In Patel V, Rogers R and Haux R (Eds.) 10th World Congress on Medical Informatics (MedInfo'2001), September 2 – 5, London. IOS Press, Amsterdam.

Extracting Clinical Cases from XML-Based Electronic Patient Records for Use in Web-based Medical Case Based Reasoning Systems

Selvakumar Manickam, Syed Sibte Raza Abidi

School of Computer Sciences,Universiti Sains Malaysia,Penang,Malaysia.

Abstract

Development and usage of Case Based Reasoning (CBR) driven medical diagnostic system requires a large volume of clinical cases that depict the problem-solving methodology of medical experts. Successful usage of CBR based systems in healthcare is constrained by the need for a continuous supply of current and correct clinical cases (in an electronic medium) from medical experts. To address this constraint we present a strategy to pro-actively transform generic Electronic Patient Records (EPR) to Operable CBR-oriented Cases (OCC) that are compliant to specialised CBR-based medical systems. EPR-OCC transformation methodology is based on XML parse-trees, Unified Medical Language Source (UMLS) meta-thesauri and medical knowledge ontologies. The featured work involves the implementation of a Java-based computer system for the automatic transformation of XML-based EPR—originating from heterogeneous EPR repositories accessible over the Internet/WWW—to specialised OCC that can then be seamlessly incorporated within Intelligent CBR-based Medical Diagnostic Systems.

Keywords

Case-Based Reasoning; Electronic Patient Records; Knowledge Discovery; XML; Medical Ontology; Internet

Introduction

Case-Based Reasoning (CBR) techniques [1] now are routinely used in 'intelligent' medical decision-support systems for a variety of tasks including diagnostic support, WWW-based patient-centric consultation, health planning and so on [2][3][4]. Functionally speaking, CBR based medical systems provide 'analogy-based' solutions/diagnosis to clinical problems by manipulating knowledge derived from similar previously experienced situations, called *Cases*. According to the CBR methodology, a new problem is solved by finding similar past cases, and reusing their problem-solving strategy to derive a solution to the new, yet similar, problem situation. Note that, each case is described by a set of case-defining attributes, and is associated to a solution (or decision) suggested by a medical practitioner [5].

Development and usage of a CBR-based medical diagnostic system requires a large volume of diagnostically unambiguous clinical cases, which are to be provided by acknowledged medical experts. This implies the need for (i) a continuous supply of up-to-date and correct clinical cases from medical experts; (ii) collection of a wide variety of clinical cases originating from more than one medical expert and medical site; (iii) expert-level validation of the content of the clinical cases to determine relevance and accuracy; (iv) manual transcription of clinical cases to the native information format of the CBR system (i.e. case structure); and (v) terminological and conceptual standardisation of heterogeneous clinical cases to a common information standard. Despite the natural propensity of CBR technology towards solving medical diagnostic problems, it can be argued that the efficacious utilisation of CBR-based medical systems is comprised due to the highlighted constraints. Indeed, it is a daunting and time-consuming undertaking on part of medical experts, who are already pressed for time and resources, to satisfy the above constraints in order to improve the efficacy of CBR-based medical diagnostic systems.

To address the abovementioned problem, we propose a strategy for (a) the automatic collection of clinical cases culled from general-purpose Electronic Patient Records (EPR); and (b) the transcription of the collected clinical cases to a dedicated case format—*Operable CBR-oriented Cases* (OCC)—compliant to a CBR-based medical system. The rationale for our approach derives from the fact that EPR can be regarded as an 'alternate', yet implicit, source of clinical problem-solving knowledge, systematically compiled by physicians during episodic visits by patients [6][7]. Transactional EPR constitute physician-generated descriptions of the diagnostic process, hence OCC derived from EPR is deemed equivalent to leveraging the collective expertise of physicians within the featured sample. Note that typically an EPR comprises the kind of information—i.e. longitudinal patient history, illness-related symptoms & signs, pathological finding, diagnosis (or prognosis) by physicians

and a treatment plan—that is included in the definition of OCC.

