

Design and Implementation of Anatomic Pathology Database System

Changwoo Yoon*, James K.Massey**, William H. Donnelly**, Douglas D. Dankel II***

*Computer and Information Science and Engineering, University of Florida
Shands Hospital at University of Florida, Box 100275 Dept of Pathology,
1600 SW Archer rd., Gainesville, FL32610
cwyoon@cise.ufl.edu

**Department of Pathology, University of Florida
Room MG-35, PO Box 100275, Gainesville, FL 32610-0275
massey@pathology.ufl.edu, donnelly@pathology.ufl.edu

***Computer and Information Science and Engineering, University of Florida
E301 CSE, C.I.S.E. PO Box 116120, Gainesville, FL 32611-6120
ddd@cise.ufl.edu

Abstract

This paper describes the design and implementation of the University of Florida's Anatomic Pathology Database System. The first phase of the system consists of the patient record parser and DB generator. The second phase includes application development to facilitate the clinical and research needs of pathologists. The parser separates the patient record into meaningful blocks of information. The final application will provide web based service to pathologist using the DB tagged with XML and the knowledge maintained by SNOMED Semantic Network Knowledge Base (SSN-KB). We describe the conceptual architecture of the SSN-KB which is structured for fast and efficient information searching and information extraction.

Introduction

The University of Florida's Anatomic Pathology Database (APDB) system was undertaken to facilitate the clinical and research needs of pathologists. Between 1980 and 2001, the Department of Pathology archived more than 470,000 surgical pathology patient records totaling more than 19.7 millions lines of text. All records are stored on magnetic tape written as text files. Each record has multiple fields defining the patient's characteristics, diagnosis, and SNOMED retrieval codes written by pathologist..

The first phase of APDB system development involves parsing the patient records and storing the obtained information to a database. Each patient record is in plain text format, which must be separated into meaningful units of information without examining the underlying semantics (Emile 1985). We describe the terms used for the separation of these records in Section 3.1. After the parsing is completed, the meaningful information units are stored in the database.

The second phase involves further parsing of the demographic, descriptive, and diagnostic elements of each

record to create XML tags. The final system includes encoding terms into records using a separate XML encoding program (Cyber 2001).

The objectives of the system are to:

- Provide the pathologist with a useable data base for routine patient care and quality assurance
- Develop an XML database for biological research for other basic and clinical investigators.
- Create an XML DB for informatics research.
- Maintain an expanding DB to protect against changes in computer reporting systems.
- Provide routine quality assurance and clinical trials reporting and notification.

This paper describes the metadata set definition used in the parser and the conceptual design of the knowledge base structure of SNOMED Semantic Network Knowledge base (SSN-KB) that will support the improvement in speed and accuracy of the information search and information extraction.

SNOMED

Surgical Pathology, cytology, and autopsy reports are highly structured documents describing specimens, their diagnoses, and retrieval and charge specification codes. The Systematized Nomenclature of Medicine (SNOMED) developed by the College of American Pathologists is used for a retrieval code. This was developed in collaboration with multiple professional societies, scientists, physicians, and computer consultants (Systematized 1979). SNOMED II is a hierarchically organized and systematized multiaxial nomenclature of medical and scientific terms. There are six main axes based on the nature of man, which begin with a hierarchical listing of anatomical systems (Topography). Any change in form of those structures throughout life is characterized in the (Morphology) axis. Causes or etiologies

Copyright © 2003, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

for those changes are listed in the (Etiology) axis. All human functions, normal and abnormal are listed in the (Function) axis. Combinations of Topography, Morphology, Etiology, and Function may constitute a disease entity or syndrome and are classified in the (Disease) axis. Using the T, M, E, F, and D axes it is possible to code nearly all-anatomic and physiologic features of a disease process.

$$\begin{array}{rcccccc} T & + & M & + & E & + & F & = & D \\ \text{Lung} & + & \text{Granuloma} & + & \text{M.tuberculosis} & + & \text{fever} & = & \text{Tuberculosis} \end{array}$$

Figure 1. "Equation" of SNOMED disease axes

A Procedure field allows identification of services or actions performed in behalf of the patient with the problem.

Pathology reports typically consist of useful, apt, and concrete terms in sentence or template format. The diagnostic terminology in reports and SNOMED involves standard terms and acceptable synonyms, both have the same SNOMED code number (e.g. Pneumonia and pneumonitis are coded T28000 M40000 (lung + inflammation). Pathology reports usually contain a specific field for SNOMED codes. Certain anatomic pathology computer systems include SNOMED files that allow code selection, but automated encoding programs are uncommon. Pre-coded synoptic templates of diagnostic terms allow a certain consistency for diagnostic encoding, but many diagnostic statements contain descriptive language, semantic forms, and linguistic nuances that make automated coding difficult. The need for error checking is constant.

