Combination of visual similarity retrieval & textual retrieval from medical documents

Ivan EGGEL[b], Henning MÜLLER[ab]
[a]*Medical Informatics Service, University & Hospitals of Geneva, Geneva, Switzerland*
[b]*Business Information Systems, University of Applied Sciences, Sierre, Switzerland*

Abstract. Medical visual information retrieval has been an active research area over the past ten years as an increasing amount of images is produced digitally and has become available in patient records, scientific literature, and other medical documents. Most visual retrieval systems concentrate on images, only, but it has become apparent that the retrieval of similar images alone is of limited interest and, and rather the retrieval of similar documents is an important domain. Most medical institutions as well as the World Health Organization (WHO) produce many complex documents. Searching them, including a visual search, can help finding important information and also facilitates the reuse of document content and images. The work described here is based on a proposal of the WHO that produces large amounts of documents from studies but also for training. The majority of these documents are in complex formats such as PDF, Microsoft Word, Excel, or PowerPoint. Goal is to create an information retrieval system that allows easy addition of documents and search by keywords and visual. For text retrieval, Lucene is used and for image retrieval the GNU Image Finding Tool. A Web 2.0 interface allows for an easy upload as well as simple searching.

Keywords. Content-based medical image retrieval, multimodal information search

1. Introduction

Medical images play an important role in diagnosis, research, and teaching. They are used in a variety of contexts. Images are rarely useful without context information, though, and for teaching or research purposes most images are embedded in complex formats such as PDF (Portable Document Format), Word, Excel, or PowerPoint. The work described in this paper relies on a concrete proposition of the World Health Organization (WHO) consisting of small, independent entities sharing a common infrastructure. Images are an important part of documents in research, results communication, and also for preparing teaching materials for various countries in several languages. Images and much of the material itself could be reused if the data were managed better and images could be found easily in the haystack of data. Thus the idea was to extract images and text from documents and then allow for a textual search as well as for a visual search in a single and combined web-based interface.
Information retrieval (IR) has traditionally concentrated on textual information and a large number of systems exist [1]. Image retrieval started much later and then concentrated on text close to the images searched or manual annotation of images themselves [2]. The next step was (visual) content-based image retrieval that relied solely on visual characteristics in the images for retrieval [3], leading to other problems

such as the gap between simple visual features used and the high-level semantics a user issearching for. In the medical domain, visual retrieval was proposed several times [4,5] but real clinical applications are scarce [6]. It has also become increasingly clear that neither visual nor textual retrieval can solve all the problems alone. Rather, a combination of media is required to optimize performance of IR systems [7,8].
This article describes an approach for multimodal (text and images) medical IR using open source tools limiting the time required for development and also costs. For the extraction of images and text from complex document formats existing tools were reused. An interface using JSF (JavaServer Faces) and AJAX allow for an easy use.

2. Methods

The data used for this article consisted of CDs of teaching material from the WHO in a variety of formats. Another test was run with articles made available in the context of the ImageCLEF[1] competition 2008 (Part of CLEF, the Cross Language Evaluation Forum) consisting of 67'000 medical images from several thousand scientific articles (of the journals Radiology and Radiographics). The techniques were all used and evaluated in the competition itself and thus a repetition of the evaluation is avoided but can be found in [9]. A few results are explained to have a clearer idea of the performance compared to other research IR systems.
Goal of the presented project was to reuse well established existing tools. For text retrieval Lucene[2] was used that is easy to integrate and adapt to a variety of scenarios [10]. Lucene also allows searching in more than one field. Documents can be searched by free text but also by author name. Many other options of Lucene were not used in the first prototype described in this paper. For visual retrieval the GIFT[3] (GNU Image Finding Tool) was chosen that has equally been in use for almost ten years and that has shown to deliver stable visual research results. Another important part was the availability of Apache APIs (Application Programming Interfaces) to extract visual and textual information separately from Microsoft Office and PDF documents. Other libraries exist for the XML-based formats of OpenOffice as well. POI[4] (Poor Obfuscation Implementation) was used for the extraction from Office documents and PDFBox[5] for the extraction of images and text from PDF. As application server Glassfish was used. We relied on Java and JSF for the integration.

