Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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.

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008

The EM procedure works fine if the data is complaint to the generative assumptions of the model, but sometimes these assumptions are violated, producing poor performance. The authors present two extensions of the algorithm to improve the classification accuracy, according to two conditions: a weighting factor to modulate the contribution of the unlabeled data; and the use of multiple mixture components per class.

In [12] an automatic text categorization and summarization approach for analyzing the input text structure is presented. The authors presented a text analyzer which is developed to derive the structure of the input text using rule reduction technique in several stages: Token Creation, Feature Identification and Categorization, and Summarization.

The work proposed in [13] studies the problem of building text classifiers using positive and unlabeled examples. The key feature of this problem is that there is no negative example for learning. This work presents a two-steps classifier. In first step a naive Bayesian technique to identify a set of reliable negative documents from the unlabeled set is used. In the second step the naive Bayesian classifier is used to build a set of classifiers by iteratively applying a classification algorithm and then selecting a good classifier from the set.

The authors of [14] describe a method for improving short text strings classification using a labeled training data combination plus a secondary corpus of unlabeled, but related, longer documents. This work shows that such unlabeled background knowledge can greatly decrease error rates if the number of examples or the size of the training strings are small. This is particularly useful when labeling the text is a labor-intensive job and when there is a large amount of information available on a particular problem in the World Wide Web.

In this work, a text based indexing system for mammographic image retrieval and classification is presented. The system is able to interact with the common information systems used in a Department of Radiology, since it can be interfaced with the DICOM (Digital Imaging and COmmunications in Medicine) and HL7 (Health Level Seven) standard protocols. The presented system can be used as Medical Decision Support System for medical diagnosis or as case-based learning system for education. New DICOM structured reports and images can be added to the initial knowledge base, increasing its dimension and its awareness with the application domain. The system is based on Information Retrieval techniques and Machine Learning methods. Information Retrieval techniques have been used to refine the DICOM structured reports labeling process. A Bayesian Network has been developed to classify the processed DICOM structured reports. Training and testing tasks have been implemented using real cases, generated by breast physicians during their daily workflow.

3. The proposed system

The proposed system is composed of several blocks and processing phases and it is integrated with the common radiological information systems. System's main goals are the mammographic DICOM structured reports and images indexing and classification.

The structured reports, containing the free text medical reports, are collected and stored in the RIS and they can be addressed with unique *Patient ID* and *Accession Number*. The *Accession Number* is also used to extract the corresponding PACS mammographic images. The structured reports are created by breast physicians during the daily workflow and they are in Italian language.

The system architecture is shown in the Fig. 1. Continuous lines underline the workflow of the indexing process of an examination. The free text medical report is processed through six steps, identified as blue rectangles. These steps are needed to reduce and standardize medical terms and improving the information retrieval process results.

Indexed documents are stored in a database, and each of one is linked to the images of the examinations through a unique key. Dashed lines underline the workflow of the information retrieval process of an examination.

The query text is typed via a web form and it is manipulated as in the indexing process. The resulting query text is exploited by the Search Engine for extracting the most similar examinations. The query results are shown in a web page.

From an architectural point of view, the system acts as a Web Service and it can be interfaced with the existing RIS and PACS servers. Given a query, the service returns the compliant patient reports and images, selecting and extracting them from the RIS and PACS.

Information Retrieval techniques have been applied to the DICOM structured documents through the following steps [15]:

1. sentences identification;
2. semantic analysis;
3. stop-word removal and synonyms identification;
4. stemming;
5. indexing;
6. classification.

3.1. Sentences identification

In this phase a transformation is applied to the stream of input characters (the physician produced original text), generating a stream of words or tokens, or a sequence of characters having a specific meaning. In the free text medical report, the words can be easily identified by the presence of spaces, newlines, and punctuation marks. After that, a word (token) recognition process is performed using a vocabulary. In this phase, the morphology of each word has been analyzed and the role of each word in the sentence has been evaluated.

The processed dataset is composed of real clinical reports produced by several breast physicians. From these reports, the most significant words through a frequency analysis [16] have been selected. Fig. 2 shows a graph with the words sorted in descending order according to their frequency on the x-axis and the number of occurrences of each word on the y-axis.

Words occurrence analysis has been performed through the analysis of Luhn [17]. The Luhn's study analyzes each sentence in a document. The significance factor of a sentence is based on a combination of two measurements: words occurrence frequency and the relative position of a word within a sentence. The former parameter furnishes a useful measurement of words significance; the latter is the parameter for determining the meaning of a sentence.

