Knowledge Discovery Methodology for Medical Reports

Vitor Pinheiro and Victor Alves

Departamento de Informática,
Universidade do Minho
vitor.amares@gmail.com, valves@di.uminho.pt

Abstract. Medical reports contain valuable information, not only for the patient that waits for the results but also the latent knowledge that is possible to extract from them. The recent introduction of standard structured formats like the Digital Imaging and Communications in Medicine Structured Report and the Clinical Document Architecture Health Level Seven provide an efficient generation, distribution, and management mechanism. Also, they provide an intuitive and effective manner of information representation, unlike the traditional plain text format. In this paper we present a knowledge discovery methodology for structured report interchange based on plain text medical reports using YALE, a leading open-source data mining tool and Open-ESB platform that provides conversion, parsing, different protocols and message formats interchange capabilities.

Keywords: SOI, SOA, Open-ESB, BPEL, Data Mining, YALE, DICOM, HL7.

1 Introduction

Medical reports are the support and contact point between the diagnosis, the knowledge transmission of the specialist physician, and the patient or other physician. In the majority of the studies, these reports are plain text information without any standardized structure or contents sharing, that could be achieved by the usage of medical images attached to the reports. A medical report is an individual case study without any type of general correlation or association with others studies or diagnosis. These cases can be considered as epidemic studies for a given population, population sample or other option.

With the establishment of open medical standards, like DICOM - Digital Imaging in Medicine [1] and HL7 - Health Level Seven standard [2] it is now possible to integrate heterogeneous systems and medical information [3]. DICOM defines standards for exchanging, storing and printing medical imaging related information that can be used by medical equipment. HL7 defines standards to interconnect and control clinical and administrative data, of the whole life cycle of a patient clinical documentation, using the HL7 Clinical Document Architecture (CDA).

In addition to the DICOM standard definition of acquisition and storage of waveforms and images, DICOM also includes Structured Reporting (SR), a great expressive and hierarchical representation of structured medical information, containing text with links to other data such as images, waveforms, and spatial or temporal coordinates [4]. However, the medical report as we know it is static, delivered on paper and sometimes without any kind of format or in accordance with any standard. In general, this kind of medical information is treated only by the physician that asked for the medical exam and the physician who diagnosed it. There's no correlation or other kind of data analysis or knowledge discovery within this medical information.

One initial goal for the SR was that the information encoded in such reports would become more readable and thus easier to extract, then an unstructured plain text or paper report [5], making it easier to index and to selective retrieve information, without having to rely on Natural Language Parsing (NLP) [6]. Also, SR encoding and structure allows queries and data mining operation such as a query for all documents where a malignant mass of a specific dimension is reported. These operations are supported because every element of information is described by a code. A code value that can be unambiguously identified enabling data mining [4] using a data mining tool like YALE (Yet Another Learning Environment), as formally called, a data mining platform (environment) that simplifies the construction of experiments and the evaluation of different approaches. Using YALE is possible to build different experiments models to discover relations and patterns within this data, in the medical report could provide new medical knowledge, applying methods that have been developed to discover this hidden medical knowledge [7].

Nevertheless, for external communications of these medical reports with other departments it is necessary to use HL7. According to [8] DICOM SR is a matter of primary interest to HL7 CDA for different reasons. In the practical level, generally, the end users of SR are referring physicians using HL7-based systems. The SR usability requires a method of exporting these results to the HL7 domain. Although, HL7 and DICOM have joined efforts to adjust CDA and SR to avoid incompatibilities, there still is no mechanism of bi-directionally trans-coding, SR to CDA, with full fidelity defined by either group, and may never be defined, nor do they exists for plain text to structured report [5]. All these problems are common in Enterprise Application Integration (EAI). EAI technology enables incompatible protocols and messages formats to be exchanged by different entities [9]. Open-ESB [10] is a platform built using open standards like JBI [19] that can be used as a platform for both Enterprise Application Integration and Service Oriented Architecture applications development. The Open-ESB architecture enables communication between different protocols and messages, synchronous and asynchronous, interoperability and scalable applications. This advantages can solve problems related with medical reports, for example, like the trans-coding bi-directionally mechanism between DICOM and HL7.

