ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(S1), 340-345, January 2015* 10.17485/ijst/2015/v8iS1/59327

A Method of Extraction of Non-text Contents for Extending the Applicability of National R&D Reports

Kiseok Choi, Kwangnam Choi and Jaesoo Kim*

NTIS Center, KISTI, Taejon, Korea; choi@kisti.re.kr, knchoi@kisti.re.kr, jaesoo@kisti.re.kr

Abstract

Background/Objectives: A research report is textual information on performance. With the value of science and technology, it is very critical for industrial and economic purposes such as follow-up studies, technology transfer and commercialisation. **Methods/Statistical Analysis:** The research report information retrieval service provided by the Korea Institute of Science and Technology. Information (KISTI) offers optimal keywords for search conditions by indexing the report contents. However, various forms of non-textual contents such as tables and figures are often left out from the information retrieval without being included in the indexing. In terms of search accuracy and efficiency and user convenience, therefore, it is hard to support them efficiently. **Results:** Hence, this study developed a method to extract non-textual contents from a research report and use them in information retrieval with a goal of improving the accuracy and efficiency of information retrieval. **Conclusion/Application:** This study suggested a development plan for a non-textual content processor which can extract and store tables and figures and provide search services. It appears that there would be more opportunity to use high-quality national R&D report database.

Keywords: Information Retrieval, Non-Ttext Contents Extraction, R&D Report Management, XML Data Management, XML Data Parsing

1. Introduction

The academic meaning of research performance is defined as unique and valuable knowledge which is created during the research and becomes available publicly[1]. In other words, it refers to the achievement in science and technology such as patent and thesis and other tangible and intangible economic, social and cultural results. In particular, a research report is textual information on performance with the value of science and technology. It is very critical for industrial and economic purposes such as follow-up studies, technology transfer and commercialisation[2,3]. A national R&D report is a program promoted by the central administrative body with government budget or funds for R&D in science & technology[6]. Specifically, it means the records of the results of national R&D programs stipulated in Article 11 of Framework Act on Science and Technology. For effective collection and use of research reports, the KISTI was named 'research outcome management institute' for national R&D reports. Since then, it has collected research reports and developed database to provide services[8]. The current report information retrieval service provided by the KISTI offers optimal keywords for search conditions by indexing the report contents. As a result, researchers were able to search the data they wanted by entering

* *Author for correspondence*

keywords[4,5,9,10]. However, various forms of non-textual contents such as tables and figures are often left out from the information retrieval without being included in the indexing[11]. In terms of search accuracy & efficiency and user convenience, therefore, it's been hard to support them efficiently[12]. Hence, this study developed a method to extract non-textual contents from a research report and use them in information retrieval with a goal of improving the accuracy and efficiency of information retrieval. This study is structured as follows: In chapter 2, the overview of non-textual contents is introduced. For example, the definition and examples of non-textual contents are stated, and the restrictions targeted to extract non-textual contents from the PDF documents among various forms of non-textual contents in research reports are described. In chapter 3, a processor designed to extract non-textual contents from research reports is explained. The primary design for handling non-textual contents is introduced, and then the workflow and functions of the processor are defined. In chapter, conclusion and future work are given.

2. Non-Textual Contents

2.1 Definition of Non-Textual Contents

In this study, 'non-textual contents' refer to all objects but general texts, which are being entered by researchers while preparing a research report using a word processor. They usually include tables and figures, and each of them is captioned. A caption is made of words that explain a picture or table. It provides information with which an object can be identified during structural analysis after being converted to a PDF file format. In addition, the contents of the caption are used in indexing words for information retrieval. The Figure 1 reveals the example of non-textual contents in national R&D reports. The figure on the left refers to image data while a table on the right side is common non-textual contents.

Figure 1. Example of non-textual contents.