In this paper we will discuss the methodological issues pertaining to the automatic transcription of heterogeneous EPR to OCC, together with the computational implementation of an experimental system to perform the same. The featured computer system is able to autonomously transform generic eXtensible Markup Language (XML) based EPR—originating from heterogeneous EPR repositories accessible over the Internet/WWW—to specialised OCC that can then be seamlessly incorporated within Intelligent CBR-based Medical Diagnostic Systems.

Methodology: EPR-OCC Transformation

In our work the transformation of an EPR to OCC is not seen as a straightforward mapping of attributes-value pairs from the EPR to OCC. We argue that such an approach would lead to complexities due to the presumed heterogeneous origin of EPR, whereby structural, terminological and conceptual differences may exist both across different EPR schemes and with respect the a posteriori specified OCC standards. For that matter a correspondence between EPR and OCC attribute-value pairs is established based on equivalence within a multi-layer descriptive framework, with distinct analysis at the: (1) object, (2) terminology, and (3) concept levels; as illustrated in Figure 1.

Analysis at the (most basic) object-level is facilitated via the EPR descriptive framework which utilises: (1) Health Level (HL) 7 for the transactional logic context, and (2) XML for the actual document-object. The basic idea is to use the HL7 message-types for construction of XML Document-Type Definitions (DTD), which are essentially templates against which EPR document-objects can be parsed for data-structural correctness. Application of the XML framework allows for error detection-correction, and (more importantly) straightforward comparative analysis between EPR document-objects and OCC meta-structures.

The heterogeneous origin of the EPR leads to the inevitable usage of functionally synonymous terms to denote the same clinical concept. Hence, it is necessary to establish terminology-level standardisation. This is achieved by the deployment of a medical meta-thesaurus—in our case ULMS—that constitutes the system-supported medical vocabulary, and enables operationally flexible many-to-one mappings between EPR-OCC attribute-value pairs. Such an analysis is equivalent to the aggregation of individual EMR DTDs into a thesaurus-specified meta-DTD.

Finally, concept-level equivalence of medical concepts given in the EPR with respect the standard OCC concept vocabulary is achieved via medical concept-specific ontologies. The medical ontology-set defines the generalised conceptual framework associated with EPR and OCC attributes, and is therefore useful for the abstract transformation of 'non-standard' concepts given within EPR to the standard OCC conceptual framework. Note that the determination of concept-level equivalence, within an ontology-specified medical context, thereby allowing for EPR-to-case information transfer beyond the (necessarily) crisp boundaries defined by semantic equivalence analysis.

