Indexing the medical open access literature for textual and content-based visual retrieval

Ivan Eggel[a], Henning Müller[ab]

[a] Business Information Systems, University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland
[b] Service of Medical Informatics, Geneva University Hospitals and University of Geneva, Switzerland

Abstract

Over the past few years an increasing amount of scientific journals have been created in an open access format. Particularly in the medical field the number of openly accessible journals is enormous making a wide body of knowledge available for analysis and retrieval. Part of the trend towards open access publications can be linked to funding bodies such as the NIH[1] (National Institutes of Health) and the Swiss National Science Foundation (SNF[2]) requiring funded projects to make all articles of funded research available publicly.

This article describes an approach to make part of the knowledge of open access journals available for retrieval including the textual information but also the images contained in the articles. For this goal all articles of 24 journals related to medical informatics and medical imaging were crawled from the web pages of BioMed Central. Text and images of the PDF (Portable Document Format) files were indexed separately and a web-based retrieval interface allows for searching via keyword queries or by visual similarity queries. Starting point for a visual similarity query can be an image on the local hard disk that is uploaded or any image found via the textual search. Search for similar documents is also possible.

Keywords:

Visual information retrieval, open access literature, medical information analysis, information retrieval

Introduction

Images play an increasingly important role in medical practice. They are used in a large variety of contexts such as screening, diagnosis or treatment planning and also with an increasing variety of modalities and radiology protocols. Through the electronic patient record images are not only accessible to radiologists but for all clinicians, and so new tools providing aid particularly for less experienced clinicians in interpreting images seem necessary [1].

Many open access publishers have become available on the Internet, also through initiatives by the NIH (National Institute of Health) and the SNF (Swiss National Science Foundation) to oblige researchers to make articles of publicly funded research available. BioMed Central[3] is surely the most well known publisher but others such as BenthamOpen[4] or Hindawi[5] also provide open access publishing possibilities. Most (medical) articles containing text and images are provided in the form of a complex format such as PDF (Portable Document Format), although some are also accessible in HTML (Hyper Text Markup Language) format. Large journals often provide several thousand articles in a linked hierarchy on their web pages with bibliographic information and abstracts very often being available directly on the HTML pages. Web pages of journals allow for text search in the articles but a particular image search, or search for articles containing similar images is very often not possible. Textual search for images is also provided by ImageFinder[6] and Biosearch[7] but currently no visual search is possible. Such a visual data access has shown to well complement textual search [2].

To obtain all articles of a particular journal, simple web crawlers can be developed to parse the HTML web pages and then follow the links to the full PDF versions of the articles that can then be analyzed further. To obtain articles from BioMed Central's journals it was necessary to crawl their website to obtain meta information on articles, download the corresponding PDF files and extract their texts and images separately for indexing. In a next step all obtained information was indexed using the Lucene[8] text retrieval engine and the GIFT[9] (GNU Image Finding Tool) visual retrieval system [3].

Information retrieval (IR) has traditionally rather concentrated on textual information and a large number of text retrieval systems exist [4]. Image retrieval started much later and then concentrated on text close to the images searched or manual annotation of images themselves [5]. The next step was (visual) content-based image retrieval that relied solely on visual characteristics in the images for retrieval [6], leading to other problems such as the gap between the simple visual features used and the high-level semantics a user is normally searching for. In the medical domain, visual retrieval was proposed several times [7,8] but real clinical applications are scarce [9]. Access to the medical literature was also proposed in [10]. It has become increasingly clear that neither visual nor textual retrieval can solve all the problems alone. Rather, a combina-

[1] http://www.nih.gov/
[2] http://www.snf.ch/
[3] http://www.biomedcentral.com/
[4] http://www.bentham.org/open/
[5] http://www.hindawi.com/
[6] http://krauthammerlab.med.yale.edu/imagefinder/
[7] http://biosearch.berkeley.edu/
[8] http://lucene.apache.org/
[9] http://www.gnu.org/software/gift/

tion of media is required to optimize performance of IR systems [2,11] in the medical field.

This paper describes an approach for multimodal (text and images) medical IR using open source tools limiting the time required for development and also the costs. For the parsing of web pages, the extraction of images and text from the PDFs, as well as for indexing textual and visual information, existing tools were reused. A web interface using JavaServer Faces (JSF), Javascript, and AJAX (Asychronous Javascript and XML) allows for an easy use for the final interface.