2.2 Non-Extraction Types and Restrictions of Non-Textual Contents

Non-textual contents randomly appear in a research report in table and figure formats. Depending on how these tables and figures are positioned in the contents, they could be later detected or not when non-textual contents are extracted. This section attempts to unveil the types formed in these non-textual contents and clarify the parts which are hard to be extracted so far.

The non-textual contents in research reports were established to use non-textual contents such as tables and figures to provide high value-added services in national R&D outcome. In extracting these contents, a non-textual content extractor extracts non-textual contents (images, tables) from the PDF files and separately stores them by database and file for the purpose of retrieving information and using them as non-textual contents for the report. In 2013, the non-textual contents were extracted from disclosed reports among the national R&D report database. Specifically, a total of 28,616 texts in a PDF format were obtained. The non-textual content extractor uses a PDFbox library which extracts full texts from the PDF file. In addition, it acquires tables or images based on location information using the captions of non-textual contents. The extracted non-textual contents store storage path and XML information on tables and images in a table structured by the report control number. The actual table and image files are classified by the type of contents and stored in the storage comprised by the report control number and report page. Regarding the extraction of non-textual contents from the PDF file, XML information is created. Then, created XLM data are shown in the Figure 2. Non-textual contents include information on the page in which the contents are located and storage paths of the images and tables. In terms of the location of non-textual contents, the location in which caption is stated in the PDF file is recorded. Based on the caption information, then, the location information of images and tables are recorded.

Figure 2. Non-textual data extraction XML.

In 2013, non-textual contents were extracted from 28,616 texts (disclosed texts in PDF format). In 1,871 texts (6.5%), however, it wasn't able to extract non-textual contents because of failure to extract particular information or read files such as absence of location information in the PDF file, undecided font, absence of standardized PDF and PDF conversion error in raw report files (hwp, doc, etc.). The cases of failing to extract non-textual contents from the DPF file are as follows:

1) Absence of caption on tables and images in PDF file
Because there is no caption on tables and images, the location information is unknown when tables or images are only inserted without entering caption on tables or images at the preparation of a report. Therefore, even though tables or images exist in the PDF file, the location information cannot be read unless there is caption.

Figure 3. Example of caption-less non-textual contents.

2) When non-textual contents cannot be extracted due to a composition error in tables or images
As shown in the Figure 4, because tables and images are dual-structured in preparation of a report, the non-textual contents may not be extracted. If a table or image is included in the table, for example, they cannot be extracted without finding location information.

Figure 4. Images in the table (wrong caption).

3) Unless a font is decided in the PDF file
At PDF conversion with source files (ex: hwp, doc, etc.), a font shall be substituted by basic font (PDF basic font or system font) when the font in the source file is used, or a font in the original file is a certain font (the one which is not supported in the PDF file). The location information which extracts tables and images is developed based on the texts in the PDF file. In some reports, a font prepared at PDF conversion isn't defined yet. Therefore, location information on tables and images weren't properly read. In addition, it becomes hard to extract unintended tables or images

4) In case it is not a standard PDF file
In case a domestic processor (ex: hwp, etc.) is converted into a PDF file prior to 2002, the access information in the text and table becomes inaccurate because of a PDF composition error after failing to observe PDF standards. In this case, even though the location information of text is read, it is unable to read the location information of tables and images properly because of inaccurate access information to them. As a result, an unintended table or image would be extracted, or it is hard to extract the location information.

5) If the PDF's header information is damaged
In case header information in which a PDF's composition (version, conversion type, etc.) with the file converted into a PDF file prior to 2002 is damaged, or it is unable to read the information because of an old version, the extractor may not be able to read the PDF's basic information or get access to the PDF file either.