Knowledge Base Structure

Current World Wide Web (WWW) development enables the building of a global information repository. This repository will provide meaning and value to the various data it contains. The operations of handling information range from information search to information extraction. Many investigators are trying to develop systems that extract exact information from dummy files (DAN 2000), (Steven 2000), (Rohini 2000), and (Dragomir 2000).

Two important characteristics of the APDB patient records are their fixed data and closed domain. The system's target data are patient records from 1980 to the present, which we consider as fixed or static, meaning that any dynamic features of the system can be minimized. The nomenclature used in a patient report is restricted to the domain of anatomic pathology and related areas of medicine, making it a relatively closed domain. These features provide a good situation and structure for constructing a knowledge base.

Among several forms of knowledge representation, the semantic network is widely used for representing simple hierarchical structures. Because the SNOMED has a hierarchical architecture, we adopted the semantic network for the knowledge representation method.

Metadata Set Definition

Appendix A. shows the metadata set definition used to parse the patient surgical pathology records. There are 25 terms that must be located and separated in the current patient record. These terms serve as attributes in the database table. Because some term names have changed through the years, several synonyms exist for same term. For example, "SURGICAL PATH NO", "ACC#", and "CYTOLOGY NO" have the same meaning: the sequential number of the patient record in the set.

Some terms can be used directly as XML tags such as NAME, ROOM, AGE, and SEX. Others, such as GROSS DESCRIPTION and DIAGNOSIS must be parsed further to create XML tags.

The parser, a batch program, processes the patient record and creates an output file containing separated patient record fields. The DB loader reads the output generated by the parser then stores the results to the DB. The parser also generates an index file that has proximity information among words inside the gross description and diagnosis. This can be used in multiple keyword information searches. The proximity information is needed to rank the relevant results.

Information Processing: Retrieval and Extraction

There are several distinct advantages in processing the pathology patient data. First, the patient record data from 1982 is unique to the University of Florida. This reflects the unique character, both regionally and periodically. Thus, when the parsing is finished, the analysis of frequency of words and multiple word terms has significant meaning. Second, because the patient reports are expressed in standard medical language (which will vary slightly from physician to physician), the terms used are sometimes not an exact match to the SNOMED terms. This makes it useful to analyze the patient reports based on the SNOMED terms. Patient reports also have a <Transcription Codes> field that shows matching SNOMED codes with the <Diagnosis>. The analysis of the SNOMED code frequency throughout the patient records can give a valuable research sense to the pathologist. These kinds of analysis can be done statically and can be reported all at once (Robert 2001) (Moore 2001).

While this static analysis is extremely useful, most information processing should be done dynamically. We cannot imagine or anticipate all requests that might be made of this knowledge base. Currently, we envision two types of requests.

One is information searching. A DB query can be made having several keys to the existing relational database. The system displays the searched results matching the input keys. Results are sorted by relevancy using pre-calculated proximity data. Figure 2 is an example of proximity data. If we want to search LOCAL HOSPITAL, the system lists the result patient record using the proximity value 6 of LOCAL and 8 of HOSPITAL.

```

"1984100001",3,1,"RECEIVED"
"1984100001",3,2,"SLIDES"
"1984100001",3,3,"LABELLED"
"1984100001",3,4,"83MCS-4769"
"1984100001",3,5,"83MCC-1300"
"1984100001",3,6,"LOCAL"
"1984100001",3,7,"COMMUNITY"
"1984100001",3,8,"HOSPITAL"
"1984100001",3,9,"ANYWHERE"
"1984100001".3.10."FLORIDA"

```

Figure 2. Example of proximity data

Information extraction is the other type of information processing. Here the pathologist gives a complete question to the system. The system searches for possible results then generates an exact answer for the question. For information extraction, a fast and efficient knowledge base structure is needed. To speed the search, there are several suggested data structures. One is a TRIE structure. The Trie is a tree of degree $m \geq 2$ in which the branching at any level is determined not by the entire key value, but by only a portion of it. For example, we can make a trie consisting of high frequency SNOMED codes from the patient reports. In this case we can search more efficiently only using frequently used term.