Fig. 3 shows the *Rank of the Words* in documents vs. the *Probability Density* (see Eq. (1)) and the *Cumulative Probability Distribution* (see Eq. (2)). The continuous curve is the normalized frequency analysis of words given the rank; the dotted curve shows the evolution of the probability density of the Gaussian distribution of the dataset with $\mu = 6.43$, $\sigma = 16.44$, $Z_{\min} = -0.33$, and $Z_{\max} = 11.53$. The Z values are computed using the Eq. (3). The cumulative probability is depicted by the dashed curve.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

$$F(x) = \sum_{x \le x_i} f(x_i) \tag{2}$$

$$Z = \frac{x - \mu}{\sigma}. \tag{3}$$

According to the analysis of Luhn, it is deduced that the area identifying the most significant words of the dataset is located between the intersections of the continuous and dotted curves. As

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008

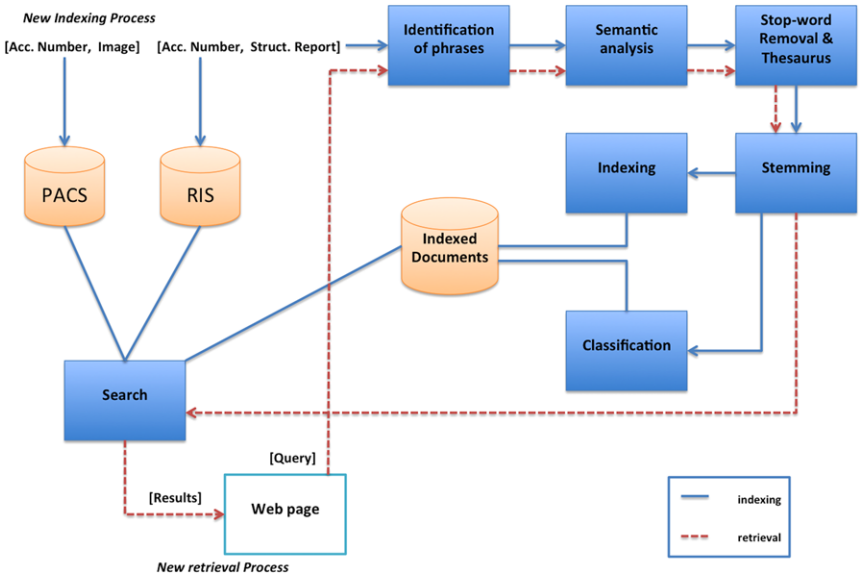

Fig. 1. The architecture of the proposed system. Continuous lines underline the workflow of the indexing process of an examination. The free text medical report of a DICOM structured report is manipulated through six steps, identified as blue rectangles. Dashed lines underline the workflow of the retrieval process of an examination. The query text is typed via a web form and query results are shown in a web page.

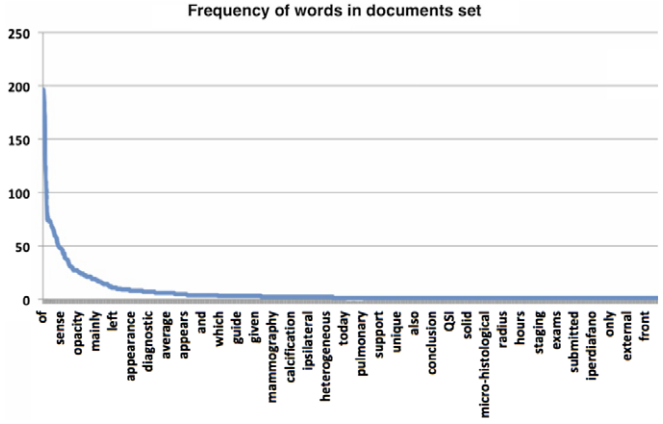

Fig. 2. Frequency of words in the processed document set.

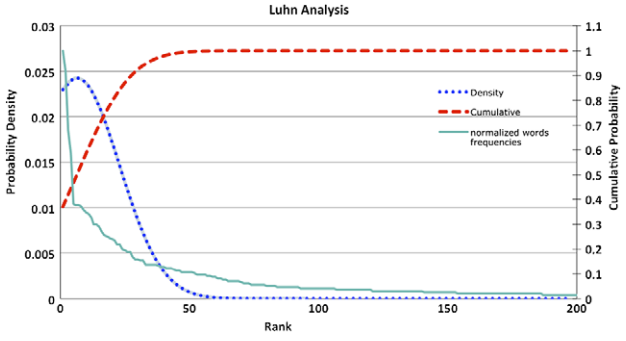

Fig. 3. Luhn Analysis: the intersection between the normalized frequency of words (continuous line) and the probability density (dotted line) marks the cutoff values used to identify the range of significant words.

depicted in Fig. 3, the occurrences of cutoff frequencies of upper-bound and lower-bound are identified by rank = 5 and probability density = 0.023 and rank = 39 and probability density = 0.0035, respectively.

The result of this analysis and the intermediate frequencies identifying the most significant terms in a sentence are shown in Table 1. In the table, the high and low frequencies terms are removed, since high frequency terms, such as prepositions, articles, and auxiliaries, are not strictly significant. They are often grammatical terms used to build every morphosyntactic sentence. At the same time, the low frequency terms are uncommon words used in few cases.