3. Design of Non-Textual Content Processor

To extract tables and images from the PDF file and store them, the non-textual content processor implemented in this study is designed to develop information extraction and storage functions on the tables and images in the report texts, extract non-textual content objects in a JPEG format and store them in repository and compose indexing and tagging information using caption on the objects. The design aimed to handle non-textual contents is shown in Table 1, and the overview of the designed non-textual content processor is shown in Figure 5:

Table 1. Design of non-textual content processor

1.Extract non-textual content objects in a JPEG format and store them in repository.
2. Extract the object caption up to the line 2 and compose indexing and tagging information.
3. Add XML information in which chapter, section and paragraph information is included and use it in searching the report.
4. Add tagging information (chapter/section/paragraph, page location information, etc.) to the XML information and use it in the detailed search of reports.
5. Add page to all page-unchecked XML chapter, section and paragraph.
6. Add object path information to conventional XML and page tag-less XML, revise schema and add sources.
7. Test from the start chapter, section and paragraph to the next chapter, section and paragraph in <p> tag.
8. In case of two or more caption lines, extract and store indexing information.

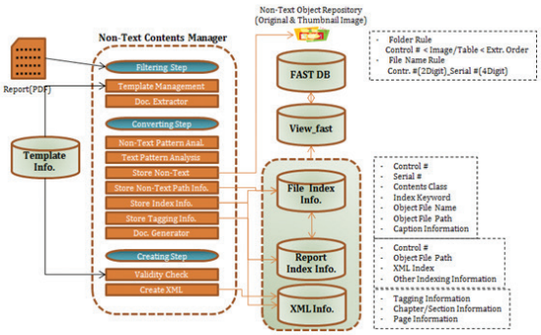

Figure 5. Overview of non-textual content processor.

3.1 Design of the Flow of Non-Textual Content Processor

To extract tables and images from the PDF file, an extraction process was designed as shown in Figures 6 and 7.

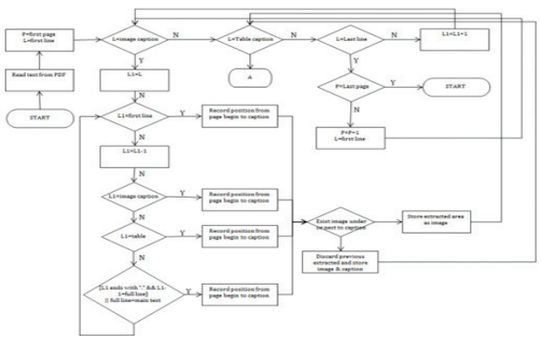

Figure 6. Non-textual data extraction process I.

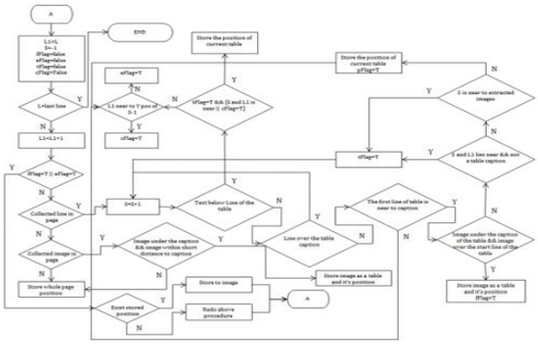

Figure 7. Non-textual data extraction process II.

To provide search functions after storing non-textual contents, it was designed to store the thumbnail file of non-textual contents separately for professional search. Then, UI was designed to use indexing/tagging information extracted from the non-textual content processor for the search based on the non-textual content location information in an XML file[13,14]. This kind of non-textual content search processor is shown in Figure 8.

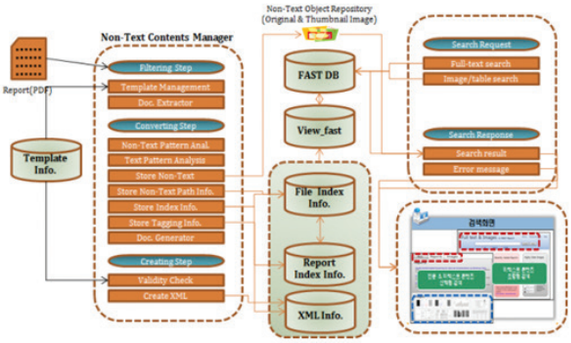

Figure 8. Non-textual content retrieval process.

