Scenario-Oriented Information Extraction from Electronic Health Records

Anis Yousefi, Negin Mastouri[*], Kamran Sartipi
Department of Computing and Software
McMaster University
Hamilton, ON, Canada
{yousea2, mastoun, sartipi}@mcmaster.ca

Abstract

Providing a comprehensive set of relevant information at the point of care is crucial for making correct clinical decisions in a timely manner. Retrieval of scenario specific information from an extensive electronic health record (EHR) is a tedious, time consuming and error prone task. In this paper, we propose a model and a technique for extracting relevant clinical information with respect to the most probable diagnostic hypotheses in a clinical scenario. In the proposed technique, we first model the relationship between diseases, symptoms, signs and other clinical information as a graph and apply concept lattice analysis to extract all possible diagnostic hypotheses related to a specific scenario. Next, we identify relevant information regarding the extracted hypotheses and search for matching evidences in the patient's EHR. Finally, we rank the extracted information according to their relevancy to the hypotheses. We have assessed the usefulness of our approach in a clinical setting by modeling a challenging clinical problem as a case study.

1 Introduction

One of the most important challenges in integrated healthcare delivery is providing comprehensive, reliable, relevant, and timely patient information to health providers [5]. On a daily basis, physicians encounter complex clinical scenarios that require: collecting initial data from patient, formulating diagnostic differentials (hypotheses), collecting evidences from patient's health record relevant to the given hypotheses, investigating the hypotheses (paraclinics, interventions and follow-up actions), diagnosis and treatment.

The accuracy and promptness of hypothesis formulation play a significant role in effective clinical management.In this context, EHR is a powerful asset which ties together documentation of the patient visit (clinical information) and diagnosis and treatment procedures [10].

Retrieval of scenario specific information from an extensive EHR record is a tedious, time consuming and error prone task. One approach to summarize EHR data is equipping EHR systems with reporting facilities which represents the information using meaningful visualizations such as graphs or tables [2]. Organizing EHR into a reasonable structure also assists in more efficient browsing of information [14].

In this paper, we propose a model and technique for automatic extraction of situation-specific health information from patient EHR. We use concept lattice analysis to develop diagnostic hypotheses followed by discovery of relevant information from patient's EHR. We also apply a ranking mechanism to indicate the degree of relevancy of each information item to the clinical scenario.

Specifically, we consider widely-accepted *clinical syndromic approach* to verify the proposed model. Syndrome is a set of signs and symptoms which tend to occur together and reflect the presence of a particular disease. There are a large number of major clinical syndromes that can be modeled according to our proposed technique. As a case study, we modeled syndromic approach to *Fever of Unknown Origin* (FUO) due to its importance and complexity in medical domain. There are a large number of cases with FUO which are remained undiagnosed despite hospitalization, costly paraclinic requests and invasive procedures [3].

The rest of this paper is organized as follows. Section 2 provides an overview of related work. Section 3 presents the model and mechanism of relevant information extraction. In section 4, we study the applicability and usefulness of our approach in clinical settings by examining the case of FUO. Discussion and future work is provided in Section 5.

2 Related work

Organizing EHR information into a meaningful structure assists physicians in finding relevant patient information. From early 1980s commercial and non-commercial healthcare organizations have developed their own proprietary structure for representing electronic health records. These

[*] MPH, M.D, Infectious & Tropical Disease Specialist

products vary from a simple *time-oriented* list of medical concepts [9, 8] to relatively sophisticated *problem-oriented* [17, 9] and *source-oriented* [14] medical records. In time-oriented structure, patient information is categorized into groups of events (e.g., lab results in a specific date). In problem-oriented structure, the information is grouped under one or more *problem headings*: Subjective, Objective, Assessment, and Plan (SOAP). In *source-oriented* structure, the content of a record is arranged according to the method by which the information was obtained (e.g., notes of visits, X-ray reports and blood tests).

Most EHR systems support query-response mechanism for information retrieval [16, 15]. In this mechanism, a domain expert is needed to specify the set of queries that care givers would require. To find relevant information, care givers use those predefined queries through system interfaces. Both browsing and query-response methods suffer from the fact that user must know what he/she is looking for and find the appropriate section or query interface to get the information.

