Int J CARS (2009) 4:71–77
DOI 10.1007/s11548-008-0269-7

ORIGINAL ARTICLE

Indexing and retrieving DICOM data in disperse and unstructured archives

Carlos Costa · Filipe Freitas · Marco Pereira · Augusto Silva · José L. Oliveira

Received: 10 January 2008 / Accepted: 23 September 2008 / Published online: 28 October 2008

Abstract
Objective This paper proposes an indexing and retrieval solution to gather information from distributed DICOM documents by allowing searches and access to the virtual data repository using a Google-like process.
Methods and materials The medical imaging modalities are becoming more powerful and less expensive. The result is the proliferation of equipment acquisition by imaging centers, including the small ones. With this dispersion of data, it is not easy to take advantage of all the information that can be retrieved from these studies. Furthermore, many of these small centers do not have large enough requirements to justify the acquisition of a traditional PACS.
Results A peer-to-peer PACS platform to index and query DICOM files over a set of distributed repositories that are logically viewed as a single federated unit. The solution is based on a public domain document-indexing engine and extends traditional PACS query and retrieval mechanisms.
Conclusion This proposal deals well with complex searching requirements, from a single desktop environment to distributed scenarios. The solution performance and robustness were demonstrated in trials. The characteristics of presented PACS platform make it particularly important for small institutions, including educational and research groups.

Keywords PACS · DICOM · Information retrieving · Search engine

C. Costa (✉) · F. Freitas · M. Pereira · A. Silva · J. L. Oliveira
University of Aveiro, DETI/IEETA,
3810-193 Aveiro, Portugal
e-mail: carlos.costa@ua.pt

Introduction

Over the last decade, the use of digital medical imaging systems in healthcare institutions has significantly increased, and they constitute valuable tools supporting the medical profession, both in decision support and in treatment procedures. Research and industry efforts to develop medical imaging equipment, which evolved gradually to a grid of networked imaging resources, have been the major driving forces towards the wide acceptance of the picture archiving and communication system (PACS) concept.

PACS empowers healthcare practitioners with the capability to remotely access multimedia patient information and to setup telemedicine, telework and collaborative work environments [1]. Currently, there are PACS solutions with different architectures and services, from simple models, typically used in small laboratories, to enterprise-wide platforms, mostly used in large hospital networks. The PACS concept encompasses several technologies that include hardware and software for acquisition, distribution, storage and analysis of digital images in distributed environments [2].

The digital imaging and communications in medicine (DICOM) standard architecture was a major contribution to the exchange of structured medical imaging data [3]. Currently, almost all medical imaging equipment manufacturers provide embedded DICOM (Version 3) digital output in their products. As a result, large volumes of DICOM data have been produced in the last few years, creating enormous sets of clinical data that, in most of the cases, have been stored in local archives without remote indexing and retrieval facilities.

In this paper, we present a novel approach to handle this data that is based on an indexing and retrieval software service that can be easily installed in any DICOM server or in any computer that stores DICOM files.

Materials

Typically, the core element of PACS is a central server that stores the images and the database that contains complementary information about patients and studies (Fig. 1). This system implements the *DICOM Storage Service* that allows any imaging equipment to directly send the acquired images to the centralized PACS archive. The access to the stored images is then supported by the *Query and Retrieve Service* [4] or *Web Access to DICOM Persistent Objects* (WADO) [5]. DICOM Version 3 is currently a well-established standard in the medical field, and its characteristics facilitate the interoperability between modality equipments and information systems (Fig. 1).

PACS were originally conceived to store the huge amount of images that are generated in a hospital, and the searching mechanisms were rather small. They were specifically chosen to allow a medical specialist to retrieve images based, for instance, on patient name or on patient ID. However, the spread of the PACS through the internet has created unforeseen scenarios for the use of this technology.