The next section describes the materials and methods used for this project. Then, the results of the web crawling as well as the indexation step are given with a description of the user interface. The article finishes with a critical discussion.

Materials and Methods

Data used

The data used for the system described in this article consists of articles from 24 journals (217 are available in total) from the online open access publisher BioMed Central. The chosen journals were in the fields of medical informatics and medical imaging. Each journal contains between 16 and 2500 scientific articles in PDF format, as BioMed Central is still a very young publisher. All PDF documents were publicly accessible and are free of charge. Information taken from the articles was of textual and visual (images) nature. The mentioned journals were crawled on August 20-21, 2009.

Table 1- Overview of the amount of data indexed.

Measure	Value
Number of journals	24
Number of articles	9403
Min. number of articles per journal	16
Max. number of articles per journal	2495
AVG number of articles per journal	392
Total number of images (after a cleaning step)	37940
Min. number of images per journal	28
Max. number of images per journal	13618
Average number of images per journal	567
Min. number of images per article	0
Max. number of images per article	1659
Average number of images per article	4
Size of all images total	2.81 GB
Average image size	77 KB

Textual information was parsed in HTML format from the description of each article on BioMed Central's webpage consisting of the title, abstract, journal name, publication date, author names and the URL of PDF documents. Extracted information of the PDF documents consists of the entire text in the document and the contained images. The average number of extracted images per article was around 4 (with a total of 9403 indexed articles and 37940 extracted images). The minimum number of images per journal was 28 with a maximum of 13618 and an average of 567. The total size of all extracted images was about 2.8 GB. An overview of the data is given in Table 1.

Technologies used

Goal of the project presented in this article was to reuse well-established existing tools to limit the development time. For text retrieval, the open source Java library Lucene was used, which is easy to integrate and adapt to a variety of scenarios. With Lucene providing the possibility to index more than one field per document it allowed searching in several data fields such as text content, author name, and article title. Several other options Lucene offers were not used in the first prototype described in this paper.

For visual retrieval, the GIFT was chosen that has equally been in use for almost ten years and that has shown to deliver stable visual research results.

The separation of images and text from PDF documents was performed using Apache PDFBox[10]. To parse basic meta information (on the web page) of each article NekoHTML[11] was used. NekoHTML is an open source Java HTML scanner and tag balancer library that enables developers to parse HTML documents and access the information using standard XML (Extensible Markup Language) interfaces. As application server Glassfish v2.1 was used. We relied on Java and JSF for the integration. Other Technologies used on the client side were pure Javascript and AJAX.

For the work described in this article it was not necessary to start from scratch since a previous system for the separation of text and images had already been developed [3]. The current system is an extended version of the previous system with an additional web crawler and changes in the user interface.

The server was a rack server with two Intel Xeon Dual Core 1.6 GHz processors with 2 GB of RAM and total disk space of 244 GB in a RAID array.

Results

This section describes the main results concerning mainly the implementation of our prototype with its search interfaces.

System setup

The work described had six main goals and for all of them existing tools could be combined with an ergonomic web interface that was developed in this project (Figure 1):

[10] http://incubator.apache.org/pdfbox/
[11] http://sourceforge.net/projects/nekohtml/

(1) crawling and parsing HTML pages on BioMed Central and writing the collected information into XML files (one file per journal with information on each article in a journal),

(2) downloading all PDF documents (web addresses of each article's PDF document saved in the XML files),

(3) extracting images and free text from the PDF documents,

(4) indexing the meta information of the articles (saved in the XML file) and of information directly extracted from the PDF documents with Lucene,

(5) indexing the extracted images with GIFT, and

(6) combining the extraction and the retrieval systems in a single interface based on JSF.

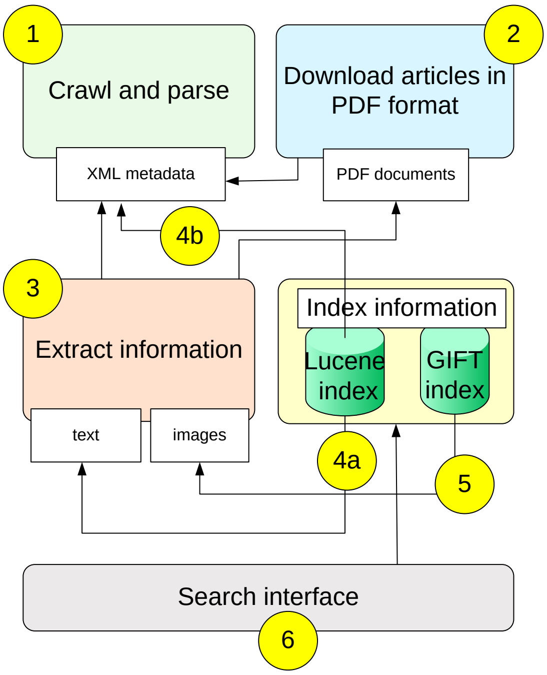


Figure 1 – The six main goals of work in a diagram with the basic system architecture.

