Future Generation Computer Systems 37 (2014) 243–251

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A text based indexing system for mammographic image retrieval and classification

Alfonso Farruggia, Rosario Magro, Salvatore Vitabile *

Dipartimento di Biopatologia e Biotecnologie Mediche e Forensi, Universitá degli Studi di Palermo, Viale del Vespro, 90127, Palermo, Italy

HIGHLIGHTS

- Text based indexing system for mammographic image retrieval and classification.
- Accurate information extraction from large amount of data.
- Bayesian Naive classifier to improve Search Engine results.
- Web service for mammographic structured reports indexing and related images labeling.
- Medical Decision Support Systems.

ARTICLE INFO

Article history:
Received 15 April 2013
Received in revised form
4 November 2013
Accepted 17 February 2014
Available online 12 March 2014

Keywords:
Information retrieval
Medical documents indexing and classification
Medical images indexing and classification

ABSTRACT

In modern medical systems huge amount of text, words, images and videos are produced and stored in ad hoc databases. Medical community needs to extract precise information from that large amount of data. Currently ICT approaches do not provide a methodology for content-based medical images retrieval and classification. On the other hand, from the Internet of Things (IoT) perspective, the ICT medical data can be produced by several devices. Produced data complies with all Big Data features and constraints. The IoT guidelines put at the center of the system a new smart software to manage and transform Big Data in a new understanding form. This paper describes a text based indexing system for mammographic images retrieval and classification. The system deals with text (structured reports) and images (mammograms) mining and classification in a typical Department of Radiology. DICOM structured reports, containing free text for medical diagnosis, have been analyzed and labeled in order to classify the corresponding mammographic images. Information Retrieval process is based on some text manipulation techniques, such as light semantic analysis, stop-word removing, and light medical natural language processing. The system includes also a Search Engine module, based on a Bayes Naive Classifier. The experimental results provide interesting performance in terms of Specificity and Sensibility. Two more indexes have been computed in order to assess the system robustness: the A_z (Area under ROC Curve) index and the σ_{Az} (A_z standard error) index. The dataset is composed of healthy and pathological DICOM structured reports. Two use case scenarios are presented and described to prove the effectiveness of the proposed approach.

1. Introduction

Information and Communication Technologies (ICT) are changing the way to organize and manage medical diagnosis. Emerging Internet technologies, increasing computational power, and fast pervasive digital communications can be applied to the medical domain for introducing new paradigms in clinical data analysis and management. Virtualization and cloud computing are redesigning both the ICT architectures and the nature of ICT services. From the Internet of Things (IoT) perspective, the ICT medical structures are composed of several data producers. In addition, data produced complies with all Big Data characteristics: high volume, quickly produced and of a different nature. The IoT guidelines put at the center of the system a new smart software entity to manage the data gathered from several sources. In the medical scenario, the goal of this entity is to manage, transform, and represent Big Data in a new understanding form for designing innovative Medical Decision Support Systems. So, physicians could have a smart tool to support their decision processes in analyzing real-time data.

In this work, a software architecture to integrate existent RIS (Radiology Information System) and PACS (Picture Archiving and Communication System) functionalities within a Department of

* Corresponding author. Tel.: +39 091 655 2378; fax: +39 091 655 2325.
E-mail address: salvatore.vitabile@unipa.it (S. Vitabile).

http://dx.doi.org/10.1016/j.future.2014.02.008
0167-739X/

Radiology is presented. The system is composed of two modules: the *Indexing Engine* and the *Search Engine*. The first one allows for collecting, processing and indexing DICOM (Digital Imaging and Communications in Medicine) mammographic structured reports including the free text of a medical report [1]. The second one allows for mammographic images classification and retrieval using the previous results. The DICOM structured report is an electronic document composed of several fields related to a patient, such as the Patient ID, the Accession Number, and the free text medical report produced by the physicians. In this work, the last field is manipulated using the Natural Language Processing (NLP) techniques for indexing and extracting medical knowledge. The considered structured reports, containing the free text medical report, are produced by physicians during the daily workflow and they are in Italian language. The structured reports are collected in the RIS and they can be extracted using unique *Patient ID* and *Accession Number* information. *Accession Number* is also used to access the corresponding mammographic images in the PACS.

As each Machine Learning classification problem, documents classification is related to the parameters estimation of an approximating model [2]. Currently, a precise and exhaustive technique for medical documents classification does not exist. On the other hand, the Naive Bayes classifier [3] is one of the most widely used approaches to classify a text into categories. It provides a good methodology to address the problems from different points of view. Analyzing the DICOM structured report contents, the classifier labels the medical reports as healthy or pathological on the basis of its content.

