Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-841

Combination of Visual and Textual Similarity Retrieval from Medical Documents

Ivan EGGEL [a], Henning MÜLLER [a,b,1]
[a] *Business Information Systems, University of Applied Sciences, Sierre, Switzerland*
[b] *Medical Informatics Service, Hospitals & University of Geneva, Geneva, Switzerland*

Abstract. Medical visual information retrieval has been an active research area over the past ten years as an increasing amount of images are produced digitally and have 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 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 in this paper 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 content. For text retrieval, Lucene is used and for image retrieval the GNU Image Finding Tool (GIFT). 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 many contexts. Images are rarely useful without contextual information, though. For teaching or research purposes many images are embedded in complex formats such as PDF (Portable Document Format), Word, Excel, or PowerPoint. The work described in this paper is based on a proposition of the WHO towards us. Images are important at the WHO for documents in research, dissemination, and for preparing teaching material for various countries and languages. Images and text could be reused if the data were managed well and if images could be found easily in the haystack of data. Thus the idea is to extract images and text from documents and then allow for a textual search as well as for a visual search in a single combined web interface. As this study is rather on teaching material and complex document formats, images are in standard image formats (jpeg, gif, or png) and not in the medical DICOM format that yields a higher quality and DICOM header information but does not allow for an easy integration into office documents. An important standard for teaching files is MIRC (Medical Imaging Resource Center) but it is not a focus of this article that concentrated on office formats.

[1] Corresponding Author: Henning Müller, HES SO, TechnoArk 3, 3960 Sierre, Switzerland; Tel.: ++41 27 606 9036; E-mail: henning.mueller@hevs.ch.

Information retrieval (IR) has traditionally been focused on textual information and a large number of systems exist [1]. Image retrieval started much later and first used text close to the images searched or manual annotation of the images themselves [2]. The next step was (visual) content-based image retrieval that relied solely on visual characteristics of 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 is searching 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 used. An interface using JSF (Java Server Faces) and AJAX creates a good usability.

2. Methods

The data used for this article consists of five CDs of teaching material from the WHO in a variety of formats (as the evaluation is qualitative, any set of documents can be used for such a test). Another test was run with articles made available in the context of ImageCLEF[2] 2008 (part of CLEF, the Cross Language Evaluation Forum) consisting of 67,000 medical images from several thousand scientific articles (journals Radiology and Radiographics) including full text articles and image captions. The techniques were used and evaluated in the competition itself and thus results are only referenced here [9]. A few results are explained to compare the presented work to other systems. Goal of the presented project was to reuse well established existing tools. For text retrieval, Lucene[3] was used that is easy to integrate and adapt to various scenarios [10]. 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[4] (GNU Image Finding Tool) was chosen that has equally been in use for almost ten years and that has shown to provide 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 but are not integrated in our project at the moment. POI[5] (Poor Obfuscation Implementation) was used for the extraction from Office documents and PDFBox[6] 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

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

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.

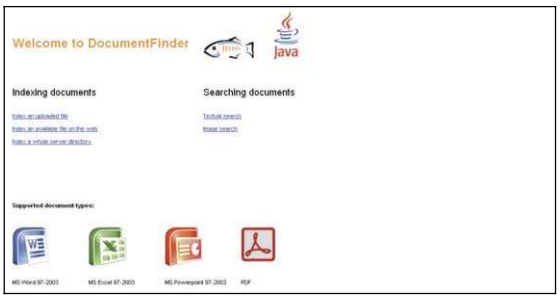

Figure 1. Overview of the implemented system components for the extraction of images and text from the complex documents and than the retrieval of the indexed documents

Figure 1 shows the entry page of the user interface. The upload of documents allows submitting four types: PDF, Microsoft Word, Excel, PowerPoint. These documents can be uploaded as single files. With the POI system, Microsoft office documents are separated into text and images. With PDFBox the PDF documents are treated. The documents currently extract the document title, the author, the full text, and the images separately. The full text, author, and title are stored directly in a Lucene index.

A zip file containing 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 uploaded to the server using a URL. To avoid having extremely small images indexed in the dataset (such as small arrows on lines) the minimal width and height of images have to be 16 pixels. Smaller images are not stored and indexed. The system offers several text search options, all based on Lucene. Text search is possible for the full free text, document titles, and by author. Stop words (very frequently occurring words) are automatically removed from the retrieval. An English stop word list is used but stop word removal does exist for other languages as well.