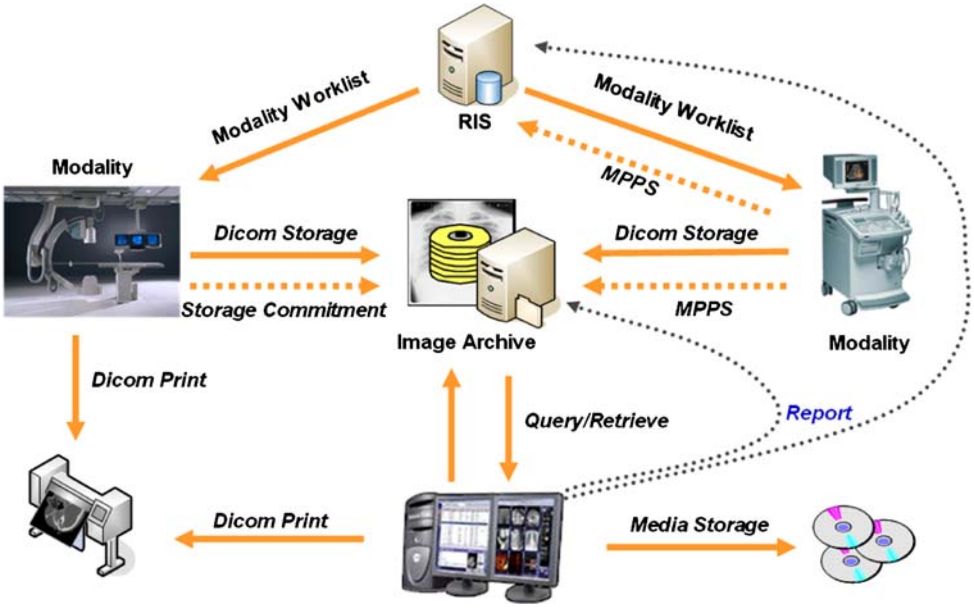
Query and retrieval

A typical DICOM persistent object contains the pixel image data together with several descriptive information blocks. This metadata includes attributes like the image modality, equipment reference, acquisition parameters, image resolution, measurements or clinical trial study data. Moreover, the DICOM structured reports (SR) object supports both conventional free-text reports and structured information.
The PACS may gather all this information (image, descriptive information, report) that can be retrieved later by authorized users. However, because the archive only typically contains a small number of fields to support the *DICOM Information Model* Query/Retrieve (DIM Q/R) model, a great portion of this information is not searchable. Also, with traditional PACS, we do not have a content-based retrieval system available.

Small-sized and distributed environments

Medical imaging modalities are successively becoming more powerful and less expensive. The result is the proliferation of equipment acquisition by imaging centers, even in small ones. A good example is the new generation of ultrasound equipment that is extremely compact and has high-performance imaging capabilities. Typically, these small centers do not have large enough requirements to justify the acquisition of a traditional PACS. They usually have a single reviewing workstation with the capacity to store a few months of exams or, at most, a free small PACS solution. As a major consequence, any healthcare institution, even with limited human or financial resources, is currently able to generate and collect their own medical imaging data [6]. Nevertheless, expensive computational tools and human skills are usually concentrated in a small number of specialized medical centers. In those centers, properly trained professionals are able to provide the correct diagnosis to the patient's physician [7–9]. To provide the technical infrastructure to this workflow, besides the network, we can envision a set of methodologies such as PACS services, simple DICOM data transfer using file servers (ftp, http) or email, or specific distributed applications supported by Web Services or similar programming interfaces. Independent of the transfer mode, any DICOM image can be finally uploaded in the central PACS

Fig. 1 PACS DICOM services and processes workflow



archive using a private connection or the DICOM Storage service.

Research and educational PACS

Besides clinical practices, it is common to find research and educational institutions that possess a huge volume of disperse DICOM images. These images are documents from different clinical areas, pathologies and image modalities, but all are of great diagnostic interest. These images are anonymous and can be shared with other groups to help enrich the global knowledge about a particular pathology. However, the lack of remote access and content-based query mechanisms make the process impractical.

In a personal computer, one can also have a large set of images, but the installation of a PACS in every computer is not a cost-effective solution. This is because this would involve non-trivial installation procedures and increase the system maintenance overhead. The problem felt, sometimes, by professionals is that "they never find that interesting study when it is needed".