In this phase is also executed the Part-of-Speech (PoS) process [18]. This process allows for associating a word to its grammatical meaning. The list of the most significant terms are automatically tagged with their grammatical meaning using the online Italwordnet [19,20] vocabulary. The results of this process are shown in Table 2. This list will be used in the next phase of the semantic analysis, presented in Section 3.2. Italwordnet is a large semantic database within the Italian national project SITAL [20]. It shows a set of integrated resources and tools for the automatic processing of the Italian language.

3.2. Semantic analysis

The aim of this phase is to identify the meaning of the whole sentence, starting from the meaning of each term and the relationships between them. The meaning of a sentence is not only given by the contained words, but also by the knowledge of the rules for building a sentence in a given language. The combination of words, the order in which they appear in the sentence, and the links binding the words with other terms determine the sentence meaning.

In our trials, a simple semantic analysis to solve the *negation problem* in a sentence has been carried out. Negative terms are a very common phenomenon in medical terminologies [21]. Some terms are defined by opposition to the others (not otherwise specified, not Hodgkin's Lymphoma), other terms comprise the negation of relational adjectives that can be described using negative slots (not genetic hemochromatosis, not-A hepatitis, not-B hepatitis, not glomerular lesions). In a sentence containing a negative term, it is measured the distance between the negative term and the closest adjective. Adjectives are contained in the list created in the previous phase. A distance function calculates the number of words between the negative term and the closest adjective. If the distance is less than an experimentally fixed threshold, the sentence is modified by deleting the negative term and replacing the adjective with its opposite. The threshold value has been fixed considering the average distance between the negative terms and the relative adjectives in the training set.

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008

Table 1
Main significant words and the relative frequencies.

Word	Freq	Word	Freq
structure (struttura)	75	qse (qse)	38
suggest (consiglia)	74	right (destra)	37
check (controllo)	73	opacity (opacita)	37
breasts (mammelle)	70	eteroplastic (eteroplastico)	33
images (immagini)	68	microcalcifications (microcalcificazioni)	31
pathological (patologiche)	67	left (sn)	31
fibroadiposa (fibroadiposa)	59	breast (mammella)	30
sense (senso)	59	centimeters (cm)	30
appreciable (apprezzabili)	57	in the (nel)	27
radiographically (x-graficamente)	53	examination (esame)	27
periodic (periodico)	48	diameter (diametro)	27
absence (assenza)	47	etero-formative (eteroformativo)	27
annual (annuale)	46	millimeters (mm)	27
ultrasound (ecografica)	43	about (circa)	26
integration (integrazione)	43	fibro-glandular (fibroghiandolare)	25
presence (presenza)	39	mainly (prevalentemente)	25

3.3. Stop-word removal and synonyms identification

In this phase high frequent words will be deleted from the processed dataset. The deletion is needed because several statistical analyses have demonstrated that high frequency words lead to the performance degradation of a search engine. As mentioned in Section 3.1 and in Section 3.2, negative terms, articles, prepositions, and so on have been removed after the related analysis.

The method adopted in this phase uses a statistical approach for the terms frequency analysis in the dataset. The text has been manipulated using a clinician's thesaurus to replace each term with the synonym having the maximum length. In this way, the words having the same meaning are replaced with the same synonym term. This process increases the frequency value of substituted terms and decreases the total number of terms in the dataset. High frequency terms highlight the words with low discriminatory power, including prepositions, auxiliary verbs and other common used words. As an example, stop-words removal allows for the reduction of 30% ÷ 50% of used terms.

3.4. Stemming

In this phase, the words are reduced to their semantics roots. Stemming algorithms are fusion procedures reducing all the words, having the same root, to the stem. Words starting with the same set of characters or having a character sequence in common can have the same etymological origin and similar information content. Generally, this procedure removes the end of the words leaving a common stem. Because of its high performance, the Italian version of the Porter algorithm has been used in this work [22]. The algorithm reduces inflected and, sometimes, derived words to their stem base or root form.

3.5. Indexing

In this phase each diagnostic report is indexed. The process collects, parses, and stores indexed data to facilitate fast and accurate information retrieval operations on mammographic reports and images. The Information Retrieval module is based on term-text information, i.e. terms extracted after the natural language processing phases. In the training phase, the reports are processed off-line to obtain a useful representation capturing its content. Physician or student queries are processed on-line, selecting the set of mammographic reports and images that are presented in descending order with respect to the degree of similarity with the submitted query.

The model used in this work is based on vector spaces and it is called the vector-space model. Each element is weighted, while documents and queries are represented by vectors of length k. Different techniques can be used to determine the weight of each element. Classical indexes computing the frequency of each term in a single document, i.e. the Term Frequency (TF), and the entire collection, i.e. the Inverse Document Frequency (IDF) have been adopted [23].

Table 2
The grammatical tags of the main significant words in the dataset. Part of the terms has been tagged using Italwordnet.

