society. 2004; 7(1):132–44.
5. Androutsos D, Plataniotis KN, Venetsanopoulos AN. Image Retriveal Using the Directional Detail Histogram. Stroage and Retrival for Image and Video Database SPIE. 1998:129–39.
6. Hong S, Lee C, Nah Y. An Intelligent Web Image Retrival System. Proceeding of SPIE: Internet Multimedia Management System II; 2001:106–15.
7. Chu WW, Leong IT, Taira RK. A Semantic Modeling Approach for Image Retrieval by Content. VLDB Journal. 1994; 3:445–77.
8. National R and D Report Registration Management System. KISTI. Available from: http://nrms.kisti.re.kr
9. NDSL Research Report. KISTI. Available from: http://www.ndsl.kr
10. Final report of national science and technology knowledge information service program. KISTI; 2012.