In this paper we present a methodology for knowledge discovery for structured report interchange based on plain text medical reports using YALE [11] based on the Weka [12] tool and Open-ESB [10] platform that provides conversion, parsing, different protocols and message formats interchange capabilities.

2 Medical Reports

DICOM Structured Reports can be used for different purposes with different levels of complexity. These different levels of complexity are related with the target report or with the diagnosis complexity. So we have three different types of structured report classes:

- *Basic Text* – minimal code use, for document title, subtitles and hierarchical subtitles tree
- *Enhanced* – superset of the Basic Text, numerical measurements with representative codes for units and measurements, images and waveforms references.
- *Comprehensive* – superset of the Basic Text and Enhanced, references between elements.

The information in SR is grouped in nine modules, in which the items of information are related. There is a module for information on the patient, such as date of birth and weight, a module for the general information compliance with the document, such as names of people responsible for verifying the document and flags that indicate whether the document was found complete, and so on. Although, SR documents are not necessarily on a patient, may be on a sample as a sample of human tissue for analysis. The information contained in the contents module of the document is divided in "contents items". A content item consists of a pair name-value, where the name is a code of a selected dictionary of terms like the SNOMED CT - Systematized Nomenclature of Medicine-Clinical Term [13] and the value is a type among the fourteen types of value defined by default. These types are *text* (for text), *num* (for numbers, percentages), *image*, *date* and *waveform*. The items are hierarchically organized, so that the information in the highest levels of the hierarchy contains, or is derived from, information on items below the same. The DICOM SR standard specifies eight different types of relationships, among them are *contains* (node father of the information is contained in node child), *has properties* (information to the node is a child's property information node of the father), *has obs. Context* (the information in the node is a child comment on the information of the node father). The following sentence and examples adapted from [14], in natural language, could be divided into items of information and organized in a hierarchy:

```
Personal Habits: In average the patient drank 7 beers
per day in the last 2 years. The patient is obese, and
often passes away…
```

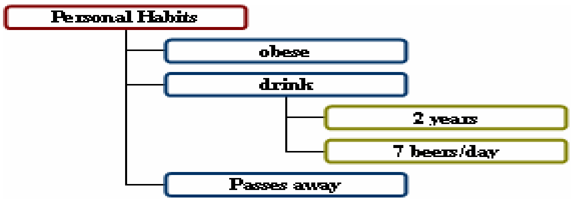

Fig. 1. Personal Habits example – the information was divided according to the most important concepts.

In DICOM SR each item should be a pair name-value, each containing a relationship with its father item and a value type, indicated above the item (Fig. 2).

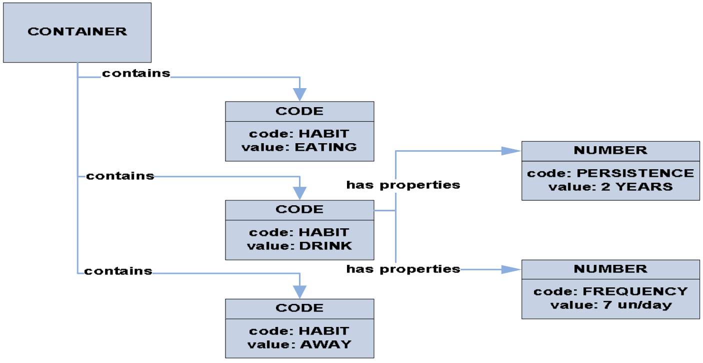

Fig. 2. Items hierarchy according to DICOM SR.

The biggest question in the structured reports is the integration and acceptance by the physician for this way representing the diagnosis. Most of them have there own templates, defined in a dot file type, and do not want to be restrained by any standard format. In our case study we use CT (Computer Tomography) reports, in text mode, that have an associated template.