Methods

During the last decade, we have been developing engineering solutions to support medical imaging transmission in critical scenarios [10–12]. One of the results of this work was a web-enabled PACS to support cardiac imaging modalities [11,12], where the main challenge is the handling of dynamic imaging modalities (films), such as cardiac ultrasound (US) and X-ray angiography (XA). Although the system is being used in several hospitals, we realize that for specific scenarios, such as the ones discussed previously, there are still a large set of DICOM files that are outside PACS management.

With the aim to create a simple peer-to-peer PACS solution, we have developed Dicoogle, an indexing and retrieving framework for distributed DICOM resources. This system is based on a public domain document indexing engine, which implies reduced installation and maintenance costs.

Dicoogle

Associated to every PACS infrastructure, there is a database engine to support the DICOM Information Model—real world (DIM) [13] or, at least, a "dicomdir" structured file containing information related to patients, studies, series and images. For instance, when a *DICOM Storage SCP* receives an exam from an equipment modality (Fig. 1), one needs to only store the images in the file system and to update the PACS database with elements extracted from the study.

The main idea of Dicoogle was the replacement or the extension of the traditional database by an indexing and retrieving engine. With this solution, since one can index any document type besides the storage and searching of DIM fields, it is possible to add all other DICOM data elements (text-based) without the need to create new fields, new tables, and new relations that would be necessary in the database supported approach. Moreover, since Dicoogle is mainly a storage device, the PACS can use it to take advantage of its free text search features to retrieve images over distributed repositories.

Dicoogle is based on Apache Lucene [14,15], a Java search library that is used in a large variety of applications, including Wikipedia. Index servers' technology is very popular in digital libraries and information retrieval systems. However, in the medical imaging field, its usage has been sparing. For instance, in [16,17], this solution was used for classification and annotation to improve medical imaging context-based retrieval.

Data interfaces

The developed solution has two main input interfaces to DICOM data. The first is a DICOM Storage SCP service to receive data objects from imaging equipment or other DICOM nodes. The second interface is based on a file system monitor and on a document indexing service. In this last interface, all events (file creation, change, deletion, etc.) are intercepted, and when coming from DICOM files, specific indexing actions will be triggered. This process can be subject to several configuration rules. For instance, one can specify particular computers, disks, file systems, paths, files and subsets of DICOM SOP classes.

Indexing schemas

Dicoogle explores the two different types of indexing modes supported by Lucene: full text and metadata. The full-text (or content based) index mode maintains for each word a list of the files containing that word and a statistical measure of how important that word seems to the document (TF-IDF [18]).

In the hierarchical DIM Q/R context, the most important mode is the metadata index because it allows the association of a DICOM file with a keyword/value pair. For example, the "Test.dcm" file can have a key "PatientName" with the value "Demo Patient". This way, the metadata index is able to process metadata queries and efficiently return the files that satisfy each query. In Dicoogle, the metadata keywords used are the DIM Q/R mandatory fields (Patient ID, Patient Name, Study Instance UID, Study ID, Study Date, Study Time, Accession Number, etc.) [4]. All DICOM metadata is parsed and indexed according with predefined rules.

Finally, thumbnail images extracted from DICOM files are also stored in the index server as no-indexed fields.

Query and retrieval mechanism

The query and retrieval requirements follow different users' needs. In general, users need not have a complete knowledge of the structure of the DICOM images to retrieve the wanted information. However, skilled users can fine tune search sentences to improve the retrieved results.

Dicoogle supports different types of queries. It is possible to query by keywords that must occur in the context of specific structural parts of the DICOM standard. The current Dicoogle version supports queries in DIM Q/R fields, but this can be easily extended to support the lowest level of the DICOM structure, i.e. the attributes. For instance, it could be possible to query for keywords in a specific DICOM tag element. Second, it is possible to query content only, as in a common information retrieval system. Here, we can make the analogy with a search engine. The matching of keywords can be performed through exact matching or using approximate terms. Using the two previous search types, one can build mixed queries where content and the DICOM structure are combined.

The new paradigm introduced with the Dicoogle index engine is that we can execute queries over a set of distributed DICOM repositories, which are logically indexed as a single federated unit. This lets us run queries against all Dicoogle nodes as though they were a single index. Users can enter search criteria once and access several search engines simultaneously in real time.