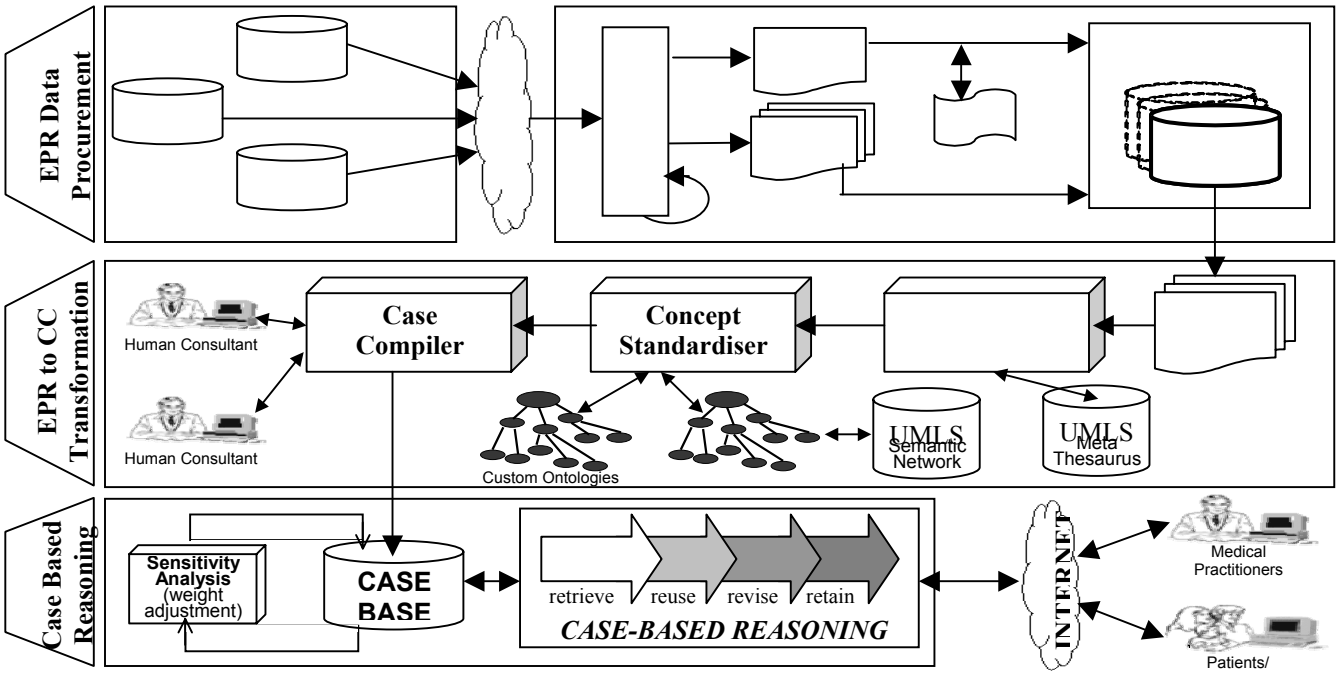

Figure 1 - Functional architecture that comprises 3 main phases and their respective modules

Method: EPR-OCC Transformation Stages

Functionally, EPR-OCC transformation is carried out in three stages: Stage1-EPR procurement over the Internet; Stage2-EPR transformation to specialised OCC; and Stage3-Web-enabled diagnostic services using the derived OCC. To support the above functionality we have implemented a computer system comprising three functionally distinct layers. Each layer comprises a number of modules, each responsible for a certain task. Figure 1 shows the functional architecture and the process workflow of the system. We briefly explain the functionality involved in the three stages.

Phase 1 : Internet Mediated EPR Procurement

This phase is responsible for (i) procuring EPR from diverse sources over the Internet; (ii) cleansing the procured EPR; and (iii) uploading the cleansed EPR in an intermediate database for subsequent activities.

(a) EPR procurement from heterogeneous, Internet-enabled EPR repositories is managed by a dedicated computer server whose responsibility is to periodically monitor the selected (XML-based) EPR repositories for new EPR; and if new EPR are found to extract them.

(b) The procured EPR are next cleansed by way of removing undesirable information, for instance patient name, address, next-of-kin, etc. The list of undesirable information fields is maintained within the system. The tags of the XML document—the EPR representation format—is scanned against the said list and matching tags are removed from the XML document in order to cleanse the EPR.

(c) The cleansed EPR is finally uploaded into the intermediate database. In case the EPR is procured from a new EPR repository, we need to dynamically map the EPR structure to the intermediate's database structure which in fact models the OCC's format. This direct mapping is achieved by exploiting the XML document's *Document Type Definition (DTD)*—the logical structure of the EPR—defining the EPR's semantic tags, which are then used to perform a mapping from the EPR's DTD to the OCC structure. Once the database structure has been created, the data corresponding to each XML tag is then uploaded into the intermediate database.

Phase 2 : EPR to Specialised OCC Transformation

This phase is responsible for the automatic transformation of EPR, collected over the Internet from diverse EPR repositories, to specialised OCC, as per the specification of the OCC case-base. In Figure 1, the intermediate layer shows the various modules that are used during phase 2, and here we will briefly discuss their functionality.

The data entities involved in the EPR-OCC transformation are as follows: (a) The entire set of OCC Attributes (say OCCA); (b) The range of allowed OCC Values (say OCCV) for each OCC attribute; The entire set of the EPR Attributes (say EA) and EPR Values (say EV). Additional resources used to facilitate the transformation process are: (a) OCC MetaMap—the description of the OCC structure and allowed contents; and (b) Transformation Map—a dynamically generated text file to record all the transformations determined by the system. The EPR-OCC transformation process is carried out according to the following scheme:

1. *Direct Mapping of OCCA to EA*: For each OCCA a mapping to a matching EA is determined by checking the EAs against the OCC MetaMap. IF an exact match is found then it is recorded in the Transformation-Map as an *Exact Match*.