3 State of Art

The introduction of open standards like HL7 and DICOM present expressive representations of hierarchical structures of medical information, capable of containing text with links to other structured data such as images and described by codes, that can be unambiguously identified enabling data mining [4]. However, the physicians continue to write their reports ignoring these standards or simply do not know of their existence. Thus, their reports can suffer from the ambiguous terms of natural language [15] and sometimes do not address the key clinical question [16], containing clinically important errors [17] despite the fact that a study showed that referring physicians strongly prefer concise well-organized radiology reports [18].

There is no well known mechanism of bi-directionally trans-coding SR to CDA with full fidelity formally defined [5]. This problem is common in Enterprise Application Integration (EAI). EAI technology enables incompatible protocols and messages formats to be exchanged by different entities [9].

In [25] we can find a proposal for a text mining system to extract and use the information in radiology reports that consists of three main modules: a medical finding extractor, a report and image retriever, and a text-assisted image feature extractor. In our approach these modules are included as the information integration from different sources, protocols, message formats or standards using service oriented integration architecture.

3.1 Service Oriented Integration

Common problems in enterprise application integration are the incompatible protocols and messages formats. In response to this kind of problems, the actual industry path is based on standards definition for business integration and standard metadata in the web services stack [9].

The JBI 1.0, JSR-208 [19], specification is an industry-wide initiative to create a standardized integration platform for Java and business applications addressing Service-Oriented Architecture (SOA). JBI employs concepts similar to J2EE to extend application packaging and deployment functionality to include JBI Components. JBI Components are an open-ended class of components that are based on JBI abstract business process metadata [19].

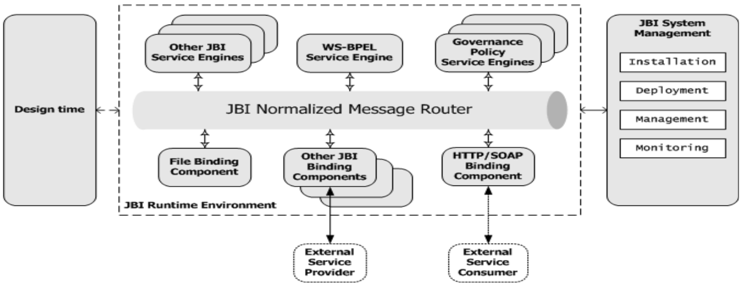
Design time
Other JBI Service Engines
WS-BPEL Service Engine
Governance Policy Service Engines
JBI Normalized Message Router
File Binding Component
Other JBI Binding Components
HTTP/SOAP Binding Component
JBI Runtime Environment
External Service Provider
External Service Consumer
JBI System Management
Installation
Deployment
Management
Monitoring

Fig. 3. The JBI Environment [20].

According to [20] the key pieces of the JBI environment (Fig. 3) are:

- *Service Engines* - enable pluggable business logic;
- *Binding Components* - enable pluggable external connectivity;
- *Normalized Message Router* – directs normalized messages from source components to destinations according to specified policies.

Java Business Integration (JBI) provides a foundation for building a SOA and is the foundation of the Open-ESB platform. It allows anyone to create JBI-compliant integration plug-in components and integrate them dynamically into the JBI infrastructure. Despite this, JBI alone does not have a single point of administration for the entire system and each operation in the system requires knowledge of the system topology. The Open-ESB solve these problems using the Java Open Enterprise Service Bus built with JBI technology enabling a set of distributed JBI instances to communicate as a single logical entity that can be centrally managed. Using the Open-ESB it's possible to integrate existing business functions as services, and decoupled interaction between service providers and consumers. It provides direct support for composite application creation through the mechanism of JBI service assemblies, which allow applications to be composed directly from the service-based interfaces of JBI service units and BPEL (Business Process Execution Language) orchestrated [20].

This direct support for composite application construction atop a service-oriented architecture and standards-based messaging infrastructure makes JBI an ideal foundation for constructing service-oriented applications and accomplishing service-oriented integration of existing systems using normalized messages in XML [9].