The query result includes an image preview using the stored thumbnails. To retrieve the whole study, one can select the image or the series, and a DICOM viewer will be opened to present the transferred data. The remote retrieval of DICOM image objects is performed using the DICOM Q/R Service.

Results

The medical informatics scenarios empowered by index engines are promising, and they can even change the traditional PACS usage paradigm. Dicoogle supports two main scenarios (Fig. 2):

- Traditional PACS Archive Unit;
- Personal computers with unstructured archives of DICOM files.

In both usage scenarios, Dicoogle can receive data via a DICOM Storage SCP service or scanning a file system volume for DICOM persistent objects.

Dicoogle works in two phases. The first time when the application is executed, it is necessary to scan all the files in the computer. During this process, all detected DICOM files are indexed, much as it happens with those files received by Storage SCP. Once the complete scan is finished, the application turns into a "steady state" mode (file system watch), where it is only necessary to detect incremental file changes

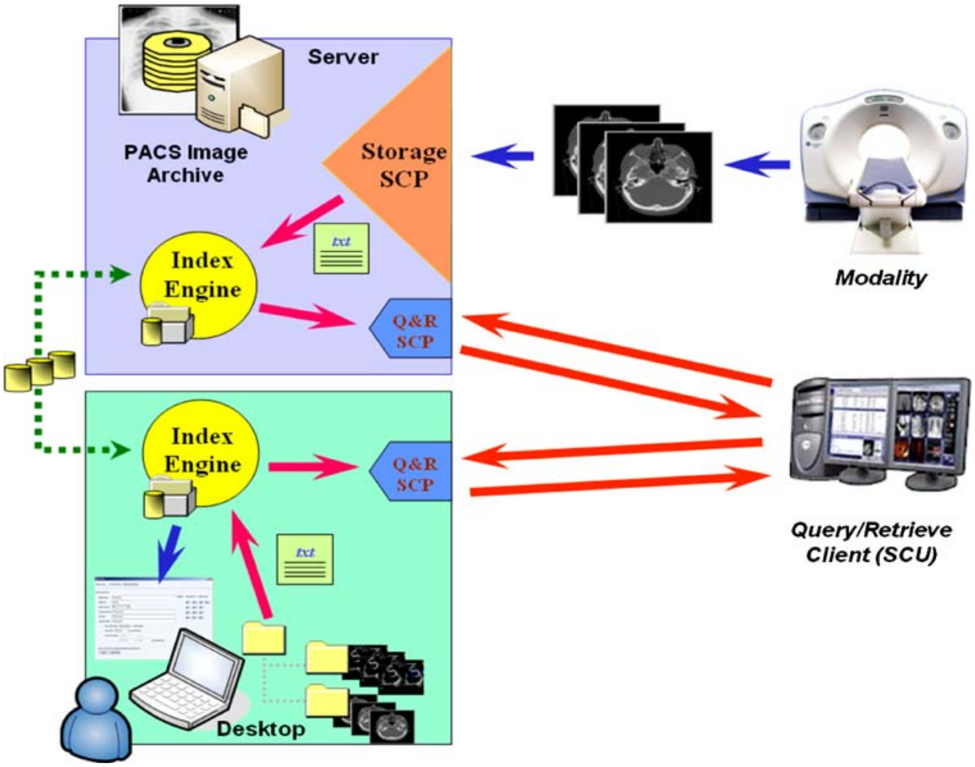


Fig. 2 Dicoogle scenario

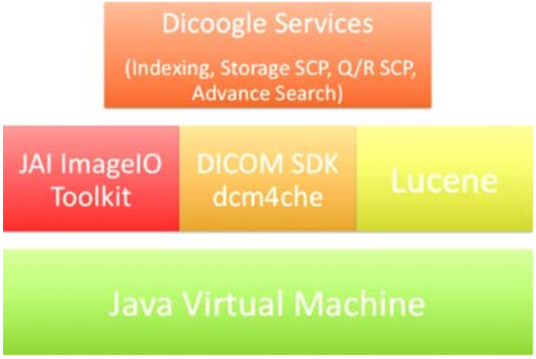


Fig. 3 Dicoogle software components

(create, update and delete). Dicoogle only needs to monitor directory and files changes using the operating system call back functions that notify the user directory changes. Using Java native interface (JNI) to wrap these native function calls, every time a file is changed an event is generated, and Dicoogle updates the metadata index.