To extract non-textual contents, a basic logic was designed as shown below: In the conventional XML converter, the functions to process paragraphs (more than 1 line) and words, store a conversion log and extract non-textual contents were developed. To extract non-textual contents, an XML conversion is conducted with control number and report title as input values. Then, the non-textual contents in the report are extracted. After getting control number and report title, the first XML conversion is carried out[15–17]. The first XML convertor's flow is shown in Table 2 and Figure 9.

Table 2. First XML conversion

1.Extract non-textual content objects in a JPEG format and store them in repository.
2. Extract the object caption up to the line 2 and compose indexing and tagging information.
3. Add XML information in which chapter, section and paragraph information is included and use it in searching the report.
4. Add tagging information (chapter/section/paragraph, page location information, etc.) to the XML information and use it in the detailed search of reports.
5. Add page to all page-unchecked XML chapter, section and paragraph.
6. Add object path information to conventional XML and page tag-less XML, revise schema and add sources.
7. Test from the start chapter, section and paragraph to the next chapter, section and paragraph in <p> tag.
8. In case of two or more caption lines, extract and store indexing information.

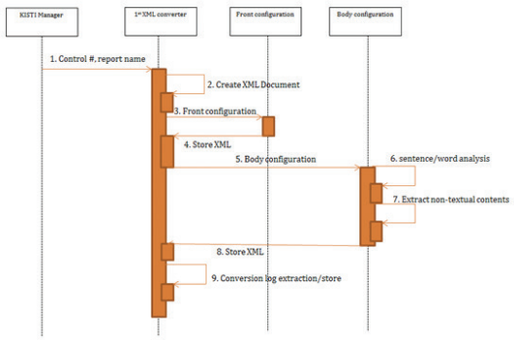

Figure 9. Operation of first XML converter.

The basic flow of the second XML converter is shown in Table 3 and Figure 10. Then, the target files should be those which aren't stored in a PDF/A format.

Table 3. Second XML conversion

1. Get the control number of report title.
2. Create an XML document.
3. Read metadata from the database.
4. Configure the front.
- Configure an XML document from the database.
5. Compose the body.
- Review all lines in all pages.
- In case of table/figure patterns, create related XML categories.
- In case of chapters and sections, create related XML categories (in case of chapters and sections, they are received up to the 3rd stage).
- In case of text patterns, create related XML categories.
- Through analysis on paragraphs and words, create XML based on chapters and sections. - Create images on non-textual contents based on caption on tables and figures.
6. Store a conversion log.
7. Store the created XML in the database.

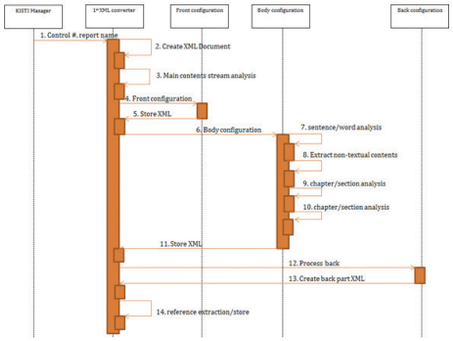

Figure 10. Operation of second XML converter.

3.2 Definition of the Functions of Non-Textual Content Processor

The function-definition documents designed to extract non-textual contents are defined in Table 4. To extract non-textual contents, the database manager performs an XML conversion as follows:

1. The database manager marks (selects) the report for XML conversion and automatic extraction of metadata on the XML conversion list.
2. Click the XML batch conversion button.
3. Once the batch conversion is carried out, the metadata extraction selection page appears. Then, open a PDF file and select the automatic metadata extraction information

Table 4. Definition of functions of non-textual contents

Category	Function
Paragraph/ word analysis	Search of paragraph/section patterns
	Creation of chapter/section reference XML
Management of conversion log	Extraction of conversion log
	Storage of conversion log
Extraction of non-textual contents	Caption search
	Designation of caption-based image coordinates
	Extraction of images
	Storage of images
	Storage of tagging and indexing information