Integrating computing entities using only service interactions in a service-oriented architecture is defined as Service-Oriented Integration (SOI). Service-oriented integration solutions deal better with integrating legacy problems and inflexible heterogeneous systems using more often functionalities that were hidden in different applications as reusable services. The main advantages, for the traditional enterprise application integration (EAI) are the application of standards to define standards interfaces, the opaqueness of the functionality that is hidden from the service interface and the flexibility of the service in the perspective of the consumer and producer that can change except the description of the service.

4 Knowledge discovery methodology

The free plain text medical report cannot just simply be abandoned and substituted by the structured report. This does not just happen and will not happen until the physician is comfortable with standards or with tools that generate or help to write a structured report [14]. Beside this, the physician will have there own template for each possible diagnosis with his own personal mark. Also we must not forget the existent legacy plain text file medical reports that contain medical information that are valued information in future exams. Knowing that and in accordance with the presented standards for integration and medicine it's possible to suggest an information integration and knowledge extraction corresponding to the physician desires, the new structured report, all the platforms and medical systems related to this scope and obtain any extra information with the knowledge discover possibility.

We present a methodology definition for knowledge discovery supported by Service Oriented Integration architecture, the set of rules or principles that should be followed to discovery latent information in the medical reports, like patterns and association rules:

1. Use a service oriented integration architecture able to connect different services, that can use different protocols and messages;
2. Obtain a encode and decode service of a free text medical report to a normalized format, based in a model;
3. Achieve a mapping between medical terms used in the medical report and a codification non ambiguous solving problems concerning the use of language natural in medical reports (denial);
4. Apply knowledge discovery algorithms to the normalized integrated information e stored in the database.

Considering the medical reports as legacy files stored in a FTP (File Transfer Protocol) server we can orchestrate in BPEL a business process were the medical reports are encoded and the structured information saved on a database for future data mining operations using YALE.

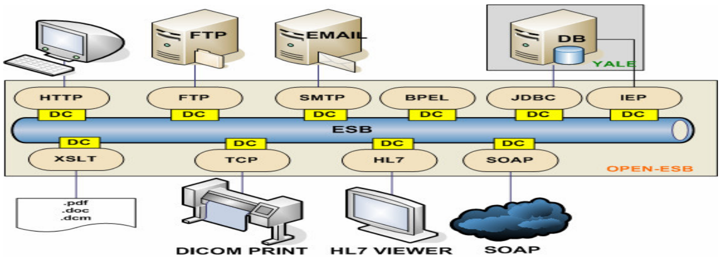

Fig. 4. Overall of the Medical Report Integration Architecture.

One possible integration process (we can have many others) is based on the architecture illustrated by Fig. 5 that is supported by the Open-ESB platform. The first step for this integration process (Fig. 6) is the oriented template parser definition. Since they all have been created based on a template we can parse the medical report using a defined XSD for each physician template and return all the information in a structured format from the FTP server. The encoded operation permits the interchange of the information between the systems and the possible conversion to a SR. Once the information is normalized it will be published to the Java Open Enterprise Service Bus (using the Direct Channel connection) and will be subscribed by different destinations (bindings) or used in other business logic (service engines). It will be possible to transform the medical reports in different file formats, send them by email or exhibit them in a HL7 viewer.

Once the information is encoded we need to describe the natural language, Portuguese, medical terms by a code. A code value that can be unambiguously identified enabling data mining [4] making easier the indexing and selective retrieval, without having to arrange to NLP - Natural Language Parsing [6].

The codification could be done using the SNOMED CT coding system but this is a proprietary medical term library. Thus we decided to code the medical terms with internal mapping defining a *CODE* service for this codification.