Alternatively, we could use a hash table as shown in Figure 3. In this structure the SNOMED terms are organized hierarchically in a relational tree. This hierarchy is then referenced through a hash table. Because the total number of SNOMED codes is 56,000 we can use hash functions to speedup searching.

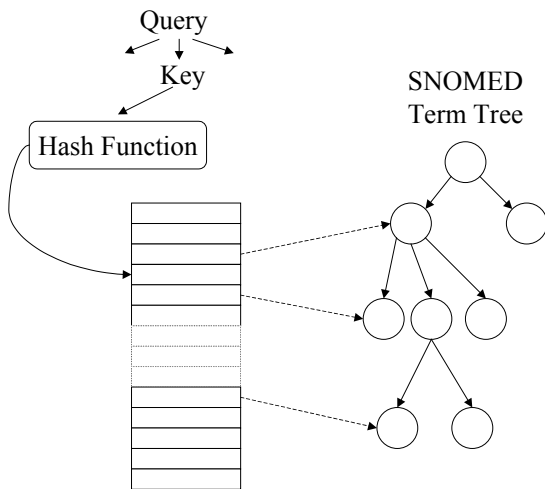


Figure 3. Hash structure of knowledge base

Accumulated Knowledge-base

SNOMED consists of six categories: Topography, Morphology, Etiology, Function, Disease, and Procedures. The patient report has <retrieval codes> terms showing matching SNOMED categories and numbers. With the first five axes it is possible to code most of the anatomic and physiologic elements of disease process, both normal and abnormal, and often sum up these elements as a codable class of disease or as a recognized syndrome by using the disease axes, basically what is called the SNOMED "equation" shown in Figure 1.

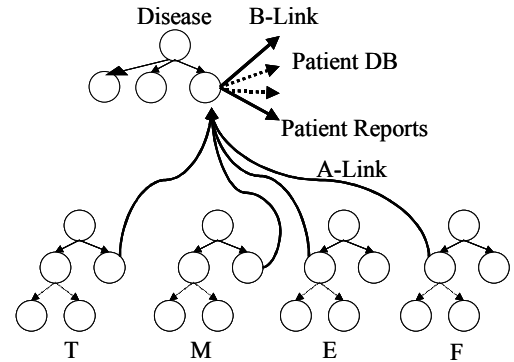


Figure 4. Dynamic knowledge base structure using semantic network

Some of the relations are straightforward but often cases have unique relationships based on the patient report. It is possible to develop a unique knowledge base using these relationships. Figure 4 shows a possible dynamic structure of the knowledge base using semantic network, which is named the SNOMED semantic network knowledge base (SSN-KB). The SSN-KB is made by traversing the retrieval codes of pre-loaded patient records. This kind of knowledge can be obtained from practical records only not by the original SNOMED nomenclature. Links represent the relationship between a disease and other terms that can be obtained from individual patient reports. Most common link type is IS-A and PART-OF. These links can be modified and grown dynamically. Link B represents links to a relevant patient report, knowledge obtained by the previous search or information extraction. It is possible to get additional information from the existing SSN-KB because the semantic network is a subset of first-order logic so we can infer further knowledge using inference rules of logic.

Sets created by A-link retrieval analysis enable investigators to apply new technologies as they develop. Data derived from coordinated use of multiple technologies open new areas for diagnosis, treatment, epidermology, and long term follow-up. They can be added to the knowledge base to enrich its value, and also serve to alert diagnostic pathologists to their use. The B-linkages allow exploration of data added from other data have, increasing the value of the A-linkage.

Future Plan

Future plans include phase II development of the system. This includes parsing of gross description and diagnosis using detailed XML tags and making a web based user interface enabling remote access to the system (Friedman 1999) (Ralf 2000). The support of a more efficient and intelligent knowledge base for information extraction is included. The research regarding collaborative system among different organizations is also of interest.

Conclusion

The anatomic pathology patient report has important meaning to pathologists both clinically and scientifically. We are building a knowledge base to aid pathologists in finding significant facts from previous cases, which has significant potential as an aid to patient care.

The patient record data is described as a closed domain with our knowledge fixed. We suggested a fast and self-growing knowledge structure for use in information extraction. This structure uses the hierarchical and inter-related structure of SNOMED terms and disease equation that can be found in unique patient reports.