Word	TAG	Word	TAG
low (bassa)	Name	undefined (indeterminato)	Adj
right (destra)	Name	deserving (meritevole)	Adj
present (presente)	Adj	structural (strutturale)	Adj
trasformation (trasformazione)	Name	presence (presenza)	Name
likely (verosimilmente)	Adverb	morphology (morfologia)	Name
nodule (nodulo)	Name	pulmonary (polmonare)	Adj
absence (assenza)	Name	incidence (incidenza)	Name
cyst (cisti)	Name	data (dato)	Name
waiting (attesa)	Name	data (dati)	Name
indicate (segnalare)	Verb	date (data)	Name

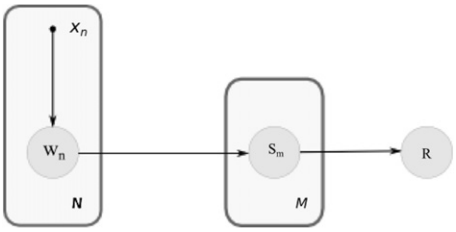

Fig. 4. The Bayesian Network used in this work. The random variables W are associated to the significant mammographic words. The S variables are associated to the status of the sentence. The R variable is associated to the structured report status and it can hold one of two possible states: *healthy concept*, *pathological concept*.

3.6. Classification

The current phase deals with the structured reports classification. Classification tasks have been performed using the Naive Bayesian Classifier. Fig. 4 shows the Bayesian Network used in this work [24]. The Bayesian Network is able to infer the status of each sentence of a structured report. Sentence classification has been performed selecting some significant words through the Luhn Analysis [17].

In the training phase, expert breast physicians have clustered the processed sentences in two classes: *healthy concept*, *pathological concept*. The random variables W are associated to the significant mammographic words (see Table 1). A variable $w \in W$ is

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008

accounted in the Bayesian Network when the associated significant word is in the sentence. The information about the word state *is/is not* is held by the input X. The S variables are associated with the status of the sentence and the values of these random variables are two: *healthy concept, pathological concept*. The final classes suggest the predominant concept of the sentence. The R variable is associated to the structured report status and it can hold one of two possible states: *healthy concept, pathological concept*.

The Naive Bayesian Classifier choice is motived. The problem addressed in this work follows the assumption given by the Naive Bayesian Classifier. Given the class variable, it can be assumed that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. In other words, the Naive Classifier is based on the assumption that, given a value of the objective function, the attribute values are conditionally independent. So that, given a new item q_n, the classification is computed as shown by Eq. (4).

$$\begin{aligned} c_{q_n} &= \arg\max_{c\in Y} P(c|q_1, q_2, \ldots, q_n) \\ &= \arg\max_{c\in Y} \frac{P(q_1, q_2, \ldots, q_n|c)P(c)}{P(q_1, q_2, \ldots, q_n)} \\ &= \arg\max_{c\in Y} P(c)\prod_j P(q_j|c) \end{aligned} \tag{4}$$

where Y is the set of the possible classes that the variable q_n can assume.

In this work, the Eq. (4) is computed in two steps: initially the status of each sentence S is computed and then the structured report status R is inferred. In the first step, the sentence status S is inferred using the Eq. (5). The set H is the set of the classes assumed by the variable S. The labels of the two classes are: *healthy concept, pathological concept*.

$$\hat{S}_{n c\in H} = P(c)\prod_j P(w_j|c). \tag{5}$$

Considering the Bayesian Network depicted in Fig. 4, if and only if x_n is *true*, i.e. the word w_n is contained in the sentence, the variable w_n is considered in the network. The result of the Eq. (5) is a 2-dimensional vector $\hat{S}_n$ with continuous values for the labels *healthy concept, pathological concept*.

In the second step, the computed 2-dimensional vector is used to infer the status of the structured report R, as shown in Eq. (6).

$$R = \arg\max_{c\in H} P(c)\prod_j P(s_j|c). \tag{6}$$

The bigger value of the two labels infers the structured report status. To infer the status of the structured report R, computing the Eq. (6), the vector $\hat{S}_n$ is used. In fact, the 2-dimensional vector is the conditional probability $P(s_j|c)$.

The used training dataset is composed of 100 cases: 50 structured reports of healthy patients and 50 structured reports of pathological patients. So, the a-priori probabilities $P(c)$ for the *healthy concept, pathological concept* classes are fixed to 50%. Conditional Probability Tables (CPT) have been computed using the *frequentist approach*. The trained CPT are associated to the conditional probability $P(w_j|c)$, the values for the two classes are computed using the Eq. (7).

$$\begin{aligned} P(w_i|c = \text{`}healthyconcept\text{'}) &= \frac{\#w_i \in T_h}{|T_h|} \\ P(w_i|c = \text{`}pathologicalconcept\text{'}) &= \frac{\#w_i \in T_u}{|T_u|} \end{aligned} \tag{7}$$

where $\#w_i$ is the number of words W contained in the dataset, T_h is the structured report set of healthy patients, and T_u is the structured report set of pathological patients.

As mentioned above, the $P(s_j|c)$ is computed *on-line*, hence it does not need an *off-line* training setup.