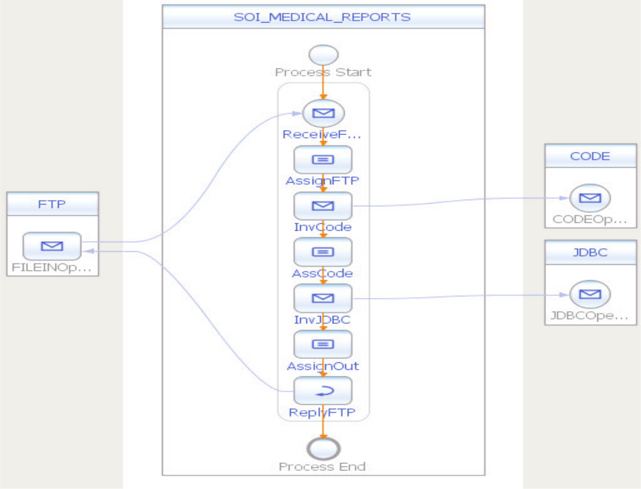

Fig. 5. BPEL Medical Reports Integration example.

The major codification difficulty is the negation terms common in natural language and in the medical reports. According to [21] the referring physicians that just say “normal” are considered as somehow inferior or incomplete. Therefore the referring physicians can also list things that are not found, even on entirely normal examinations. For example consider a report where the referring physician does not find any atrophy but found a stroke. Using our internal codification we unambiguous indentify the medical terms and encode also negation terms. For example the *CODE* service can receive the following (part) message:

```
<finding att="Not Found">atrophy</finding>
   <finding att="Found">stroke</finding>
```

The mapping that we have to these medical terms is:

Table 1. Internal map Medical Terms (part).

Medical Term	Found	Not Found
Atrophy	110	110N
Focal atrophy	1101	1101N
Stroke	120	120N
Lacunar stroke	1201	1201N

So, after receiving that message and in accordance to Table 1 the *CODE* service will reply the following (part) message:

```
<finding>110N<finding>
<finding>120</finding>
```

Therefore, the information saved on the database will facilitate the use of YALE since it is now identified by unambiguous and unique code without the usage of natural language and encoding the negation terms.

The integration process here presented is one of many possible others. For example, we can have two parallel paths for this integration process. One for taking care of the message codification before being saved in the database. And another to save the information as it is in the database without coding. Facilitating data mining and maintain history of the processed information.

After this integration process using Data Mining techniques and Machine Learning algorithms that YALE provides we can discover association rules between different medical reports. This technique of Data Mining makes it possible to identify patterns in large databases. This identification of patterns helps the gathering and interpretation of obtained results, to acquire the specific knowledge to a conclusion or assumption for the case study.

The YALE association rules discover model uses a database connector, a pre-processing operator chain, a learner unsupervised item set and association rule generator algorithm. The major pre-processing operators are the frequency discretization, filter operator nominal to binominal operator and missing value replenishment. The frequency discretization, discretizes numerical attributes by putting the values into bins of equal size. The filter operator nominal to binominal creates for each possible nominal value (YALE considers the negated terms as nominal) of a polynomial attribute a new binominal (binary) feature which is true if the example had the particular nominal value. These pre-processing operators are necessary since particular learning schemes can not handle attributes of certain value types. The next operator is the frequent item set mining operator *FPGrowth* [22]. A major advantage of *FPGrowth* compared to *Apriori* algorithm is that it uses only 2 data scans and is therefore often applicable even on large data sets [23]. This operator efficiently calculates attribute value sets often occurring together. From these so called frequent item sets the most confident rules are calculated with the association rule generator.

4.1 Case study

In our case study we have three different types of CT reports from three different physicians, available in html format. Only 100 records were kindly provided by CIT (Centro de Imagem da Trindade) for this analysis and all the records are from different patients without repetition. The information of these reports is previously extracted and normalized in the first step of the integration process to facilitate the codification process. The age of the patients were grouped by the following range:

- Baby [0;4]
- Child [4;12]
- Young [12;26]
- Adult [27;65]
- Senior [65;]

Since we are dealing with natural language, in Portuguese, there are many possible ways to represent the same word, with or without accents and lower or upper case characters we transform all characters to lower and convert the accents to the corresponding ASCII character, and e.g., the “Ç” convert in a “c”. In general, these medical reports contain a patient module: patient id, sex and age. Also indicates the number of series, modalities and notes. Follow that we have the protocol(s) and the finding(s) with conclusion or not. So each item set consist of one or more patient module, series, modalities, notes, protocols, finding and conclusion codes. With this information we try to find some association rules between the followed procedures and the diagnosis. Despite the number of medical reports, each one can have many types and more than one note, protocol or finding that increase the complexity for the find association rules.