In [5, 6, 7] an information retrieval system is presented which provides situation-specific EHR information related to the current activity within the patient's care workflow. The primary care workflow is modeled with an ontology, where for each activity in workflow relevant EHR headings and linguistic degree of relevance are pre-assigned.

In [13] adding knowledge to EHR records is proposed. This knowledge can be used by applications to provide patient-related functionality. We extend this idea by proposing the required domain models for relevant information extraction.

3 Scenario-oriented clinical information extraction

In this section, after defining a number of terms used in our proposed model, we discuss the scenario-oriented clinical information extraction process illustrated in Figure 1.

3.1 Definition of terms

The following terms are used in the proposed model.

Relevancy: measure of how "useful" the retrieved information is with regard to its application. In this work, we have considered the usefulness and relevancy of the extracted information in the process of diagnosis.

Scenario: a narrative that captures the interaction between a patient and a physician in a visit in terms of patient's symptoms and signs. In our approach, we represent an abstraction of a scenario as the set of *symptoms* and *signs*.

Hypothesis: a tentative explanation for the cause of a clinical scenario that can be tested by further investigation. Diseases that share the same set of signs and symptoms with a scenario are referred to as hypotheses of that scenario.

EHR item: we represent an abstraction of EHR as a set of $(item,\ value,\ date)$ tuples. EHR items are symbols representing clinical manifestations, predisposing factors and comorbidities, findings in physical examinations, laboratory information, imaging results, and other items which can be found in the EHR.

EHR category: we group EHR items into 7 categories: Demographic data, Past Medical History, Medications History, Allergy/Vaccination/Diet, Habitual History, Psychosocial History, Lab/Imaging/Procedure.

DiseaseAtt: a disease attribute refers to any EHR item or sign/symptom.

Evidence: an $(item,\ value,\ date)$ tuple from EHR which *strengthens* or *weakens* a hypothesis.

3.2 Proposed model

The proposed model for extracting scenario-oriented clinical information consists of the following two phases as illustrated in Figure 1:

Phase 1 (*off-line processing*): the relationships between diseases, symptoms/signs (*SymSign*) and EHR items are modeled as *Disease-Graph*. Then, we apply *concept lattice analysis* to represent Disease-Graph as highly associated groups of *Disease*'s and *SymSign*'s.

Phase 2 (*on-line processing*): consists of the following operations: i) extracting a set of probable hypotheses related to a specific scenario using the concept lattice analysis discussed in Phase 1; ii) indicating the set of relevant *SymSign*'s and *EHRiv*'s from the Disease-Graph; and iii) extracting set of matched *EHRiv*'s from the patient's EHR to discriminate among hypotheses.

3.3 Phase 1: off-line processing

The offline processing is independent of specific patient information and is comprised of two steps as follows.

Disease-Graph modeling

A *Disease-Graph*, as illustrated in Figure 1(a), is modeled as a weighted typed graph $G = (V,\ E,\ W)$ where the set of vertices V is the union of *Disease*'s, *SymSign*'s, and *EHRiv*'s. The set of edges $E \subseteq V \times V$ represents the relevancy of *SymSign*'s or *EHRiv*'s to the *Disease*'s. The set of weights W represents the degree of relevancy of the edges in E, where a weigh $w_{ij} \in W$ is a quantity that we assign to an edge to indicate the support of *SymSign*$_j$ or *EHRiv*$_j$ in the diagnosis of *Disease*$_i$. Weight w_{ij} is a combination of three weights: $w_{ij} = w_{ij_1} \times w_{ij_2} \times w_{ij_3}$.

w_{ij_1} indicates the importance of each EHR category in the process of diagnosis.

w_{ij_2} is assigned based on how much an specific *SymSign* or *EHRiv* contributes to the likelihood of a specific hypothesis based on expert opinion.