Crawling and indexing the dataset

To crawl and parse the pages of BioMed Central, a trivial crawler was developed as a Java console application. The application takes the journal's starting URL and the directory to save the generated XML files as arguments. Starting the application results in the crawler finding all articles of the journal following well-defined links, parse the necessary information and save it to an XML file.

A simple download tool was also developed in Java being able to download all corresponding PDF files of a journal's articles from the URLs stored in the XML file. With all PDF documents downloaded, the extraction and indexation can take place. For this task we use a web interface that lists all XML files in the server's directory of journals. By choosing one of the listed files the application parses the content of a single XML file and each article's information is processed. First, the article's PDF document is located and subsequently the extraction of text and images takes place. Images are stored directly on the hard disk also generating thumbnails at the same time. The next step adds all textual information of the article to the Lucene index. One single step indexes/stores the following information on an article:

- auto generated id,
- title,
- abstract background,
- abstract methods,
- abstract results,
- abstract conclusions,
- online Article URL (if available),
- PDF document URL,
- authors,
- journal name,
- publication date,
- text content,
- image names,
- homepage URL.

The average duration of extracting and indexing one article was about 1 second. After indexing one article, the next article's information is extracted and indexed. This step is repeated until all the XML files are processed. Subsequently, images are indexed with GIFT, taking about 6 hours (around 0.5 seconds per image) with a total of 38’000 images.

One problem regarding the images was the existence of TIFF (Tagged Image File Format) images inside the PDF documents since TIFF is not supported directly by web browsers. We solved this problem by converting each TIFF image automatically to a PNG (Portable Network Graphics) image with the aid of Java Advanced Imaging (JAI). This task is performed directly after extracting a TIFF so before saving the picture and before adding information to the index.

After indexing several journals it was obvious that many articles contained the same or similar BioMed Central logos, which can be considered as irrelevant. To avoid storing the logos, the system discarded all images that were equal to a small set of example images selected manually (comparison of binary data). Another problem in our first extraction phase was a large number of small logos and graphical elements in the articles resulting in over 200’000 images to index. To avoid this problem, only images higher and larger than 32 pixels were considered for indexation, which reduced massively the amount of data to index visually.

It was also found that it was not always possible to extract all images from a PDF automatically. The reasons for this have not been totally investigated but it can be due to the production of the PDF that can be formed in a way not understood by our extraction module PDFBox. A very small number of mirror-inverted images also occurred during the extraction phase due to PDFBox, and we have not found a way to avoid this. However, these problems are expected to be fixed since PDFBox is improved and extended regularly.

Another challenge was parsing the HTML documents on BioMed Central. Although each page had a uniform design and layout we found several differences concerning the position of certain HTML tags, which forced us to take all possible positions of an HTML element into account and not to rely on the initial structure.