4.2 Results

Using the YALE association rules discover model for the case study presented we obtained for example some interesting frequent item sets from our case study:

- 75% of the patients are adult women
- 86% are normal
- 69,4% are normal when not found median deviation (901N) and found permeable ventriculo-cisternal system (80111).

Despite finding many association rules for the medical procedures and respective diagnosis that can be considered expected we selected two particular rules:

Table 2. Association Rules Medical Reports (part).

Rule	1	2
Premises	Notes=30	protocol= 20111211
Conclusion	protocol= 20111211	conclusion= C10
Support	61%	75%
Confidence	81,5%	96,4%

The rules presented in Table 2 show us a curious fact that the medical report notes can have an important role in the definition of the protocol and indirectly influence the diagnosis. As you can see Rule 1 (notes=30 => protocol=20111211 [61%; 81,5%]) we have 61% of our samples have low back pain and the protocol was axial plans contiguous of 10mm parallels to the orbit-meatal plan and the confidence for this is 81,5 %. For Rule 2 (protocol1=20111211 => conclusion=C10 [75%; 96,4%]) we have that the protocol axial plans contiguous of 10mm parallels to the orbit-meatal plan in 75% of cases we will have a normal exam with 96,4 % of confidence.

5 Conclusion

The knowledge discover methodology for medical reports supported by a service oriented integration architecture here presented is a response to problems raised from the legacy medical report processing to enable knowledge discovery using industry, business and medical open standards.

A solution for typical problems presented by legacy medical reports like ambiguous terms of natural language, lack of structure, hard inter-application exchange due to business logic and trans-coding of incompatible protocols and messages formats was presented.

The legacy reports encoded based on the physician templates proved to be an excellent parsing format because all physicians have there own template making easier the conversion to a structured report format. In a first approach we tried to apply XSLT transformation to the OFFIS XML Schema *dsr2xml.xsd* [24] but the legacy medical reports lacks of mandatory information required by OFFIS.

The decision to code the medical terms with internal mapping defining a service for this codification without using a coding system like SNOMED CT solves the encoding negation problem using a specific code concept name or value. This concept of negation is crucial because it's equally important to be able to say "no stroke" as it is to say "stroke".

The results obtained from information integration shows some obvious results like that bone fractures are more common in older people or the normal exams are more common in younger people. However this result also allowed concluding that the notes taken before the realization of the exam influenced the protocol and this in turn the diagnosis.

Currently we are working on:

- OFFIS *dsr2xml.xsd* - mapping the entire medical report information to this structure. For future conversion the contents of a DICOM Structured Reporting (SR) document (file format or raw data set) to XML;
- DICOM Open-ESB binding component development – at the moment the communication is only possible by TCP/IP binding;
- Intelligent Event Processor Open-ESB – at the moment the data pre-processing treatment is done with YALE. Using the IEP withdraw the YALE load;
- Geographical Data Mining - enabling epidemic studies.

Acknowledgements

We are indebted to CIT – Centro de Imagiologia da Trindade, for their help in terms of anonymous medical reports, experts, technicians and machine availability.