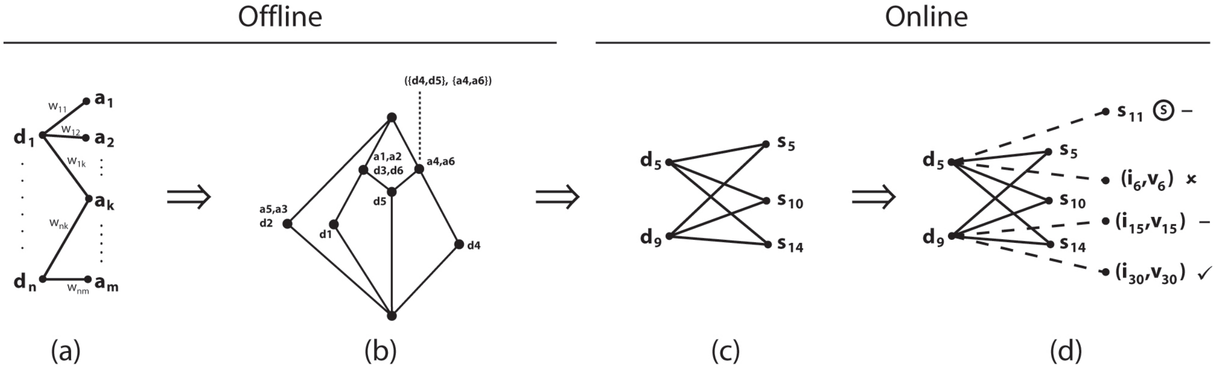


Figure 1. Scenario-oriented clinical information extraction.

w_{ij_3} is assigned by reviewing best practice evidence based resources such as UpToDate [3]. In this approach, we use the available health records of patients as a source to measure how much a specific *SymSign* or *EHRiv* is related to a specific *Disease*. There are many ways to realize such a requirement. We can use the percentage of patients with a given symptom which have been diagnosed with a specific disease or the percentage of patients with a given disease which had a specific symptom in their problem lists. An example of the first case could be: 20% of patients with symptom s_1 are diagnosed with disease d_1, and 80% of them are diagnosed with disease d_2. An example of the second case could be: 80% of patients with disease d_1 had symptom s_1 and also 80% of patients with disease d_2 had the same symptom.

Disease-Graph partitioning
In this step, we apply concept lattice analysis on the relations between *SymSign*'s and *Disease*'s from the Disease-Graph to generate groups of *Disease*'s and *SymSign*' with maximum association, as illustrated in Figure 1(b). Maximal association is a term borrowed from data mining referring to a set of objects all sharing a same set of attributes. In our approach, *Disease*'s represent objects and *SymSign*'s represent attributes.

In concept lattice [11], a binary relation between objects and their attributes is represented as a lattice which provides significant insight into the structure of the relation. Each node in the concept lattice is a *concept*. A concept is a maximal association where its object set is called *extent* and its attribute set is called *intent*. A concept lattice has the following characteristics:

- Each lattice node (i.e., a concept) is labeled with objects (*Disease*'s) and attributes (*SymSign*'s) except for the top and bottom nodes that may be unlabeled.
- Every object has all attributes that are above it in the lattice (directly above or separated by some links).
- Every attribute exists in all objects that are below it in the lattice (directly above or separated by some links).

An example of concept lattice is shown in Figure 3 of the case study in Section 4.

3.4 Phase 2: on-line processing

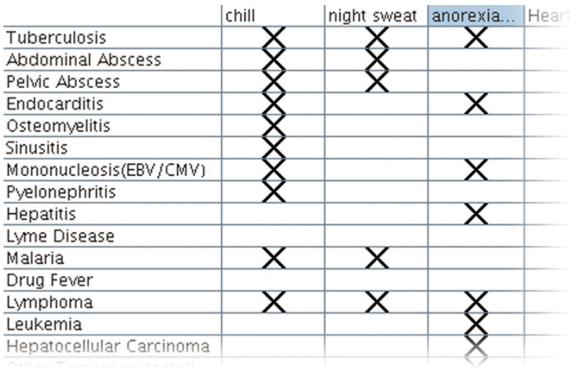
On-line processing refers to the process of extracting relevant information with regard to a specific scenario. It consists of three steps as follows.