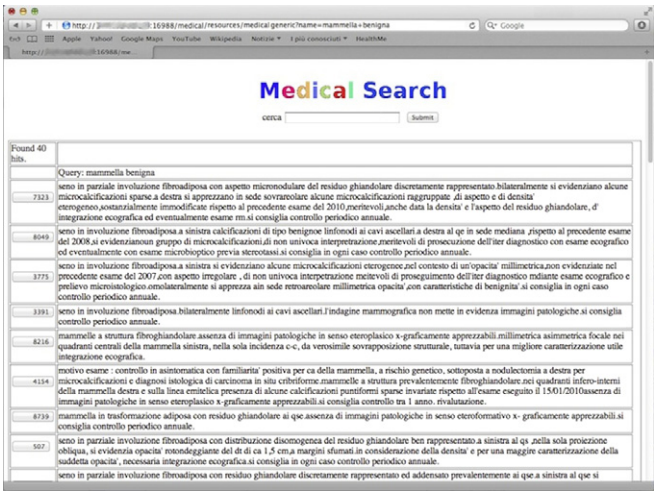

Fig. 5. The web page generated by the system. For each selected case, the structured report and the corresponding Accession Number are shown.

4. Experimental trials

The experimental framework has been developed in Java and it uses some online available libraries, such as the Apache Lucene (TM) library [25]. This library is used for full-featured text search engine in the *Indexing* and *Search* stages. Lucene is a high-performance open source library implementing the basic functions for text documents indexing and searching. The search functions are performed using an inverted index of documents, computed in the preliminary stage. The index quality is the key for the Lucene efficiency evaluation in the user queries management.

User's interaction is performed through a web interface showing the indexing and searching functionalities. The web interface is inspired by the works presented in [26,27]. It is generated through the GlassFish server [28], an open source Java EE application server. With more details, the server hosts the web service offering the indexing and classification services. The choice of this technology makes possible the service access via any browser, after the physician authentication phase. So, it avoids compatibility issues with the different systems currently used in a Department of Radiology.

Figs. 5 and 6 show the system at work. Fig. 5 includes a search bar for entering query keywords. The process starts pressing the "submit" button and it gives out the most similar structured reports including the query terms (on the right) and the corresponding accession number as link to the patient images (on the left). Fig. 6 shows the displayed mammographic images through the Oviyam web based open source software [29]. The Oviyam software has been directly interfaced with the developed web service.

Three sets of binary classification tests have been conducted to proof the effectiveness of the proposed approach. In the first phase, the *Indexing Engine* and the *Search Engine* have been tested, leaving out the Bayesian classifier. The related results are detailed in Section 4.1. In the second phase, the performance of the Bayesian classifier, without any text processing technique, has been evaluated. The related results are detailed in the Section 4.2. Finally, the classifier has been integrated in the system, and the whole system has been tested and evaluated. The related results are detailed in Section 4.3, and they show substantial performance improvements.

System performance has been evaluated by using several indexes. In Information Theory, the reliability of a binary classification test (True/False, Positive/Negative) is generally assessed in terms of Sensitivity (Se) and Specificity (Sp) [30]. The two indexes are defined by means of four parameters: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). In our trials, True Positive is the number of pathological reports classified

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008

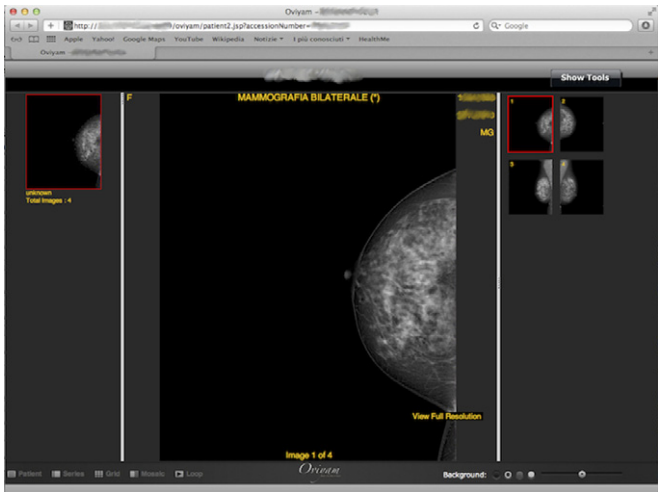

Fig. 6. Mammographic image linked to a previously selected report.

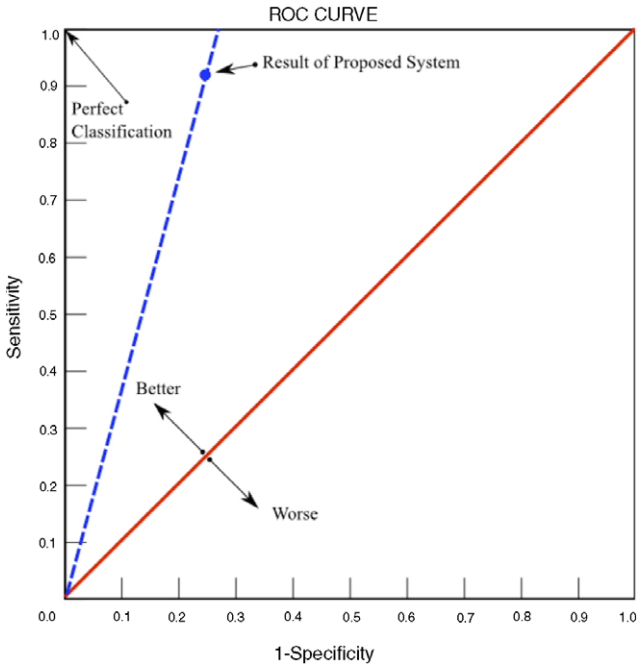

Fig. 7. The ROC curve. The continuous line shows the 50% A_z value for a random performance classifier. The dashed line shows the A_z value of the proposed system.