2. *Vocabulary Mapping of OCCA to EA*: Here we seek to establish vocabulary-level equivalence between OCCA and EA—i.e. convert the EA's vocabulary to the designated OCC vocabulary standards as stated in the *OCC-MetaMap*. The underlying idea is to determine whether the target 'un-matched' OCCA is represented in the source EPR using some variant terminology. This is achieved by the *Vocabulary Standardiser* module, which uses the UMLS meta-thesaurus as the central resource to standardise the medical terminology between the EA and OCCA. For example, suppose that the OCC MetaMap contains an OCCA named *Symptom.* However the EPR's DTD does not have an EA with the same name, but an EA with the name G*eneral Manifestation of Disorders*. By using UMLS meta-thesaurus, it is found that G*eneral Manifestation of Disorders* is a synonym of *Symptoms*, hence the said OCCA can be mapped to the EA, and the event is recorded in the transformation-map as a *Synonymous Match.*

3. *Ontological Mapping of OCCA to EA*: In case, the above-mentioned vocabulary mapping fails to establish equivalence between a given OCCA/OCCV and available EA/EV, then we attempt to establish equivalence at the conceptual level. A *Concept Standardiser* module, employing medical knowledge ontologies, is then used to standardise medical concepts between the EPR and OCC MetaMap. The rationale for the use of medical ontologies is that a medical knowledge ontology can determine the ontological equivalence between two concepts, whereby one medical concept can be determined to be the generalisation or specialisation of the other medical concept. Therefore, by using a taxonomic description of medical concepts—i.e. an ontology—we are able to establish conceptual equivalence between two concepts. In our work, two different types of medical ontologies are used: (1) standard medical ontologies available within the medical literature and (2) ontologies pro-actively derived by us

from medical coding schemes such as ICD10, MSH99 and so on (see Figures 2 and 3). If a successful ontological match is achieved then it is recorded in the transformation-map as an *Ontological Match*.

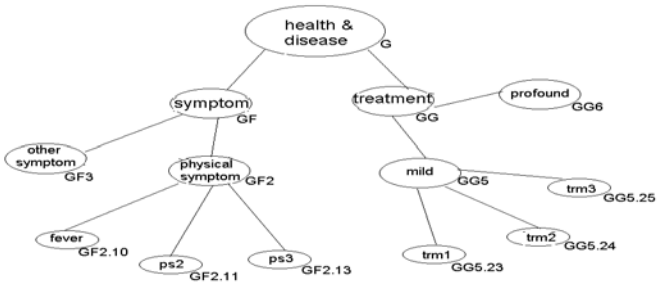

Figure 2 - Medical ontology for the concept *fever* derived from AOD95 medical coding scheme

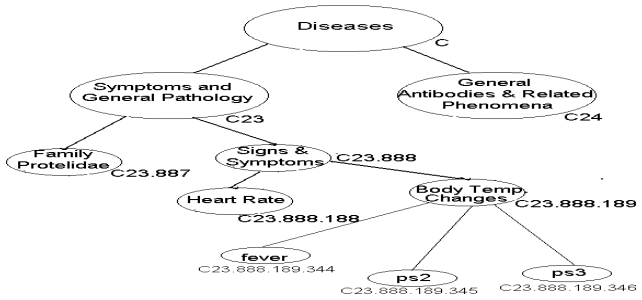

Figure 3 - Medical ontology for the concept *fever* derived from MSH99 medical coding scheme

4. *Manual Mapping of OCCA to EA* : Assuming that all the previous method of transformation fails, then the system is designed to consult a human expert to match the OCCA/OCCV based to the standard coding provided. If a successful manual match is achieved then it is recorded in the transformation-map as a *Manual Match*. However, at this stage we do not allow manual mapping.

Finally, the C*ase Compiler* module ensures the completion of the 'automatic' EPR-OCC transformation. The *Transformation Map* details the derived transformations of the EPR attributes/values to corresponding OCC attributes/values. The case compiler module makes use of the internal *Transformation-Map* to complete the transformation of the EPR to a specialised OCC. For the current version of the system, the user is provided an on-screen explanation to validate the EPR-OCC transformation, if desired. Figure 4 shows the phases of an exemplar transformation of an EPR to an OCC.