Hypothesis discovery
In this step, we discover hypotheses related to a specific scenario using the result of concept lattice analysis. Where, the extent of a concept represents a set of Hypotheses and the intent of the concept represents the set of *SymSign*'s from the corresponding scenario. Figure 1(c) illustrates a concept with extent $\{d_5, d_9\}$ and intent $\{s_5, s_{10}, s_{14}\}$.

Evidence discovery
In this step, we use Disease-Graph to discover other *SymSign*'s and all *EHRiv*'s relevant to the hypotheses (diseases) which have been discovered in the previous step. Then, we search EHR for those $(item,\ value,\ date)$ tuples whose $item$ part is the same as the item part in an *EHRiv*. Next, we compare the $value$ part of the EHR item with the value part of the respective *EHRiv*. There are two possible results. In the case of a match, the EHR item *supports* its respective hypothesis. In the case of a mismatch, the EHR item *weakens* its respective hypothesis. We also consider those *EHRiv*'s that we couldn't find in the EHR to indicate the need for further investigation by physician.

Figure 1(d) highlights *matches* with ✓, *mismatches* with $\times$, and *unfound items* with $-$. In this figure solid lines represent the relationship between *Disease*'s and the information we already know about them (i.e. initial symptoms and signs), dotted lines represent the information that we need to investigate, and Ⓢ indicates other symptoms and signs that are related to the *Disease*.

Calculating the degree of relevancy
Degree of relevancy indicates how relevant information is to a hypothesis. w_{ij} indicates the relevancy of information j to hypothesis i. The likelihood of a hypothesis in a scenario is also calculated by aggregating the weights of its supportive evidences: $w_i = \Sigma_{1 \leq j \leq m}(w_{ij})$



	chill	night sweat	anorexia...	Hear
Tuberculosis	X	X	X	
Abdominal Abscess	X	X		
Pelvic Abscess	X	X		
Endocarditis	X		X	
Osteomyelitis	X			
Sinusitis	X			
Mononucleosis(EBV/CMV)	X		X	
Pyelonephritis	X			
Hepatitis			X	
Lyme Disease				
Malaria	X	X		
Drug Fever				
Lymphoma	X	X	X	
Leukemia			X	
Hepatocellular Carcinoma			X	

Figure 2. Context table for FUO.

Note that we put some thresholds to translate quantitative values to qualitative terms *strong*, *medium-strong*, *medium*, *mild-medium* and *mild*.

4 Case study

We conducted a case study to indicate how our approach can be used in a real-world health care scenario. In this section, we illustrate how relevant information are extracted with regard to the current clinical scenario. In this case study, we modeled *Fever of Unknown Origin* (FUO)[1] as an example of a major clinical syndrome. Approaching a patient with FUO is one of the most challenging problems in medical science, mostly because of the large number of potential diagnoses [3, 12].

We have modeled 45 diseases and 64 common symptoms and signs (SymSign) associated with FUO from a highly cited medical reference by Mandell et al. [12]. Figure 2 illustrates the context table associated with our case study and represents relations between diagnostic hypotheses in FUO and their associated symptoms and signs. We employed the *Concept Explorer* tool [1] to generate and illustrate the concept lattice of different hypotheses and their corresponding symptoms and signs into 499 concepts.

In the following, we provide a sample clinical scenario to assess the proposed model: *A 68-year-old Spanish female presented with* ***anorexia***, ***malaise***, ***non-productive cough***, ***night sweats***, ***chill***, *and* ***daily fever*** *(temperature, 38.3°*C*-39.5°*C*) from 4 days ago. She recently moved to Canada and spoke English with difficulty and was not cooperative in giving a precise history. She was brought to clinic by her neighbor who was not aware of her past medical history, her medications and contact with animals or ill people. In her first visit, cardiovascular, respiratory, and breast examinations were unremarkable. She was diagnosed community acquired pneumonia by family physician who prescribed antibiotic medication for her. Over the following weeks her fever persisted. She was referred to specialist for further investigation of FUO.*