References

CyberCoder for XML Phase I Synopsis, www.csihq.com/CSI/pdf/sbir_xml_rpti%20summary.pdf, 2001, CSI

Dan Moldovan; Roxana Girju; Vasile Rus 2000. Domain-Specific Knowledge Acquisition from Text. *Conference on Applied Natural Language Processing*, 268-275

Dragomir R. Radev; John Prager; Valerie Samn 2000. Ranking suspected answers to natural language questions using predictive annotation. *Conference on Applied Natural Language Processing*, 150-157

Emile C.Chi.; Carol F.; Naomi S.; Margaret S.L. 1985. Processing free-text input to obtain a database of medical information. *Annual ACM Conference on Research and Development in Information Retrieval*, 82-90

G. William Moore, and Jules J. Berman. 2000. Chap 4. ANATOMIC PATHOLOGY DATA MINING. *Medical Data Mining and Knowledge Discovery*.

Friedman C.; Hripcsak G.; Shagina L; Liu H. 1999. Representing Information in Patient Reports Using Natural Language Processing and the Extensible Markup Language. *Journal of the American Medical Informatics Association*, 76-87

Ralf Schweiger; Ali T.; Dudeck J. 2000. Using XML for flexible data entry in healthcare. *XML Europe 2000 Conference*

Robert E. Miller, MD; John K. Boitnott, MD; G. William Moore, MD, PhD. 2001. Web-based Free-Text Query System for Surgical Pathology Reports with Automatic Case De-Identification. *Arch Pathol Lab Med*. Forthcoming.

Rohini Srihari, and Wei Li. 2000. A Question Answering System Supported by Information Extraction. *Conference on Applied Natural Language Processing*, 166-172

Steven Abney; Michael Collins; Amit Singhal. 2000 Answer Extraction. *Conference on Applied Natural Language Processing*, 296-301

Systematized Nomenclature of Medicine, (Rodger A. Cote, Editor) College of American Pathologists, 1979, Skokie, IL.

Appendix A. Primary terms which is basis for the DB attribute

TERMS	ROLES	ETC
SURG PATH NO SURGICAL PATHOLOGY NO# ACC.# ACC# CYTOLOGY NO	<ul style="list-style-type: none"> Format:NNNN-YY-T NNNN: Serial number distinct in one year, digit width may vary YY: year expressed in two digit T: Type = { C, S, O, G, M } Type “C” Consultation Rpt “S” in-house surgical Rpt 	This number also shown at the end of the line having format: YYTNNNN###Y YMMDD
NAME	<ul style="list-style-type: none"> Patient name Format: Last, First, Middle, Suffix 	
TEST NO	<ul style="list-style-type: none"> Test number 	
SPECIMEN NO	<ul style="list-style-type: none"> Specimen number 	

SPECIMEN		
MED REC NO Medical Record #	<ul style="list-style-type: none"> 6 digit unique number of each hospital format: NN-NN-NN may vary 	
ROOM WARD	<ul style="list-style-type: none"> Room number 	WARD Patient location
AGE	<ul style="list-style-type: none"> Age of patient Format: NN [Y M D] NN – number Y – represent year M – month D - day 	
SEX	<ul style="list-style-type: none"> Sex of patient Format: {M F} 	
DATE Service Date	<ul style="list-style-type: none"> Service date Format: Month Day, Year Example: JANUARY 07, 1981 	
PHYS PHYSICIAN Referring Physician Surgeon	<ul style="list-style-type: none"> Referring Physician or Surgeon 	
REPORT TYPE	<ul style="list-style-type: none"> Example: S1 Surgical 	
SERVICE	<ul style="list-style-type: none"> Date obtained Date received 	
Date Obtained	<ul style="list-style-type: none"> Date obtained 	
Date Received	<ul style="list-style-type: none"> Date received 	
HISTORY CLINICAL HISTORY	<ul style="list-style-type: none"> Clinical history Specimen(s) submitted/ Procedures ordered 	Long text
Specimen submitted	<ul style="list-style-type: none"> 	
GROSS DISCRIPTION	<ul style="list-style-type: none"> 	
MICROSCOPIC DESCRIPTION MICROSCOPTIC DESCRIPTION	<ul style="list-style-type: none"> Light Microscopy Immunofluorescence microscopy Electron microscopy Other tests: e.g. included cytogenetics, molecular biology, or flow cytometry data 	
DIAGNOSIS	<ul style="list-style-type: none"> Bone marrow, aspiration: No lymphoma detected 	
COMMENT	<ul style="list-style-type: none"> 	
PATHOLOGIST	<ul style="list-style-type: none"> 	
RETRIEVAL CODES	<ul style="list-style-type: none"> Diagnostic/Retrieval codes Modifier codes Transaction codes: JP/whd Date of transcription: 03/23/99 Electronic signatures Date Electronically signed out 	