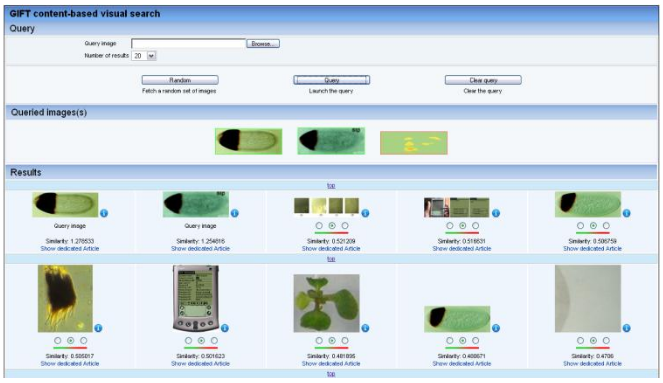


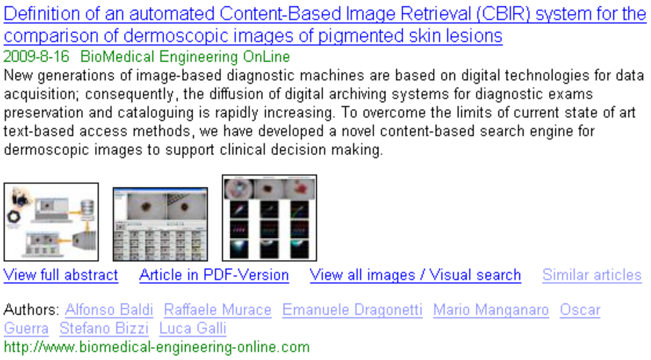
Figure 2 - Screenshot of a purely visual search

Query options

The query interface can be found for testing at MedSearch[12]. A query start is typically a keyword or a new image that can be uploaded to start a visual query. It is possible to switch between a visual and textual query interface and it is always possible to show all images of a particular article (if any images were extracted) together on a single page or to show a link to the full text PDF of this article. After displaying images contained in an article it is possible to launch a visual search by clicking the button "Similar images" below an image. After this start, positive and negative relevance feedback can be used to refine the visual query further. All found images are marked neutral at the start and can be selected as relevant (green) or irrelevant (red) as shown in Figure 2. Another way to start a visual search is to obtain random images of the image database and perform queries by relevance feedback afterwards. It is then possible for the user to select "Show dedicated article" below an image. This will cause the system to switch to the textual search interface and to automatically phrase a query searching for the article the image is contained in. In the same interface the user can phrase his own textual queries with the client afterwards performing a HTTP-GET-request to the server passing relevant query parameters. This can facilitate integration of the system into other web pages. The user may also specify the number of results to show. Once queried, the results screen shows details of the retrieved articles consisting of the following elements (see Figure 3):

- article title with link,
- publication date,
- journal name,
- the first 500 characters of the abstract,
- link to full abstract,
- download link for article in PDF format,
- link to view all images and possibly to launch a visual search,
- search for similar articles based on the text (article search based on similar images not implemented yet),
- link to author names, which causes the system to search for all articles by the same author,
- web address of original article homepage,
- thumbnails of first three images in article (if available); it can be configured to show more images.

The entire retrieval system was installed on a server of the University of Geneva, Switzerland. However it is possible to integrate both into a totally distributed system, as Lucene, the query interface, and GIFT are independent components.



Definition of an automated Content-Based Image Retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions
2009-8-16 BioMedical Engineering OnLine
New generations of image-based diagnostic machines are based on digital technologies for data acquisition; consequently, the diffusion of digital archiving systems for diagnostic exams preservation and cataloguing is rapidly increasing. To overcome the limits of current state of art text-based access methods, we have developed a novel content-based search engine for dermoscopic images to support clinical decision making.

View full abstract Article in PDF-Version View all images / Visual search Similar articles

Authors: Alfonso Baldi Raffaele Murace Emanuele Dragonetti Mario Manganaro Oscar Guerra Stefano Bizzi Luca Galli
http://www.biomedical-engineering-online.com

Figure 3 – Screenshot of a single textual search result including images of the found article.

No evaluation of retrieval quality is given in this text as no ground truth of BioMed Central is available. Both GIFT and Lucene have been evaluated on retrieval of journal articles and images in the ImageCLEF competition [2].

Speed measurements

Lucene has been used in many large-scale projects and search times with single key words are in the order of a few milliseconds. In our case it averages 40 milliseconds with more than 9'000 indexed articles, leaving room for much larger databases. Visual similarity search using GIFT was around 0.5 seconds for single image queries using our database with about 38'000 images. This allows for fast querying and good usability. When using new images to query, the feature extraction takes another 0.5 seconds. In total, results are usually shown in a second.

Conclusions

This article presents a solution for visual and textual IR from the open access publisher BioMed Central's online journals. Meta information of articles was available in HTML format, and the articles themselves were crawled directly in PDF-Format. Crawling, parsing and extracting text and images separately were essential parts. For all implemented tasks open source components could be used. The goal of the system is to collect information on scientific medical articles and to make this search interface available publicly to medical students and clinicians. A combined visual and textual search interface should optimize the reuse of all existing knowledge including text and visual parts. Currently, only 24 of the over 200 available journals on BioMed Central were indexed. Considering this and also the several other publishers of open access articles, the system is extendable to include a much larger body of knowledge. The resulting system was achieved by developing

[12] http://medgift.unige.ch:8080/MedSearch/faces/Search.jsp

a loosely coupled architecture that stores information in XML files after crawling and parsing. The component-based architecture allows the extension of the system simply by replacing the crawler or adapting it to a new structure of journal web pages. All web-components can be integrated into a distributed environment easily. Besides open access publishers there is also an increasingly large number of articles available in full text by traditional publishers, very often 6-12 months after the original publication data.

