A Figure Search Engine Architecture for a Chemistry Digital Library

Sagnik Ray Choudhury[†], Suppawong Tuarob[‡], Prasenjit Mitra[†‡], Lior Rokach[*], Andi Kirk[⋆], Silvia Szep[⋆], Donald Pellegrino[⋆], Sue Jones[⋆], C. Lee. Giles[†‡]
[†]Information Sciences and Technology, [‡]Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802 USA
[*]Information Systems Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel 84105
[⋆]The Dow Chemical Company, Spring House, PA 19477 USA
{sagnik, suppawong}@psu.edu, pmitra@ist.psu.edu, liorrk@bgu.ac.il,
{andikirk,sszep,dapellegrino,susanjones}@dow.com,giles@ist.psu.edu

ABSTRACT

Academic papers contain multiple figures representing important findings and experimental results; we present a search engine specifically focused on figures in academic documents. This search engine allows users to search on figures in approximately 150,000 chemistry journal articles though the method is easily extendable to other domains. Our system indexes figure caption and mentions extracted from the PDF in documents using a custom built extractor. Recall and precision performance of extracted figures is in the 80 to 90 % range. We give the frame work for the extraction algorithm, architecture and ranking function.

Categories and Subject Descriptors

H.4 [**Information Extraction and Retrieval**]

Keywords

Information Extraction, Figure Search

1. INTRODUCTION

In general, figures are considered to be an important part of any document; [1] showed that the text in a document was not able to convey the message in 61% of the figures. This motivates search engines specifically focused on indexing and ranking figures in PDF documents. Currently, only three digital library search engines (Yale Image Finder [1], BioText [2], and askHermes[3]) exist with this functionality, all in the biosciences. In most cases academic document repositories have only the document PDF files, from which figures and metadata will have to be extracted. However, with these systems, there has been no discussion of the process of figure and metadata extraction.

Our contributions are: 1. an automatic extractor for figures and their associated metadata (caption, mentions) from documents and 2. a scalable Solr/Lucene[4] based figure metadata search engine with a modified generic Lucene ranking function to improve the quality of search results.

2. RELATED WORK

Recent work [4] describes a methodology for extraction of images and captions from PDF files, whereby images are extracted from PDF using Xpdf[5] and captions are extracted using regular expressions and heuristics. We use regular expressions and document layout information for the same task (section 4). Previously, similar techniques were used in identifying, extracting and indexing tables[3], algorithms and acknowledged entities [2]. Our system is an continuation of this line of work and, more importantly, can be readily integrated with other search features.

3. SYSTEM DESCRIPTION

Given a PDF document, our system automatically finds and extracts related text content, figures, and figure metadata from the document using a custom built extractor, using PDFBox library[6]. For each extracted figure, a metadata file is created, which contains figure caption, figure mention and other information. Figure caption and mention are indexed as separate fields using Lucene[7]. Users can search on both caption and mention fields separately. The SERP provides a ranked list of all figures found.

4. EXTRACTION AND RANKING

Extraction: The extraction process is summarized in algorithm 1. First, a figure and corresponding text identifier (such as, "Fig. 1") is extracted together. In most cases the caption of the extracted figure starts with the identifier. As such, a paragraph starting with the figure id f_i is a probable caption for the figure f_i. However, all such paragraphs are not the actual caption and need to be filtered out. Also,

[1] http://krauthammerlab.med.yale.edu/imagefinder/
[2] http://biosearch.berkeley.edu/
[3] http://figuresearch.askhermes.org

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
ACM 978-1-4503-2077-1/13/07.