References

1. *Digital Imaging and Comunications in Medecine.* Retrieved August 1, 2008, from http://medical.nema.org/.
2. *Health Level Seven*. Retrieved August 1, 2008, from http://www.hl7.org/.
3. Dreyer, Keith J., Mehta, Amit, Thrall, James H. *PACS: a guide to the digital revolution*, Springer, New York (2002).
4. Noumeir, Rita. *Benefits of the DICOM Structured Report*, Journal of Digital Imaging, 295-306, Springer (2006).
5. Clunie, David A. *DICOM Structured Reporting and Cancer Clinical Trials Results*, Cancer Informatics 2007:4 33-56 (2007).
6. Langlotz, C.P. *Automatic Structuring of Radiology Reports: Harbinger of a Second Information Revolution in Radiology*, Radiology 224:5-7 (2002).
7. Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J.W., ML Hage, Hammond, W. E. *Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse,* Proceedings of the AMIA Annual Fall Symposyum, Vol. 101, No. 5 (1997).
8. Behlen, Fred M. *DICOM Structured Reporting and the CDA Tutorial*, HL7 International CDA Conference, Berlin (2007).
9. Ten-Hove, Ron. *Using JBI for Service-Oriented Integration (SOI).* Retrieved July 1, 2008, from https://open-esb.dev.java.net/public/whitepapers/JBIforSOI.pdf.
10. Open-ESB. *The Open Enterprise Service Bus.* Retrieved July 1, 2008, from https://open-esb.dev.java.net/.
11. Mierswa et all. *YALE: Rapid Prototyping for Complex Data Mining Tasks*, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06).
12. Witten, Ian H. and Frank, Eibe. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco (2005).
13. SNOMED CT. *Systematized Nomenclature of Medicine-Clinical Terms.* International Health Terminology Standards Development Organisation Retrieved August 1, 2008, from http://www.ihtsdo.org/snomed-ct/.

14. Bortoluzzi, Mariana Kessler. *Desenvolvimento e Implementação de um Editor de Documentos Estruturados no padrão DICOM Structured Report.* Universidade Federal de Santa Catarina, Florianópolis (2003).
15. Kong, A., Barnett, G., Mosteller, F., et al. *How medical professionals evaluate expressions of probability*. N Engl J Med; 315:740–744 (1986).
16. Lussier, Y. A., Shagina, L., Friedman, C. *Automating SNOMED coding using medical language understanding: a feasibility study*. Proc AMIA Symp, 418–422 (2001).
17. Holman, B., Aliabadi, P., Silverman, S., et al. *Medical impact of unedited preliminary radiology reports.* Radiology 191:519–521 (1994).
18. Naik, S., Hanbidge, A., Wilson, S. *Radiology reports: examining radiologist and clinician references regarding style and content.* AJR (American Journal of Roentgenology) 176:591-598 (2001).
19. Ten-Hove, Ron, Walker, Peter. *Java Specification Requests 208: Java™ Business Integration (JBI).* Retrieved July 1, 2008, from http://jcp.org/en/jsr/detail?id=208.
20. Raj, Gopalan Suresh, et all. *Implementing Service-Oriented Architectures (SOA) with the Java EE 5 SDK.* Retrieved July 1, 2008, from http://java.sun.com/developer/technicalArticles/WebServices/soa3/ImplementingSOA.pdf.
21. Clunie, David A. *DICOM Structured Reporting*, PixelMed Publishing, Bangor, Pensilvania. ISBN 0-9701369-0-0 (2001).
22. Han, J., Pei, J., Yin, Y. *Mining frequent patterns without candidate generation.* In Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00), pages 1-12, Dallas, TX (2008).
23. YALE. *RapidMiner Developer Guide*, Rapid-I, Dortmund, Germany. Retrieved July 1, 2008, from http://downloads.sourceforge.net/yale/rapidminer-4.2-tutorial.pdf.
24. Kuratorium OFFIS. *dsr2xml: Convert DICOM SR file and data set to XML.* Oldenburg, Germany. Retrieved July 1, 2008, from http://support.dcmtk.org/docs/dsr2xml.html
25. Tianxia Gong, Chew Lim Tan, Tze Yun Leong, Cheng Kiang Lee, Boon Chuan Pang, C. C. Tchoyoson Lim, Qi Tian, Suisheng Tang, Zhuo Zhang, "Text Mining in Radiology Reports," icdm, pp.815-820, 2008 Eighth IEEE International Conference on Data Mining, 2008.