The processed dataset is composed of real DICOM structured reports, produced by several breast physicians. The training set has been labeled as healthy or pathological from the same expert breast physicians. Information Retrieval techniques, such as light semantic analysis to remove negative terms and stop-words, and a clinician's thesaurus to uniform the used medical report terms have been implemented to improve the classification process results.

From an architectural point of view, the system acts as a Web Service and it can be used in different cases. In particular, the described system is useful as a Medical Decision Support System for medical diagnosis or as a case-based learning system for education. Two use case scenarios have been analyzed to test the effectiveness of the proposed approach. In the first scenario, a physician can require additional information during the diagnosis process by means of selecting similar cases. The physician, through a web page, inserts one or more keywords to extract the useful cases from the indexed databases. As result, the system shows the selected DICOM structured reports and the related mammographic images which best fit the submitted keywords. The functionality can be used to reduce the medical error or validate the initial diagnosis. In the second scenario, the physician exploits the proposed system as a case-based learning tool for students. By inserting one or more keywords about a pathology, the physician is able to show the selected cases through web *teaching files*. In both cases, new DICOM structured reports, created during the daily workflow, can be added to the database, increasing the knowledge base dimension.

The cases studied presented in this work deal with breast structured reports and images. However, the approach and the related processing steps can be applied to different pathologies, reports, and images.

The system has been tested submitting several queries with a different degree of complexity. Specificity and Sensibility indexes have been computed to prove the effectiveness of the proposed approach. Two more indexes have been computed in order to assess system robustness: the A_z (Area under ROC Curve) and the σ_{Az} (A_z standard error) indexes [4]. The first one provides a measure of the classifier capability to separate the healthy and pathological patterns. The second one is a measure of the error in calculating the area under the ROC (Receiver Operating Characteristic) curve.

The remainder of the work is organized as follows. Section 2 presents some literature works on information retrieval of medical documents, as well as some literature works on probabilistic classifiers. Section 3 describes the features of the proposed system. Section 4 presents the technologies used to develop the proposed framework and shows the experimental results with Italian structured reports. Finally, Section 5 contains some concluding remarks and future directions.

2. Related works

Everyday several DICOM structured reports are stored in the IHE (Integrating the Healthcare Enterprise) databases. The medical domain strongly feels the need to share information and to make it easily accessible to other physicians. Available informations for users are huge with a considerable amount of data. Adopting modern medical information systems, data are created directly in electronic form and stored on huge databases containing documents, natural text, and images, such as DICOM (Digital Imaging and COmmunications in Medicine) images, and Structured Reports (SR). So, there is the need, through a smart classification methodology, to provide techniques for retrieving only those images and documents, whose contents meet some search criteria.

In this context, in the literature, there are many research works on Information Analysis, Classification and Retrieval. In [5] it is presented a web based platform for medical cases management. The work has been oriented on multimedia data management and classification, as well as on algorithms for querying, retrieving and processing different medical data types (mainly text and images). The platform develops an intelligent framework to manage medical datasets (text, static or dynamic images), in order to optimize diagnosis and decision processes, reducing medical errors and increasing healthcare quality.

The authors in [6] presented a framework of web services using Bayesian theorem and decision trees to construct a web-services-based decision support system for medical diagnosis and treatment. The process helps physicians enhancing the medical decisions' quality and efficiency. In addition, the diagnosis can be transmitted to a decision-tree-based treatment decision support service component via XML to generate recommendation and analysis for treatment decisions.

In the literature, there are several standardized clinical terminology systems, such as SNOMED CT [7] or Mesh [8]. The first one is an organized collection of medical terms that can be processed by a computer. The project has seen the gradual merging of multiple terminology collections, and today is a large dictionary with more than 344,000 clinical concepts. In [9] it is presented a comprehensive analysis of artificial methods which could be applied to documents encoded by SNOMED CT. MeSH is a huge vocabulary maintained by the U.S. National Library of Medicine (NLM) in order to index articles and the scientific literature of the biomedical field in the bibliographic database MEDLINE/PubMed and the NLM catalog books. MeSH terminology allows for retrieving information even when the scientific material is used in a different period from the period received as input. In [10] it is proposed a work based on MeSH vocabulary. It uses not-Euclidean document distance measure based on MeSH tree structures. The authors quantitatively evaluate the approach against the standard vector space approach and against an hybrid version of both.

The authors of [11] introduce an algorithm for labeled and unlabeled documents learning, based on the combination of Expectation–Maximization (EM) and a Naive Bayes Classifier. In the first step, the algorithm trains a classifier using the available labeled documents. After that, the algorithm probabilistically labels the unlabeled documents. At the end, it trains another classifier using all the documents labeled, and it iterates to system convergence.