as such, False Positive is the number of healthy reports classified as pathological, False Negative is the number of pathological reports classified as healthy, and True Negative is the number of healthy reports classified as such. Sensitivity measures the percentage of actual positives which are correctly identified as such, while Specificity measures the percentage of actual negatives which are correctly identified as such. In numerical terms, the two parameters are calculated as follows:

$$Sp = \frac{Tn}{Tn + Fp} \tag{8}$$

$$Se = \frac{Tp}{Tp + Tn}. \tag{9}$$

In addition, two more indexes have been used for the system performance evaluation. The *Sp* and the *Se* values are used to identify a point on the ROC (Receiver Operating Characteristic) curve. The ROC curve is a graphical plot that illustrates the performance of a binary classifier when its discrimination threshold is varied. The area under curve (A_z) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative' ranks). A test with an area under the ROC curve greater than 80% is considered as an appreciable diagnostic test [31]. The standard error σ_{Az} for the area A_z has been also calculated to validate the achieved results. Fig. 7 shows a detailed example for A_z and σ_{Az} computation.

4.1. Indexing and search engine testing

To test the reliability of the *Indexing Engine* and the *Search Engine*, the *Sp* and *Se* indexes have been evaluated, considering three different classes of queries. Queries can be arranged with one or more keywords that are used to select the appropriate reports and images. For the reasons already presented, the described queries are linked to the Italian language. In the first case, the submitted query was *"mammella maligna", i.e. "malign breast"*. The system selected the pathological cases in indexed documents, with no pathology distinction. Specificity and Sensitivity values were 92% and 72%, respectively.

In the second case, the submitted query was *"mammelle con opacitá", i.e. "breasts with opacity"*, addressing a test in which mammographic images with opacity have to be extracted from the radiological information systems. In the processed dataset, 35 entries present opacity: 27 are pathological cases, while 8 are healthy cases. The values of Sensitivity and Specificity are 77% and 89%, respectively.

In the third case, the submitted query was *"mammelle fibroadipose benigne", i.e. "benign fibroadipose breasts"*, addressing a test in which no pathological mammographic images with a *fibroadipose* structure have to be extracted from the radiological information systems. In the processed dataset, 41 entries are cases of breasts with *fibroadipose* structure: 26 are healthy cases, while 15 are pathological cases. The Sensitivity is 100%, so that all cases of benign breast with *fibroadipose* structure are correctly identified. However, the related Specificity is about 40%.

The described tests show relevant values of Sensitivity, underlying that the system correctly detects the information in the pathological structured reports. The system also presents low values of Specificity, such that a healthy patient is positive to the test, but this does not exclude the presence of some outliers in the results.

Finally, in the third case (query 3), the area under the ROC curve is 98%, while the standard error is $\sigma_{Az} = 2.9e^{-3}$.

Table 3 summarizes the values of *Specificity* and *Sensitivity* of the three tests.

4.2. Bayesian classifier testing

The present section describes the experimental tests for the Bayesian Classifier evaluation. This module has been tested using a set of 200 unclassified structured reports, with no application of the previously described Information Retrieval techniques. After the training phase, the system labels each new structured report as *healthy concept* or *pathological concept*. The *TP*, *FP*, *FN*, and *TN* values have been computed and the classification results have been validated by an expert breast physician. Table 4 summarizes the values of *Specificity* and *Sensitivity*.

The conducted trials show relevant values of Sensitivity, i.e. the developed classifier correctly labels the pathological structured reports as *pathological concept*. The system presents acceptable values of Specificity, such that a structured report of a healthy patient is labeled as *healthy concept*.

Fig. 7 shows the point identified by the *Se* and *Sp* values. In Fig. 7, a continuous line for the first diagonal of the graph underlines this value and the A_z area below it. The area A_z is identified by the *Se* and *Sp* reported in Table 4. In numerical terms, the area A_z has a value of 92%, while the standard error is $\sigma_{Az} = 2.3e^{-5}$.

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008

Table 3
Specificity and sensitivity of the indexing and search engine.

	Query 1	Query 2	Query 3
Sp	92%	89%	40%
Se	72%	77%	100%

Table 4
Specificity and sensitivity of the Bayesian classifier.

	True	False
Positive	35	3
Negative	36	109
Sp	75,2%	
Se	92,1%	

Table 5
Specificity and sensitivity of the complete system.