3. Results

The work described had three main goals and for all of them existing tools could be combined with an ergonomic user interface that was developed: (1) extraction of images and free text from complex documents, (2) indexation of the document text with Lucene and of the images with GIFT, and (3) combination of the extraction and the retrieval systems in a single interface based on JSF and AJAX.

[1] http://www.imageclef.org/
[2] http://lucene.apache.org/
[3] http://www.gnu.org/software/gift/
[4] http://poi.apache.org/
[5] http://www.pdfbox.org/

Figure 1: Overview of the implemented system components for the extraction of images and text from the complex documents and than the retrieval of the indexed documents.

Figure 1 shows the entry page of the user interface of the implemented and combined components. The upload of complex documents allows submitting four types of documents: PDF, Microsoft Word, Excel, PowerPoint. These documents can be uploaded as single files. With the POI system the Microsoft office documents are separated into text and images and with PDFBox the PDF documents are treated. The documents currently extract the document title, the author, the full text, and the images separately. The text, author and title are stored directly in a Lucene index. A zip file with several documents can be uploaded directly and all documents in the archive are added to the index. Another upload possibility is using a URL. All file types including ZIP can be downloaded to the server using a URL.To avoid having extremely small images indexed in the dataset the minimal width and height of images have to be 16 pixels. Smaller images are not stored an indexed. The system offers several text search options, all based on Lucene. Text search is possible for free text in full text, document titles, and by author. Stop words are automatically removed from the retrieval. An English stop word list is currently used for this but stop word removal does exist for other languages as well.

Figure 2: Screenshot of a purely visual search.

The query start usually has to be a keyword or a new image to upload. It is possible to switch between visual and textual query interface and it is always possible to show all images of a particular document together on a single. Figure 2 shows a

visual similarity search using GIFT. In other search interfaces the images contain a little button “show me similar images”, starting a visual search. After displaying the contained images of a document it is possible to start a visual similarity search by clicking one of the “similar images” button. After this start, positive and negative relevance feedback can be used to refine the query. In the interface all images are marked neutral but can be selected as relevant (green) or irrelevant (red) as in Figure 2.

It is then possible to get back to the images of a document, the title and authors as well as a link to the original document file, so all the data are really integrated. This is the link "Document details" in Figure 2. The interface permits the use of a URL to submit an image or a direct upload form a local disk. The number of results to be shown on screen can also be configured here.

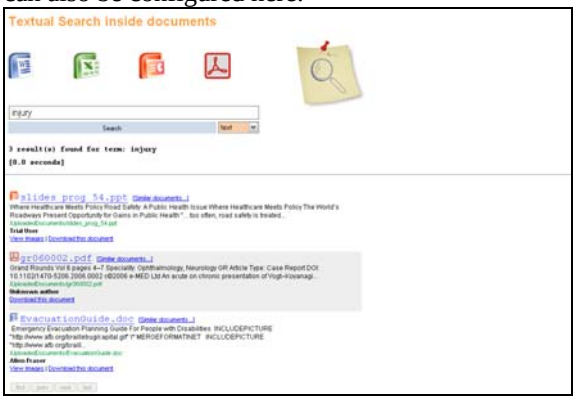

Figure 3: Screenshot of a search with keywords in the database.