On the other hand, Dicoogle can also work as a traditional PACS solution. In this case, we just need to configure an external DICOM viewer to visualize the selected images. The simplicity of this solution fits well with the requirements of a small imaging center. In an enterprise PACS, Dicoogle can also be used as an image storage archive without affecting the existent third services, for instance, the modality worklist, the visualization software or the web portal interface (Fig. 1).

Implemented modules

The developed software solution is support by open source tools that run in any common operating system (Windows, Linux, Mac OS). The implementation of DICOM standard functionalities is supported by the dcm4che library [19,20] (Fig. 3), a SDK that is used to extract DICOM data elements from persistent objects and to implement the Storage SCP and Query/Retrieve SCP services [4]. The decoded DICOM information is parsed and indexed by a Lucene server according to programmed rules. All incoming connections are logged and visible in the Dicoogle monitor window (Fig. 4).

The integration with existent information consumers, i.e. the client reviewing workstations (Fig. 2), is issued by a Query and Retrieve SCP service. To implement this service, the indexing of all mandatory Q&R DIM fields (Patient Name, Patient ID, Study Instance UID, Study Date, Study Time, Accession Number, Study ID, Series Instance UID, Series Number, Modality, SOP Instance UID and Instance Number) was crucial [4].

User interface

The Dicoogle user interface allows patients, studies, series and images in each system to be found. There are two kinds of searches available: “Free text” and “PACS mode”.

Using the “Free text” mode (Fig. 5), the user inserts the query text and the search is made in all the indexed fields. Like in web search engines, it is also possible to use Boolean operators to refine the query. Additionally, there are several other functionalities available, such as nearby terms and flexible range search:

- *OperatorName:“John Doe”* to search all studies released by a specific Clinical Operator;
- *PatientName:“Marie Pires”*∼10 to search in the Patient Name field the terms “Marie” and “Pires” separated by up to ten words;
- *StudyDate:*[*20020101 to 20030101*] to find studies with date fields values between 2002.01.01 and 2003.01.01;
- *PatientName:*{*Ana to Carmen*} to find all studies whose PatientName are between Ana and Carmen, but not includes Ana and Carmen.

Alternative to the free text search form, Dicoogle provides a graphical interface with a typical PACS view (Fig. 6).

The DICOM viewer that is used by Dicoogle is an independent application that can be replaced with any similar solution. In the example shown in Fig. 5, the Himage Viewer is used [11,21].

Performance measurements

The performance and robustness of Dicoogle was tested using a data set of 114 studies containing 23,477 DICOM files, representing 30.4 GB (1.4 MB per file). In the samples, we have different imaging modalities, but the images are predominantly Cardiac XA, US and CT. To evaluate time response, we studied four different configurations:

1. indexing only DIM Q/R fields;
2. indexing DIM Q/R fields plus document content;
3. indexing DIM Q/R plus thumbnails (64 × 64 matrix) inclusion;
4. indexing DIM Q/R fields plus content and thumbnails (64 × 64 matrix);

In each configuration, the indexing time, index file size and search time was measured. The results obtained are shown in Table 1.

The previous results effectively demonstrate that a document-indexing solution, such as Dicoogle, is very efficient in terms of time. This indexing step is naturally slower