[4] http://lucene.apache.org/solr
[5] http://www.foolabs.com/xpdf/
[6] http://pdfbox.apache.org/
[7] http://lucene.apache.org/

determining the paragraph boundary in PDF documents is a non trivial task. Therefore, we reduce the problem to two subproblems - identifying the beginning and ending line of a caption. A set of document layout information based filters are used for these tasks: 1. **Line_length(i,j)**: Returns **true** if length of line i > length of line j by a threshold, else returns **false**. 2. **FontSize_change(i,j)**: Returns **true** if average font sizes of line i and j are different, else returns **false** and 3. **Bold_font(i)**: Returns **true** if a character in the line i is written in bold font, else returns **false**. These filters are defined based on a set of simple observations: 1. font size changes take place in the caption boundary; 2. the starting line of a caption usually has a bold font character (such as **Fig. 1**); and 3. the ending line of a caption is usually shorter than the next line.

```
Data: A PDF file of a document.
Result: Figures and associated captions and mentions
        extracted from the PDF.
From a document d, extract all text lines that contain the
term “figure” or “fig” in a list $L_{rv}$;
for Each line in the whole text of the document do
    Store length of the line, font size, font weight in a list) ;
end
for Each figure $f_i$ in the document $d$ do
    Extract figure $f_i$ using PDFBox ;
    Extract the text in a rectangle below the figure $f_i$ ;
    From the extracted text, find out the id $fid_i$ ;
    if no id is extracted then
        Output the image file of the figure ;
        break;
    end
    else
        Caption=Extract_Caption(id);
        Mention=Extract_Mention(id);
        if caption is null then
            caption=Mention[1]
        end
        Output image file for the figure, metadata file for the
        caption and mention ;
    end
end
```

Algorithm 1: Algorithm for extracting figures, associated caption and mentions from an article.

Ranking: As figures in documents have rich textual metadata, we used textual features such as term frequency and inverse document frequency for ranking. Instead of the whole text of a document, only caption and mentions of a figure are indexed in our system. This way, less weight is assigned to the terms which are common in the whole document, but uncommon in the figure description. The caption field is given an index time boost of five times.

5. EXPERIMENTS AND RESULTS

Extraction: The evaluation parameters for the extraction process were defined as: 1. **Figure-caption recall**: { Number of (figure, caption) pairs extracted } / { Total number of (figure, caption) pairs present in the dataset}. 2. **Caption precision** : Ratio of retrieved captions that were correct. and 3. **Hard accuracy test**: Whether the percentage of match between the extracted and original caption was $\geq 95\%$.

For scalability we ran our extractor on 150,000 chemistry documents to extract more than 90,000 figures. The accuracy evaluation of the system was done on a sample of 150 non scanned documents, containing 883 figures. We extracted figures and associated metadata using two systems: 1. Our PDFBox based extractor (E_{pdbx}) and 2. Xpdf based extractor (E_{xpdf}) reported in [4]. Results in table 1 shows that our system performs well in precision and recall.

Comparison parameter	E_{pdbx}	E_{xpdf}[4]
Figure-caption recall	0.84	0.82
Caption precision	0.95	0.90
CE hard accuracy	0.91	0.92

Table 1: Comparative results of extraction efficiency of our system (E_{pdbx}) with [4] (E_{xpdf}).

Ranking: Three chemistry graduates evaluated the quality of the search results. Average precision at 10, 20 and 30 was found to be 0.7, 0.6 and 0.4, respectively, over a range of nine queries.

6. CONCLUSION AND FUTURE WORK

We report an extraction process of figure and associated metadata from chemistry PDF documents and a novel figure search engine on the extracted metadata. Currently, very few digital libraries allow this functionality. Future work would be to improve the extraction algorithm using linguistic features and to evaluate the scalability of the extraction process on millions of documents. Effective ranking of figures in documents remains an open question, which demands further investigation.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge partial support from the National Science Foundation and suggestions by Madian Khabsa and Hung-Hsuan Chen.

8. REFERENCES

[1] S. Carberry, S. Elzer, and S. Demir. Information graphics: an untapped resource for digital libraries. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’06, pages 581–588, New York, NY, USA, 2006. ACM.

[2] M. Khabsa, P. Treeratpituk, and C. L. Giles. Ackseer: a repository and search engine for automatically extracted acknowledgments from digital libraries. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL ’12, pages 185–194, New York, NY, USA, 2012. ACM.

[3] Y. Liu, K. Bai, P. Mitra, and C. L. Giles. Tableseer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 91–100. ACM, 2007.

[4] L. Lopez, J. Yu, C. Arighi, H. Huang, H. Shatkay, and C. Wu. An automatic system for extracting figures and captions in biomedical pdf documents. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 578–581. IEEE, 2011.