Figure 3 shows a result of a textual search. The first 200 characters of each document are shown as well as title and author (optional). With a single click, all images of a document can be shown and with another click visually similar images can be searched. The search for documents with similar text is also possible. Documents can be stored locally or there URL to the original location can be kept. The system was implemented in a distributed environment. The interface including a web server and Lucene were working on a virtual machine image (with MS Windows and 1 GB of RAM assigned) at the University of Applied Sciences Western Switzerland in Sierre, whereas GIFT was installed on a sever of the University of Geneva, Switzerland. Lucene has been used in large scale projects and search times with single key words are in the order of milliseconds for several thousand documents, leaving room for larger databases. Visual similarity search using GIFT was 0.5 seconds for single image queries using the ImageCLEF database with 67’000 images. This allows for fast querying and good usability. When using new images to query, the feature extraction needs takes another 0.5 seconds. The dataset of the WHO does not contain topics and relevance judgments. Using the database of ImageCLEF, both the GIFT and the Lucene system have been evaluated. GIFT is among the average of purely visual systems but the only open source system [10], with a Mean Average Precision (MAP) of 0.025. Lucene has obtained almost the best performance for purely textual retrieval (MAP of 0.27) and had the highest early precision [11]. These good performances and the fact that the tools are open source made them the technology of our choice.

4. Conclusions and future work

This article presents a solution for visual and textual IR from collections of complex medical documents. An extraction of text and images separately is an integral part. All components are open source. Main goal of the system is the reuse of knowledge stored in existing documents in complex formats in institutions, and this goal could be reached. The current system has an easy-to-use interface and allows for an easy integration into existing environments. All components are web-based and distributed.

The current system does not use all functions of Lucene, yet. Its architecture allows for an easy expansion of the functionality, though. One of the next is language detection of the indexed documents to allow for a multilingual indexing and retrieval important for a country such as Switzerland with four official languages. Another simple extension is the indexation of html documents in the same way as the current complex document. This would allow for a direct download and indexation of web pages in the same context as other complex documents. Two slightly more complicated changes are the mix of visual and textual retrieval in the same query step ("images similar to an example and containing the word tuberculosis in the text") and the comparison of entire documents for similarity ranking, including the visual components. All in all, the implemented system has responded to the criteria of separating complex documents into visual and textual components and allowing for an efficient. The system is based on open source components and can easily be reproduced by others.

5. Acknowledgements

This work was supported by the RCSO BeMeVIS project. We would also like to thank Irma Velazquez of the WHO for her input concerning the project.

References

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

[2] PGB Enser, Image Databases for Multimedia Projects. *J Am Society for Information Science* **46**(1), (1995), 60-64.

[3] AWM Smeulders, M Worring, S Santini, A Gupta, R Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Trans on Pattern Anal Mach Intell* **22,** (2000), 1349-1380.

[4] H Müller, N Michoux, D Bandon, A Geissbuhler, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, *Int J Med Inform* **73,** (2004), 1-23.

[5] TM Lehmann, MO Güld, C Thies, B Fischer, K Spitzer, D Keysers, H Ney, M Kohnen, H Schubert, BB Wein, Content-based image retrieval in medical applications, *Methods Inf Med* **43** (2004), 354-361.

[6] AM Aisen, LS Broderick, H Winer-Muram, CE Brodley, AC Kak, C Pavlopoulou, J Dy, CR Shyu, A Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, *Radiology* **228,** (2003), 265-270.

[7] W Hersh, H Müller, J Jensen, J Yang, P Gorman, P Ruch, Advancing biomedical image retrieval: development and analysis of a test collection, *J Am Med Inform Assoc* **13,** (2006), 488-496.

[8] RK Srihari, Z Zhang, A Rao, Intelligent Indexing and Semantic Retrieval of Multimodal Documents, *Information Retrieval*, **2**(2/3), (2000), 245-275.

[9] H Müller, J Kalpathy-Cramer, CE Kahn Jr., W Hatt, S Bedrick, W Hersh, Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task, *Springer Lecture Notes in Computer Science*, (2009) – to appear.

[10] O Gospodnetic, E Hatcher, Lucene in Action, *Manning Publications*, Greenwich, (2005).

[11] J Kalpathy-Cramer, S Bedrick, W Hatt, W Hersh, Multimodal Medical Image Retrieval: OHSU at ImageCLEF 2008, *Springer Lecture Notes in Computer Science*, (2009) – to appear.