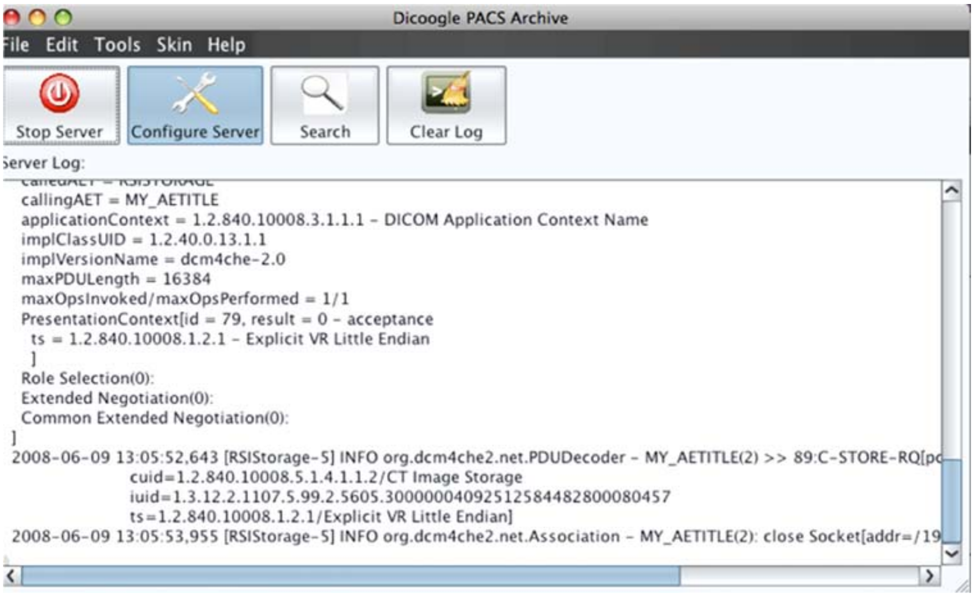


Fig. 4 Dicoogle main interface (Mac OS)

Server Log:

```
  calledAET = RSISTORAGE
  callingAET = MY_AETITLE
  applicationContext = 1.2.840.10008.3.1.1.1 - DICOM Application Context Name
  implClassUID = 1.2.40.0.13.1.1
  implVersionName = dcm4che-2.0
  maxPDULength = 16384
  maxOpsInvoked/maxOpsPerformed = 1/1
  PresentationContext[id = 79, result = 0 - acceptance
    ts = 1.2.840.10008.1.2.1 - Explicit VR Little Endian
    ]
  Role Selection(0):
  Extended Negotiation(0):
  Common Extended Negotiation(0):
]
2008-06-09 13:05:52,643 [RSIStorage-5] INFO org.dcm4che2.net.PDUDecoder - MY_AETITLE(2) >> 89:C-STORE-RQ[pc
                cuid=1.2.840.10008.5.1.4.1.1.2/CT Image Storage
                iuid=1.3.12.2.1107.5.99.2.5605.30000004092512584482800080457
                ts=1.2.840.10008.1.2.1/Explicit VR Little Endian]
2008-06-09 13:05:53,955 [RSIStorage-5] INFO org.dcm4che2.net.Association - MY_AETITLE(2): close Socket[addr=/19
```

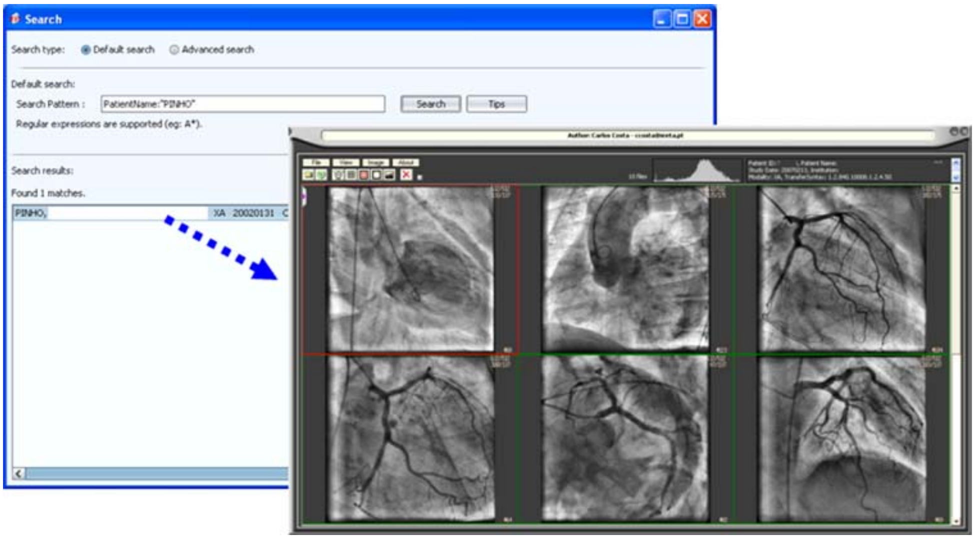


Fig. 5 Free text search (Windows)