Phase 3 : Web-based Case-Based Reasoning

This phase entails activities pertaining to the usage of the derived OCC, using CBR algorithms [1], for WWW-based decision-support services. The front-end CBR system, comprising 3 independent modules, is central to the activities reported here.

Automatic OCC Attribute's Sensitivity Assignment: Typically in CBR system the case defining attribute's relative importance—i.e. sensitivity (or weight) towards the eventual outcome—is determined manually by system designers. This approach, though valid, is not truly reflective of the knowledge encapsulated in the entire corpus of cases (i.e. the case-base). In our work, we use a neural network to analyse the relative sensitivity of each OCC attribute towards the corresponding outcome [8]. This inductively derived sensitivity measures are an actual reflection of the input-output mapping (i.e. the system's intrinsic knowledge) as documented in the case-base. The sensitivity of the OCC attributes is adjusted after each 10% growth of the case-base. We argue that the feature of our inductive approach for attribute sensitivity assignment is that it systematically analyses the entire corpus of cases in determining the weights of the OCC attributes.

Case-Based Reasoning (CBR) Engine employs standard CBR algorithms to cater for the four main CBR activities—Retrieve, Reuse, Revise and Retain. The CBR engine receives a request for decision-support in terms of the description of the clinical case in the OCC format. Next, it will retrieve the best-matching past-cases to formulate a solution for case at hand.

Web Based GUI provides an interface for users to interact with the CBR system (for diagnostic support) via the WWW. Typically, physicians will interact with the CBR system through their web browsers (by providing the CBR system's website address). A web-based form will be presented to users, asking them to fill in certain details about the case at hand. In turn, the CBR system will provide diagnostic support, again delivered over the WWW. In this way, physicians can consult the system for second opinion of confirmation and users can use this system as their informal consultant.

Implementation Details

The CBR system is largely implemented using the Java 2.0 programming language, and is hosted on a server running Windows NT. The case-base and the intermediate database(s) are implemented using the Microsoft SQL Server 7.0 (MSSQL) database. The UMLS Metathesaurus and domain ontologies are converted from native text files into indexed segments of MSSQL database for faster queries. Data exchange with donor EPR repositories is achieved using JDBC (Java Data Base Connectivity), which ensures dynamic connections with a wide range of database platforms. EPR (given as XML documents) based information extraction vis-à-vis XML document parsing is achieved using the Microsoft XML Parser. The web-based GUI is implemented using Java servlets in conjunction with Microsoft IIS web server.