The current system only indexed some of BioMed Central's journals. A next step is to index all 217 journals with their images. With data collected of the current system we can estimate a total size of all images of ~25 GB and a total size of the Lucene index of 4.8 GB. The estimated time to index all journals would be ~18 hours, crawling and downloading time excluded. Including the crawling and downloading this should still remain less than two days.

Of course there is not only a single open access publisher on medical journals on the Internet. The next step is to consider other open access publishers such as Bentham Open or Hindawi into our index. Indexing all abstracts of PubMed[13] would be another step further. As the full-text articles are linked if available, a semi-automatic crawling could be developed for this. Still, particularities of single publishers would then need to be taken into account.

Another important aspect is to keep the system up to date as new articles are published very regularly. An automatic update function has so far not been included into our system. The system would have to crawl (e.g. every week) the publisher's sites detect new articles, then download and index them.

Regarding the query options it would be good to add mixed visual and textual into the system, so search based on visual and textual characteristics combined, for example to find similar documents. A visual search could also be limited to articles containing a particular keyword.

The current implementation does not use all functions of Lucene, yet. However, it allows for an easy expansion of the functionality. Language detection of the documents to be indexed can be integrated to allow for a multilingual indexing and retrieval. Currently, this is of little importance as the published open access articles are mainly in English. Still, when adding medical texts from clinical routine in a country like Switzerland, there are documents in several languages.

The implemented system responds to the needs of crawling, parsing and separating BioMed Central's articles into visual and textual components for an efficient visual, textual and combined search. The entire system is based on open source components and can easily be reproduced.

Acknowledgements

This work was partially supported by the BeMeVIS project of the University of Applied Sciences Western Switzerland (HES-SO).

References

[1] Haux R. Health information systems - past, present, future. Int J Med Inform. 75(3-4):268-81.

[2] Müller H, Kalpathy-Cramer J, Kahn CE Jr., Hatt W, Bedrick S, Hersh W. Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task, Springer Lecture Notes in Computer Science 5706, 500-510, 2009.

[3] Eggel I, Müller H. Combination of Visual and Textual Similarity Retrieval from Medical Documents. MIE 2009, 841-845,

[4] Salton G, Buckley C. Term weighting approaches in automatic text retrieval. Information Processing and Management 24, (1988), 513-523.

[5] Enser P. Visual image retrieval: seeking the alliance of concept-based and content-based paradigms, Journal of Information Science, 2000.

[6] Smeulders AVM, Worring, Santini S, Gupta A, Jain R. Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22(12), pages 1349-1380, 2000.

[7] Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inform. 2004 Feb ;73(1):1-23.

[8] Lowe HJ, Antipov I, Hersh W, Smith CA. Towards knowledge-based retrieval of medical images. The role of semantic indexing, image content representation and knowledge-based retrieval. Proc AMIA Symp. 1998 ;882-6.

[9] Aisen AM, Broderick LS, Winer-Muram H, Brodley CE, Kak AC, Pavlopoulou C, et al. Automated Storage and Retrieval of Thin-Section CT Images to Assist Diagnosis: System Description and Preliminary Assessment. Radiology. 2003 Jul 1;228(1):265-270.

[10] Deserno TM, Antani S, Long RL. Content-based image retrieval for scientific literature access. Methods Inf Med 2009; 48(4):371-380.

[11] Müller H, Kalpathy-Cramer J. Analyzing the content out of context – features in medical image retrieval, International Journal on Healthcare Information Systems and Informatics, volume 4 number 1, pages 88-98, 2009.

Address for correspondence

Prof. Dr. Henning Müller
Business Information Systems
University of Applied Sciences Western Switzerland
TechnoArk 3
3960 Sierre, Switzerland
tel ++41 27 606 9036
fax ++41 27 606 9000
henning.mueller@hevs.ch

[13] http://www.ncbi.nlm.nih.gov/pubmed/