Search type: ● Default search ○ Advanced search

Default search:

Search Pattern : PatientName:"PINHO"

Regular expressions are supported (eg: A*).

Search results:

Found 1 matches.

PINHO, XA 20020131 C

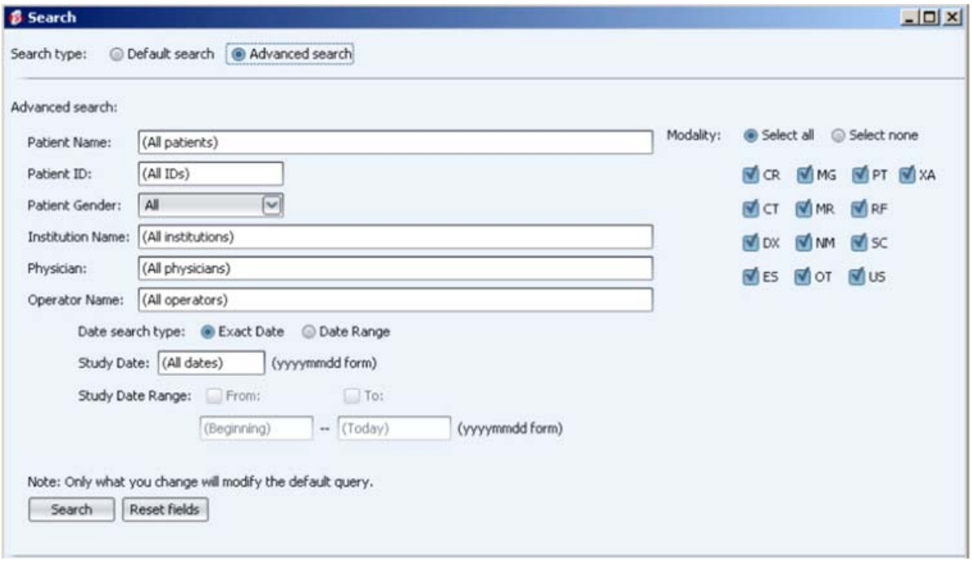


Fig. 6 Traditional search interface (Windows)

Search type: ○ Default search ● Advanced search

Advanced search:

Patient Name: (All patients)

Patient ID: (All IDs)

Patient Gender: All

Institution Name: (All institutions)

Physician: (All physicians)

Operator Name: (All operators)

Modality: ● Select all ○ Select none

☑ CR ☑ MG ☑ PT ☑ XA
☑ CT ☑ MR ☑ RF
☑ DX ☑ NM ☑ SC
☑ ES ☑ OT ☑ US

Date search type: ● Exact Date ○ Date Range

Study Date: (All dates) (yyyymmdd form)

Study Date Range: □ From: □ To:

(Beginning) -- (Today) (yyyymmdd form)

Note: Only what you change will modify the default query.

Table 1 Dicoogle quantitative measurements

Case study	Indexing time (s)	Index volume	Search time (ms)
1	1.369	7.720 KB	120
2	2.403	400 MB	230
3	2.707	49.770 KB	120
4	3.743	442 MB	230

during the setup phase (first index of the corpus), but after that, any increase in the corpus size has a small impact on the index time, due to the increment index mechanism. Moreover, the time consumed during this phase is mainly associated with the DICOM file decoding. As expected, the full contents index also increases the time of processing by approximately 72%.

The generation of an index introduces a new overhead in the DICOM corpus as far as disk size is concerned, but that is negligible compared with the original volume of data (0.4–30.4 GB). Finally, the addition of thumbnails increases, slightly, the indexing phase time (due to image resizing and JPEG encoding) and the disk size, but it has no effect in the search times.