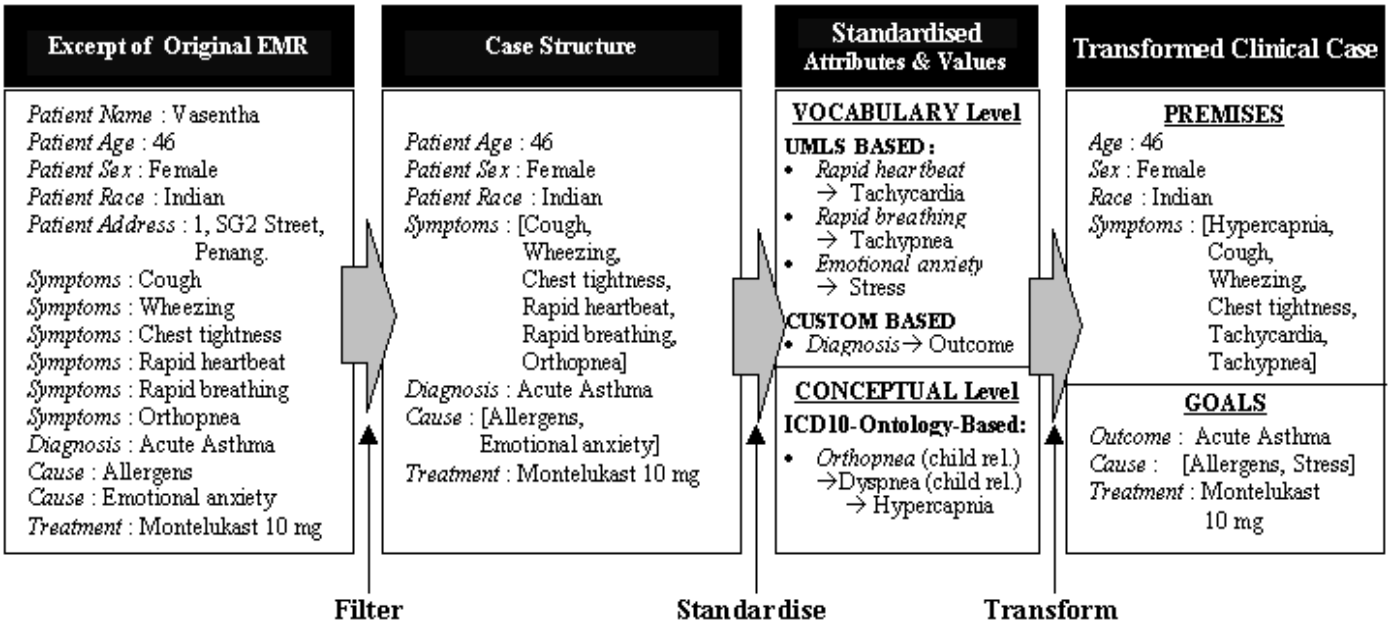

Figure 4 - Phases of an exemplar transformation of an EPR to a OCC

Conclusion

In our work, we have managed to leverage upon 'information rich' EPR, accessible over the Internet, to enhance the (medical) knowledge of traditional medical case-based reasoning systems. In this way, we have presented a novel facet and utility of routinely collected EPRs, whereby they can be transformed from mere information resource to a diagnostic decision-support resource. We have managed to demonstrate the efficacy of (a) XML as a representation language for medial information/data and (b) medical knowledge ontologies for conceptual mapping. Nevertheless, we need to add that due to the confidential nature of EPR, it is extremely difficult to find Internet-accessible EPR repositories. In this operational scenario, the success of our approach largely depends on the willingness of hospitals/EPR owners to participate in this program vis-à-vis pooling their patient information resources, i.e. EPR, and in turn benefiting from a 'rich' case-base for value-added diagnostic support services. We are working along these lines and anticipate to form a core group of EPR-donor hospitals.

References

[1] Aamodt A, and Plaza E. Relating case-based reasoning: Foundational issues, methodological variations and system approaches, *AI Communications 7, 1*, 1994.

[2] Bradburn, C., and Zeleznikow, J.,The application of case-based reasoning to the tasks of health care planning. In Wess, S.; Althoff, K. D.; and Richter, M. M., eds., *Topics in Case-Based Reasoning: First European Workshop, 365--378.Berlin: Springer-Verlag*., 1994.

[3] Mariuzzi, G., A. Nombello, L. Mariuzzi, P. W. Hamilton, J. E. Weber, D. Thompson, and P. H. Bartels. Quantitative study of ductal breast cancer --- patient targeted prognosis: an exploration of case based reasoning. *Pathology research and practice 193, 535—542*, 1997.

[4] Dysmorphic Syndromes. *Artificial Intelligence in Medicine 6*, 1994.

[5] Kolodner JL. Case-Based Reasoning, *Morgan Kaufmann*, San Mateo, 1993.

[6] Althoff K, Bergmann R. Case-based reasoning for medical decision support tasks: The INRECA Approach, In : *Artificial Intelligence in Medicine Journal, Vol. 12,* 1998.

[7] Jaulent M, Le Bozec C. Case based diagnosis in histopathology of breast tumours. In Cesnik B, McCray A & Scherrer J (Eds), *MedInfo'98*, IOS Press, Amsterdam, 1998.

[8] Poh HL, Yao JT, Tan CL, Teng. "Forecasting And Analysis of Marketing Data Using Neural Networks", *Journal of Information Science and Engineering, Vol.14*, 1998.

Address for correspondence

Health Informatics research Group, School of Computing Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia.

Email: selva@cs.usm.my