4. Conclusion and Future Work

To develop high-value added database, this study suggested a development plan for a non-textual content processor which can extract and store tables and figures and provide search services. This plan enables the automatic extraction, storage and search of non-textual contents (tables and figures) which exist in conventional reports. In terms of a processing method, the report contents are converted into an XML file, and tables and figures are extracted and stored. In addition, indexing and tagging information can be composed using the caption on the stored non-textual contents and used in information search. If the non-textual content processor proposed in this study regarding the reports on a national R&D program is used, it appears that there would be more opportunity to use high-quality national R&D report database. In addition, by sharing the report contents and metadata with research management institutes, government costs should be further reduced.

5. Acknowledgment

This research was supported by the Sharing and Diffusion of National R&D Outcome funded by the Korea Institute of Science and Technology Information.

6. References

1. Cohen WM and Levinthal DA. Innovation and learning: the two faces of R&D. Econ J. 1989.
2. Final report of national science & technology knowledge information service program. KISTI; 2012.
3. Heo T, Choi G, Park M. Analysis on economic efficiency of national R&D report management system construction program. J of the Soc of Korea Indust and Syst Engin. 2009; 32(2):45–56,.
4. Heo T, Choi G. A study of improvement of national R&D report management system. The Journal 2006 Fall Conf of the Korea Cont Assoc. 2006;10:693–97.
5. Ryu B, Choi G. A study of efficient management of national R&D outcome information and establishment of the distribution system. J Korean Libr Informat Sci Soc. 2003; 37(4):223–40.
6. Evaluation and planning, survey and analysis report of 2009 National R&D Program. Ministry of Education, Science and Technology and Korea Institute of S&T;. 2009.
7. Lee J, Chung D. A study of promotion of report distribution. Journal of the 2nd Acad Conf of 1995 Korea Soc for Informat Manag. 1995; 159–62.
8. Yoon J, Chung Y, Lee H, Lee S. A study of systems designed to promote the distribution of national r&d report information. KISTI; 7th KOSTI 2002 Journal. 2002; 21–43.
9. Kim S, Choi B, Lee M, Kang M. Standardization of work process for distribution of science & technology information. J Korea Cont Assoc. 2007; 7(12):231–7.
10. Heo T, Choi G, Kim J, Park M, Shin Y. Design and construction of registration system for exclusive management of national R&D Report. Korean Institute of Information Scientists and Engineers; Journal of 2009 Korea Computer Congress (KCC). 2009; 36(1(B)):230–5.
11. National R&D Report Registration Management System. KISTI Available from: http://nrms.kisti.re.kr
12. NDSL Research Report. KISTI. Available from: http://www.ndsl.kr
13. Nicola M, John J. XML parsing: a threat to database performance. Proc. 12th Int'l Conf. Information and Knowledge Management (CIKM 03); 2003; ACM Press; p.175–78.
14. Van Lunteren J, et al. XML accelerator engine. Proc. 1st Int'l Workshop High Performance XML Processing; 2004; Available from: www.zurich.ibm.com/~jvl/xml2004.pdf
15. L. Zhao L Bhuyan L. Performance evaluation and acceleration for xml data parsing. Proc. 9th Workshop Computer Architecture Evaluation Using Commercial Workloads (CAECW 06); 2006; Available from: www.cs.ucr.edu/~zhao/paper/caecw06_xml.pdf
16. Pan Y et al. Parallel XML parsing using meta-DFAs. Proc. 3rd IEEE Int'l Conf. e-Science and Grid Computing (e-Science 07); 2007; IEEE CS Press; p. 237–44.
17. Zhang J, Simplify XML processing with VTD-XML. JavaWorld; 2006 Mar 27; Available from: www.javaworld.com/javaworld/jw-03-2006/jw-0327-simplify.html