Conclusions

The commerce-driven World Wide Web (WWW) pushed practically all PACS suppliers to develop client applications where clinical practitioners can send or receive images using conventional personal computers. One of the most important advantages of digital imaging systems is to allow the widespread sharing and remote access of medical data between healthcare institutions.

This paper presents a new PACS storage and retrieval approach based on document indexing that can replace or extend the traditional PACS DBMS. This solution deals well with complex searching requirements, from a single desktop environment to distributed scenarios.

The Dicoogle model appears as an open source software solution that can be easily installed both in a PC and in a central server. The indexing schema is completely configurable and allows a user to search for terms that are normally inaccessible using the database query approach. It can also be applied to existing image archives, to enrich the queries that are performed within the archive.

References

1. Costa C, Silva A (2007) Oliveira JL Current perspectives on PACS and cardiology case study. In: Vaidya S, Jain LC, Yoshida H (eds) Studies in computational intelligence: advanced computational intelligence paradigms in healthcare, chap 5. Springer, Berlin, pp 79–108
2. Huang HK (2004) PACS and imaging informatics: basic principles and applications. Wiley, New York
3. Oosterwijk H (2005) Dicom basics, 3rd edn. OTech, Aubrey
4. DICOM-P4 (2007) Digital Imaging and Communications in Medicine (DICOM), Part 4: Service Class Specifications. National Electrical Manufacturers Association
5. DICOM-P18 (2004) Digital Imaging and Communications in Medicine (DICOM), Part 18: Web Access to DICOM Persistent Objects (WADO). National Electrical Manufacturers Association
6. Reiner BI et al (2005) Multi-institutional analysis of computed and direct radiography—Part II. Economic analysis. Radiology 236:420–426. doi:10.1148/radiol.2362040673
7. Bradley WG (2004) Offshore teleradiology. J Am Coll Radiol 1(4):4. doi:10.1016/j.jacr.2003.12.043
8. Larson PA, Janower ML (2005) The Nighthawk: Bird of Paradise or Albatross? J Am Coll Radiol 2(12):3. doi:10.1016/j.jacr.2005.08.002
9. Millard WB (2007) Nighthawks across a flat world: Emergency radiology in the era of globalization. Ann Emerg Med 50(5):545–549. doi:10.1016/j.annemergmed.2007.09.012
10. Silva A et al (1998) A cardiology oriented PACS. In: Proceedings of SPIE: Medical Imaging, San Diego
11. Costa C et al (2004) Himage PACS: a new approach to storage, integration and distribution of cardiologic images. In: PACS and Imaging Informatics—Proceedings of SPIE, San Diego
12. Costa C et al (2007) Enhanced PACS to support demanding telemedicine and telework scenarios. In: CARS 2007—International Congress and Exibition: Computer Assisted Radiology and Surgery, vol 2, pp S322–S323
13. DICOM-P3 (2001) Digital Imaging and Communications in Medicine (DICOM), Part 3: Information Object Definitions. National Electrical Manufacturers Association
14. Apache SF (2007) Lucene Index Server. http://lucene.apache.org
15. Gospodnetic O, Hatcher E (2004) Lucene in Action. Hanning
16. Kalpathy-Cramer J, Hersh W (2007) Automatic image modality based classification and annotation to improve medical image retrieval. Medinfo 12(Pt 2):1334–1338
17. Hersh W, Kalpathy-Cramer J, Jensen J (2007) Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006, in Lecture Notes In Computer Science. Springer, Berlin, pp 660–669
18. Jones KS (1988) A statistical interpretation of term specificity and its application in retrieval. J Doc 28:11–21. doi:10.1108/eb026526
19. dcm4che. Sourceforge project (2007) http://sourceforge.net/projects/dcm4che/
20. Warnock MJ et al (2007) Benefits of Using the DCM4CHE DICOM Archive. J Digital Imaging 20(Suppl 1):125–129
21. Costa CMA et al (2006) A demanding web-based PACS supported by web services technology. In: Cleary KR, Galloway RL Jr (eds) Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display. Proc SPIE 6145:84–92