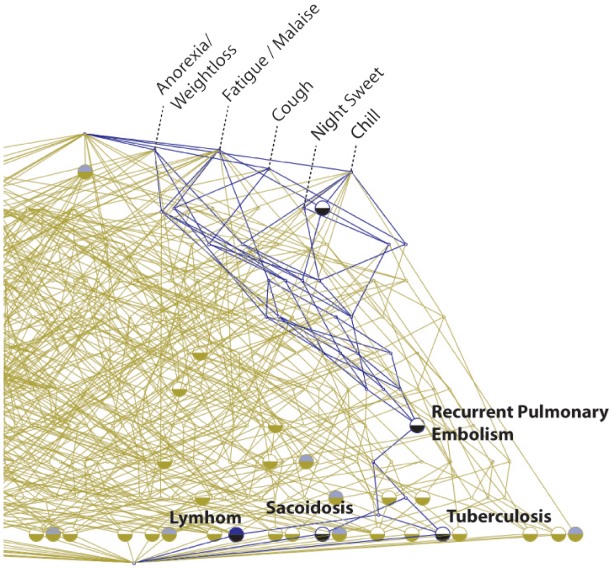


Figure 3. The concept associated with the scenario.

The set of symptoms and signs of the patient in the above scenario is {*anorexia, malaise, non-productive cough, night sweats, chill, fever*}. Using the concept lattice of FUO we identified {*Tuberculosis, Sarcoidosis, Recurrent Pulmonary Embolism, Lymphoma*} as possible diseases for the patient. Figure 3 highlights the concept associated with the scenario.

We consider the following weights for "DiseaseAtt to Disease" relations:

w_1 is weighted 0.7 for each of: past medical history of immobilization, surgery in previous three months, and diabetes. This weight is assigned according to Bates [4], where patient's history is the most important element of EHR contributing more than 70% to develop proper hypotheses. Physical examination contributes an additional 20-25%; and, laboratory testing contributes less than 10%.

w_2 is assigned based on *expert's opinion*. For example, in this case 0.7 is assigned to immobilization, 0.8 to surgery and 0.7 to diabetes.

w_3 is assigned according to evidence based resources (*meta analysis* results). For example, in our case 0.6 is assigned to immobilization, 0.6 to surgery and 0.4 to diabetes.

Using the proposed ranking system, the ranks of the hypotheses in our scenario is: 1) *Recurrent Pulmonary Emboli* (strong); 2) *Lymphoma* and *Tuberculosis* (medium-strong); 3) Sarcoidosis (medium);

At this point, we use Disease-Graph to discover the complete set of DiseaseAtt information related to the diseases in our set and their required values. Figure 4 illustrates a part of these DiseaseAtt information and the result of their matching against EHR.

Result for *Recurrent Pulmonary Emboli* is illustrated in table 1. For lack of space we have omitted the results of other hypotheses in this list.

As you see, this is the set of relevant information that physicians wish to see from EHR. This representation trig-

[1] FUO is defined as a body-temperature higher than 38.3°C that lasts more than 3 weeks with no obvious source despite appropriate medical investigations.

Table 1. Results for Pulmonary Embolism

Hypothesis	Supporting Evidences	Weakening Evidences	Unfound Items
Pulmonary Embolism (Strong)	Age (strong), Diabetes (strong), Surgery(strong), Immobilization(strong)	Normal ECG(mild), Normal CXR (medium)	D-Dimer test (strong), Lung perfusion Scan (strong) Pleuritic chest pain (strong)

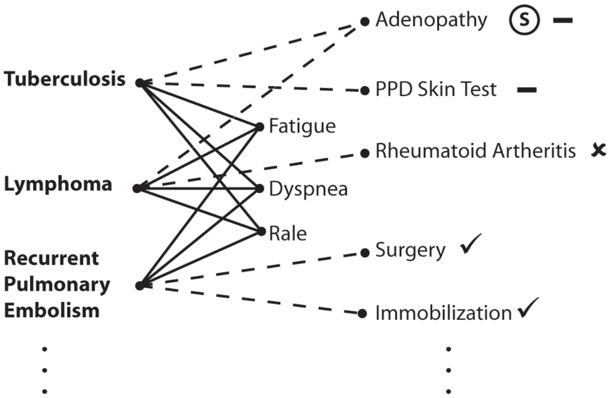


Figure 4. A part of disease-graph.

gers clinical memory, permitting the related knowledge to become accessible for reasoning. It helps physicians to be aware of clinical consequences, predisposing conditions, and other signs and symptoms that are neglected or will develop in the future. Moreover, it directs physicians to request essential cost-effective labs, images, and procedures and update EHR.