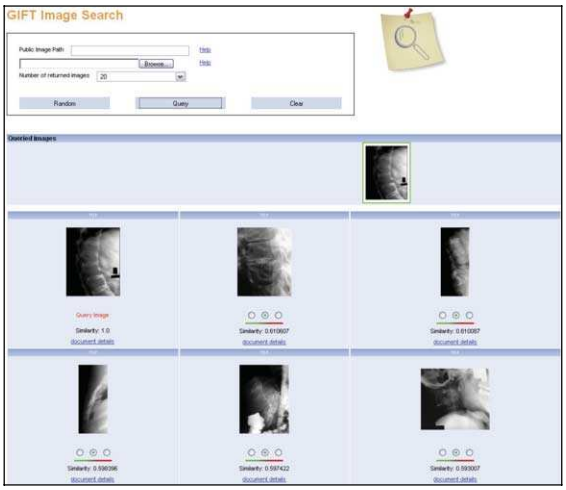

Figure 2. Screenshot of a purely visual similarity search

The query start 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 result page. Figure 2 shows a visual similarity search using GIFT. After displaying the images contained in a document it is possible to start a visual similarity search by clicking on one of the "similar images" buttons. After this start, positive and negative relevance feedback can be used to refine the query. In the interface all images are marked as neutral but can be selected as relevant (green) or irrelevant (red) as can be seen in Figure 2.

It is then possible to get back to the entire set of the images of a document to show the author or to open the original file, so all the data are really integrated. The interface also permits the use of a URL to submit an image or a direct upload from a local disk. The number of results to be shown on screen can be configured in the interface.

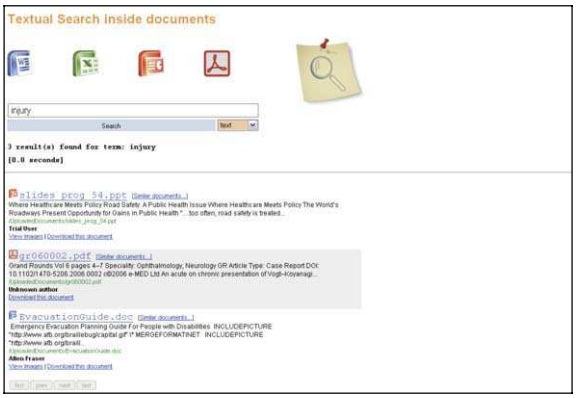

Figure 3. Screenshot of a search with keywords in the database

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 the URL to the original location can be kept. The system was fully 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. The GIFT was installed on a server 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 takes another 0.5 seconds. The dataset of the WHO does not contain topics and relevance judgments thus no quantitative evaluation can be given. Both the GIFT and the Lucene system were evaluated using the ImageCLEFmed 2008 database. 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 documents. An extraction for text and images separately is an integral part of this work. All components are open source. Main goal of the system is the reuse of knowledge stored in existing documents in various formats in institutions. This goal was 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 expansion steps is language detection of indexed documents to allow for a multilingual 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 documents. 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.

In the end, the implemented system has responded to the criteria of separating complex documents into visual and textual components and allowing for an efficient search. The system is based on open source components and can easily be reproduced by others.

Acknowledgements. This work was partly supported by the RCSO BeMeVIS project and the EU project KnowARC. We would also like to thank Irma Velazquez of the WHO for her input.

References

[1] Salton, G., Buckley, C. (1988) Term weighting approaches in automatic text retrieval. *Information Processing and Management* 24:513–523.
[2] Enser, P.G.B. (1995) Image databases for multimedia projects. *Journal of the American Society for Information Science* 46(1):60–64.
[3] Smeulders, A.W.M., Worring, M., Santini, S. et al. (2000) Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:1349–1380.
[4] Müller, H. et al. (2004) A review of content-based image retrieval systems in medicine – Clinical benefits and future directions. *International Journal of Medical Informatics* 73:1–23.
[5] Lehmann, T.M., Güld, M.O., Thies, C. et al. (2004) Content-based image retrieval in medical applications. *Methods of Information in Medicine* 43:354–361.
[6] Aisen, A.M. et al. (2003) Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment. *Radiology* 228:265–270.
[7] Hersh, W., Müller, H., Jensen, J. et al. (2006) Advancing biomedical image retrieval: Development and analysis of a test collection. *Journal of the American Medical Informatics Association* 13:488–496.
[8] Srihari, R.K., Zhang, Z., Rao, A. (2000) Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval* 2(2-3):245–275.
[9] Müller, H., Kalpathy-Cramer, J., Kahn Jr., C.E., Hatt, W., Bedrick, S., Hersh, W. (2009) Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task. In *Proceedings of Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum CLEF 2008, Springer LNCS*, in press.
[10] Gospodnetic, O., Hatcher, E. (2005) *Lucene in Action*. Manning Publications, Greenwich.
[11] Kalpathy-Cramer, J. et al. (2009) Multimodal Medical Image Retrieval: OHSU at ImageCLEF 2008. In *Proceedings of Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum CLEF 2008, Springer LNCS*, in press.