	Query 1	Query 2	Query 3
Sp	96%	91%	81%
Se	90%	82%	96%

Table 6
Sp, Se, A_z and σ_{Az} for query 3.

	Indexing and search engine	Complete system
Sp	40%	81%
Se	100%	96%
A_z	98%	96%
σ_{Az}	$2.9e^{-3}$	$1.08e^{-4}$

4.3. The complete system

In this section the performance of the whole architecture, integrating the described Information Retrieval techniques and the Bayesian Classifier, are described. The experimental trials' organization follows the guidelines presented in Section 4.1. As shown in Table 5, the previous three queries have been submitted and the related *Se* and *Sp* have been measured and compared.

In the first case, with the query *"mammella maligna", i.e. "malign breast"*, the Specificity rises from 92% to 96%, while the Sensitivity rises from 72% to 90%. These values denote that only few reports of *malign breast* are missed. In the second case, with the query *"mammelle con opacitá", i.e. "breasts with opacity"*, both Specificity and Sensitivity rise. In fact, the former becomes 91% (it was 89%), while the latter becomes 82% (it was 77%). In the last case, with the query *"mammelle fibroadipose benigne", i.e. "benign fibroadipose breasts"*, the Specificity rises from 40% to 81%, while the Sensitivity shows a light reduction of about 4%.

In the third case (query 3), the area A_z has a value of 96%, while the standard error is $\sigma_{Az} = 1.08e^{-4}$. Table 6 shows a comparison between the system described in Section 4.1 and the complete system. With more details, the Specificity (Sp), the Sensitivity (Se), the area A_z and the standard error σ_{Az} are reported for query 3.

The combined use of the Bayesian Classifier and the described Information Retrieval techniques lead to a system showing better Specificity and Sensitivity working points.

5. Conclusions

The ability to analyze and correlate Big Data is deemed to be very useful in health care domain. Fusion of unstructured heterogeneous data would be very attractive for health structures. In this paper a text based indexing system for mammographic images retrieval and classification running as Web Service has been presented. The indexing/searching process is implemented using Information Retrieval techniques for document processing and naive Bayesian classifier for document classification. The whole system shows interesting results and provides a real-time useful Medical Decision Support System to be used during the referral process. In the current version, the considered Big Data source was composed of mammographic images and structured reports, containing free text for medical diagnosis. Future directions will be aimed to integrate heterogeneous clinical data, such as biopsy reports, surgery data, follow-up reports, etc. to build a complete knowledge base addressing the breast pathologies domain. The new data will increase knowledge based consistency in order to develop innovative Big Data based Medical Decision Support Systems.

Acknowledgments

This work was partially supported by the Italian Ministero della Salute (project code RF-SIC-2007-646441) and by the Italian Ministero dell'Istruzione, dell'Universitá e della Ricerca (PON Smart Cities PON04a2_C "SMART HEALTH - CLUSTEROSDH - SMART FSE - STAYWELL").

References

[1] D. Clunie, DICOM Structured Reporting, PixelMed Publishing, 2000.
[2] C. Bishop, et al., Pattern Recognition and Machine Learning, vol. 4, Springer, New York, 2006.
[3] H. Guo, W. Hsu, A survey of algorithms for real-time bayesian network inference, in: AAAI/KDD/UAI02 Joint Workshop on Real-Time Decision Support and Diagnosis Systems, Edmonton, Canada.
[4] A.P. Bradley, The Use of the Area Under the Roc Curve in the Evaluation of Machine Learning Algorithms, vol. 30, pp. 1145–1159.
[5] C. Ogescu, C. Plaisanu, D. Bistriceanu, Web based platform for management of heterogeneous medical data, in: Automation, Quality and Testing, Robotics, AQTR 2008. IEEE International Conference on, vol. 3, IEEE, 2008, pp. 257–260.
[6] C. Chang, H. Lu, Integration of heterogeneous medical decision support systems based on web services, in: Bioinformatics and BioEngineering, BIBE'09. Ninth IEEE International Conference on, IEEE, 2009, pp. 415–422.
[7] F. Elevitch, Snomed ct: electronic health record enhances anesthesia patient safety, AANA J. 73 (2005) 361.
[8] H. Lowe, G. Barnett, Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches, JAMA: the J. Am. Med. Assoc. 271 (1994) 1103–1108.
[9] E. Ciolko, F. Lu, A. Joshi, Intelligent clinical decision support systems based on snomed ct, in: Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE, IEEE, 2010, pp. 6781–6784.
[10] J. Ontrup, T. Nattkemper, O. Gerstung, H. Ritter, A mesh term based distance measure for document retrieval and labeling assistance, in: Engineering in Medicine and Biology Society, Proceedings of the 25th Annual Int. Conference of the IEEE, vol. 2, IEEE, 2003, pp. 1303–1306.
[11] K. Nigam, A. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using em, Mach. Learn. 39 (2000) 103–134.
[12] C. Devasena, M. Hemalatha, Automatic text categorization and summarization using rule reduction, in: Advances in Engineering, Science and Management (ICAESM), Int. Conference on, IEEE, 2012, pp. 594–598.
[13] B. Liu, Y. Dai, X. Li, W. Lee, P. Yu, Building text classifiers using positive and unlabeled examples, in: Data Mining, ICDM 2003. Third IEEE Int. Conference on, IEEE, 2003, pp. 179–186.
[14] S. Zelikovitz, H. Hirsh, Improving short text classification using unlabeled background knowledge to assess document similarity, in: Proceedings of the Seventeenth International Conference on Machine Learning, pp. 1183–1190.
[15] A. Farruggia, R. Magro, S. Vitabile, Novel web service for mammography images indexing, in: The 27th IEEE International Conference on Advanced Information Networking and Applications, (AINA-2013), Barcelona, Spain, 2013, pp. 225–230.
[16] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949.
[17] H. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958) 159–165.
[18] D. Das, S. Petrov, Unsupervised part-of-speech tagging with bilingual graph-based projections, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, pp. 600–609.
[19] E. Pianta, L. Bentivogli, C. Girardi, Developing an aligned multilingual database, Proc. 1st Int'l Conference on Global WordNet.
[20] A. Roventini, A. Alonge, N. Calzolari, B. Magnini, F. Bertagna, Italwordnet: a large semantic database for italian, in: LREC.
[21] C. Jacquelinet, Opposition principles and antonyms in medical terminological systems: structuring diseases description with explicit existential quantification, in: Connecting Medical Informatics and Bio-Informatics, 2005, pp. 1261–1265.
[22] P. Willett, The porter stemming algorithm: then and now, Program: Electron. Libr. Inf. Syst. 40 (2006) 219–223.
[23] A. Aizawa, An information-theoretic perspective of tf–idf measures, Inform. Process. Manag. 39 (2003) 45–65.