According to the information above, recent surgery and immobilization in the patient's EHR are two strong evidences for the diagnosis of Pulmonary Embolism.

5 Discussion and future work

In this paper, we presented a model and a technique to extract situation (scenario) specific information from EHR for the purpose of diagnosis. We provide a tailored view of EHR by dynamically grouping its information into relevant or irrelevant based on the current clinical scenario of the patient problem. The scenario, as a core, directs us to collect additional supporting or weakening evidences form EHR.

Unlike related approaches, the selection of what is relevant is done dynamically based on the internal knowledge of disease-attribute relations and extraction of hypotheses for a specific situation. Moreover, we differentiate between the information which support a hypothesis, the information which are against a hypothesis and the information we should investigate in order to decide about a hypothesis. Our ranking is also more specific than those we have seen in the related approaches, where the ranking is static. In our work, the assignment of a rank to an information item depends on it's context (the hypothesis to which it belongs).

As a next step in our approach we continue our simulation with a group of clinicians to clearly show our methods' potentials and feasibility. Also we intend to improve our ranking system by introducing dynamic edge weights.

References

[1] Formal concept analysis toolkit version 1.3. http://sourceforge.net/projects/conexp.

[2] HL7 EHR-S records management and evidentiary support functional profile. www.hl7.org.

[3] Uptodate online version 17.1. http://www.uptodate.com/.

[4] B. Bates. *Bates Guide to Physical Examination and History Taking,Barbara bates*. Churchill Livingstone, 2007.

[5] E. Bayegan and O. Nytro. A problem-oriented, knowledge-system. In *Proc. of MIE2002*, pages 272–276, 2002.

[6] E. Bayegan, O. Nytro, and A. Grimsmo. Ontologies for knowledge representation in the computer-based patient record system. In *Proc. of ICTAI 2002*, pages 114–121. IEEE Computer Society, 2002.

[7] E. Bayegan and S. Tu. The helpful patient record system: Problem oriented and knowledge based. In *Proc. of AMIA 2002 Annual Symposium*, pages 36–40, 2002.

[8] L. J. Bird, A. Goodchild, and H. Sue. Describing electronic health records using xml schema. In *XML Asia Pacific*, 2000.

[9] K. Hayrinen, K. Saranto, and P. Nykanen. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int. Journal of Medical Informatics*, 77:291–304, 2008.

[10] R. L. Koeller. It applications in healthcare: The electronic medical record. Master's thesis, University of Maryland, European Division, Bowie State University, 2002.

[11] S. Kuznetsov and S. Obiedkov. Algorithms for the construction of concept lattices and their diagram graphs. *Lecture Notes in Computer Science*, 2168:289–300, 2001.

[12] G. L. Mandell, J. Benett, and R. Dollin. *Principles and Practice of Infectious Diseases*. Churchill Livingstone, 2004.

[13] D. Ammon et. al. Developing an architecture of a knowledge-based electronic patient record. In *ICSE 08: Proc. of the 30th int. conference on Software engineering*, pages 653–660, 2008.

[14] H. J. Tange et. al. Medical narratives in electronic medical records. *Int. Journal of Medical Informatics*, 46(1):7–29, 1997.

[15] M. Croitoru et. al. Conceptual graphs based information retrieval in healthagents. In *CBMS '07: Proc. of the Twentieth IEEE Int. Symposium on Computer-Based Medical Systems*, pages 618–623. IEEE Computer Society, 2007.

[16] T. Austin et. al. Implementation of a query interface for a generic record server. *Int. Journal of Medical Informatics*, 77(11):754–64, 2008.

[17] L. Weed. Medical records that guide and teach. *The New England Journal of Medicine*, 278(12), March 1968.