[24] A. Farruggia, R. Magro, S. Vitabile, Bayesian network based classification of mammography structured report, in: Int. Conference on Computer Medical Applications, (ICCMA 2013), Sousse, Tunisia, 2013, pp. 1–5.
[25] E. Hatcher, O. Gospodnetic, Lucene in Action (In Action series), Manning Publications Co., Greenwich, CT, USA, 2004.
[26] V. Cannella, O. Gambino, R. Pirrone, S. Vitabile, Gui usability in medical imaging, in: Complex, Intelligent and Software Intensive Systems, CISIS'09. Int. Conference on, IEEE, 2009, pp. 778–782.
[27] E. Torres, F. Fauci, R. Magro, L. Ramello, M. Fantacci, U. Bottigli, G. Masala, P. Oliva, S. Bagnasco, P. Cerello, Use of hep software for medical applications, in: Computing in High Energy and Nuclear Physics, 2003.
[28] A. Goncalves, Beginning Java EE 6 with GlassFish 3, Springer, 2010.
[29] OVIYAM Website. Last access on 15/04/2013. http://oviyam.raster.in/.
[30] C. Metz, Basic principles of roc analysis, in: Seminars in Nuclear Medicine, vol. 8, Elsevier, pp. 283–298.
[31] J. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.

Alfonso Farruggia is a post doctoral fellow with the University of Palermo, Italy. He received the Computer Engineering M.Sc. degree from the Polytechnic of Turin, Italy and the Ph.D. degree in Computer Engineering from the University of Palermo, Italy in 2008 and 2011, respectively. During the Ph.D. course, he has conducted his research on issues related to anomaly detection in Wireless Sensor Networks using Machine Learning and Computer Intelligence methods. His research interests include Wireless Sensor Networks, Computer Intelligence methods, and Information Retrieval in healthcare.

Rosario Magro received the Computer Engineering Laurea Degree and doctoral degree in Applied Physics from the University of Palermo, Italy. He has conducted research on issues related to mammographic images processing for assisted diagnosis of mammary pathologies. He also is a CEO of the CyclopusCAD Srl company, academic Spin-Off of University of Palermo, operating in the Medical and IT domain.

Salvatore Vitabile is currently an Assistant Professor with the Department of Biopathology, Medical and Forensic Biotechnologies at the University of Palermo, Italy. He received the Laurea degree in Electronic Engineering and the doctoral degree in Computer Science from the University of Palermo in 1994 and 1999, respectively. In 2007, he was a Visiting Professor with the Department of Radiology, Ohio State University, USA. He is currently a member of the Board of Directors of SIREN (Italian Society of Neural Networks). He is the Editor in Chief of the International Journal of Adaptive and Innovative Systems, Inderscience Publishers. He has also joined the Editorial Board of the International Journal of Information Technology, Communications and Convergence and of the International Journal of Space-Based and Situated Computing, Inderscience Publishers.

His research interests include biometric authentication systems, real-time driver assistance systems, multi-agent system security, and medical data processing and analysis.

Please cite this article in press as: A. Farruggia, et al., A text based indexing system for mammographic image retrieval and classification, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.02.008