Employing complex polyhierarchical ontologies and promoting interoperability of i2b2 data systems

James Richard Campbell MD1, Walter Scott Campbell PhD1, Hubert Hickman MS2, Jay Pedersen MS3, James McClay MD1

1University of Nebraska Medical Center, Omaha, NE; 2Nebraska Medicine, Omaha, NE, 3University of Nebraska at Omaha, Omaha, NE

Abstract

I2b2 is in widespread use for managing research data warehouses. It employs reference ontologies as a record index and supports searching for aggregate cases using a pattern match operator on ASCII strings representing the node traversal from root to concept (PATHs). This creates complexities in dissemination and deployment for large polyhierarchical ontologies such as SNOMED CT. We hypothesized that an alternative approach employing transitive closure tables (TC) could lead to more accurate, efficient and interoperable search tools for i2b2. We evaluated search speed, accuracy and interoperability of queries employing each approach. We found both TC-based and PATH-based queries to produce accurate results. However, we observed that TC-based queries involving concepts included in large numbers of paths ran substantially faster than PATH-based queries for the same concept. Oracle query plan resource estimates differed by one to three orders of magnitude for these queries. We conclude that a simplification of dissemination tools for SNOMED CT and revision in the metadata build for i2b2 can effectively employ SNOMED CT with increased efficiency and comparable accuracy. Use of transitive closure tables in metadata can promote network query interoperability.

Introduction

Slowly and with little fanfare, reference terminologies1 and advanced ontologies supporting structured clinical data are reorganizing and enhancing electronic health records (EHR) and providing advanced data structures in support of clinical decision making. This has been accelerated in the US by coding requirements built into the Meaningful Use program by the Office of the National Coordinator for Health Information Technology2. EHR vendors are struggling to understand and appropriately deploy these technologies and the research and public health communities have scant understanding of their importance or technical underpinnings.

With the advent of ambitious programs to develop and support large clinical research networks3 and the appearance of tools for big data bases4,5, program administrators have little to guide them in the effective deployment of ontologies within data management systems. Arguably one of the greater challenges is SNOMED CT (SNCT) which is the most comprehensive healthcare ontology with a concept space of more than 302,000 elements. This large ontology can be difficult to effectively integrate into research data management tools. The September 2014 release of US edition of SNOMED CT is published as UMLS data sets of 3.05 Gbytes. The release format is an RF2 dataset8 specified by the IHTSDO but does not include any software or detailed guidance for best practice deployment into a system such as i2b2.

SNCT employs a concept model enforced by a description logic classifier which specifies the relationships and attributes to be used for concept definition in each SNCT domain. SNCT employs polyhierarchy (Definition: a hierarchically organized set in which a node may have more than one parent) in the ontology which means that the larger domains such as 404684003|Clinical finding (finding)| and 71388002|Procedure (procedure)| may have hierarchies of 30 generations in depth and many supertype concepts (Definition: X is the supertype of Y if a directed path of IS_A relationships can be traced from X to Y). Counting the possible paths from the root to a leaf concept, required for i2b2 metadata build, can be daunting since polyhierarchy expands the combinatorial possibilities. In the 2014 release of the US edition of SNCT there are up to 7802 paths (312411001|Carcinoma of the hepatic flexure (disorder)|) and an average of 11.1 paths per concept.

A research data management toolset in common use is i2b22. i2b2 software was first developed in 2004 when state of the art for terminology standards was ICD-9-CM4 and CPT-415. In contrast to SNCT, both are monohierarchies of modest size. (Definition: A hierarchically organized set in which any node may have at most one parent) i2b2 employs a star schema information model with four reference dimensions and a central fact table. i2b2. supports aggregation or subsumption queries (Definition: a subsumption query for concept X seeks to enumerate all instances from the fact table using the SQL pattern
match operator (LIKE) and an exhaustive set of PATHs from root concept to the node. For example, one possible path for the SNCT concept 128192002|Chronic disease of cardiovascular system (disorder)| is \(404684003\text{48572001}\text{276243003}\text{128292002}\). Figure 1 shows the SNCT hierarchy for this concept and the sequence of nodes in path number one which illustrate the path computation. In total for this concept there are 4 possible paths.

Transforming the NLM release of the US edition of SNOMED CT into i2b2 metadata is difficult and time consuming, resulting in files of 4.984 Gbytes; these are troublesome to manage and share between collaborators seeking to standardize their research data. Distribution of i2b2 metadata files from the National Center for BioOntology using web services is available for SNCT but reported to be seldom employed.

Importantly in this era of evolving national research networks, the i2b2 SQL query generated by the software is unique to the path selected by the user. It is only interoperable with other sites if their metadata deployment is exactly identical. To its credit, i2b2 manages such queries as efficiently as was considered possible in the age of ICD-9-CM.

Figure 1. SNOMED CT path computation

Extending previous work of Harris\(^1\), we hypothesized that deployment of a revised i2b2 search algorithm for SNOMED CT employing simple transitive closure tables for subsumption queries would be faster and as accurate as conventional i2b2 methods. (Definition: A transitive closure set for any concept X in a IS_A hierarchy is the set of nodes which includes X and any node Y which has a path to X by following one or more directed ISA relationships\(^2\).) We further expected that a compact piece of software could install the SNOMED CT metadata using only NLM data sets. We believed that i2b2 queries employing transitive closure would be interoperable across i2b2 installations with metadata build which could be tailored for local needs.

Methods

We installed i2b2 version 1.7 in our research data warehouse on a Dell Poweredge R820 server with 16 cores and 128GB RAM; 2X744GB SAS solid state drives; 14 TB SAS 10K drives; supporting Oracle 11g.

We obtained the September 2014 RF2 release of SNOMED CT, US edition, from the UMLS Terminology Services site\(^3\) of the National Library of Medicine. We extracted the set of all active SNOMED CT concepts, relationships and descriptions from the snapshot version following the file documentation provided in the Technical Implementation Guide\(^4\). All concepts pertinent to this release were identified with moduleID values of 900000000000207008 and 900000000000012004 (International release) and 731000124108 (US-edition concept)\(^14\).

We extracted the hierarchical (IS_A) relationships from the release files and developed a Python program that used those to prepare i2b2 metadata. First we identified the top concept for all SNCT hierarchies as first order subtypes of 38875005|SNOMED CT Concept|. We then calculated i2b2 metadata paths for SNCT employing an algorithm which is a modification of the standard depth-first search algorithm for graphs\(^15\). From the path search the program further prepared a transitive closure table for all of SNCT. A portion of the table is displayed in figure 2. The i2b2 metadata we prepared was employed in our i2b2 client as two versions: one using a conventional i2b2 PATH search and the Oracle ‘LIKE’ operator for computation of subsumption, and a second using a Transitive closure (TC) table and the ‘~’ operator. Figure 3 compares the metadata tables from i2b2 for the two search protocols. We created the
two versions of metadata to prepare the test environment for our research. i2b2 allows development for integration of ontologies. Historically, while most i2b2 metadata sets have been based upon concept paths, the flexibility of

![Image](image1.png)

Figure 2. Fragment of transitive closure table SNOMEDCT_I2B2_TC

i2b2 allows the development of other types of metadata such as dimension based metadata. Integration of the transitive-closure metadata table allows the i2b2 search engine to replace path-based calculation of subsumption with simple parent-child based subsumption using the TC table. Figure 3 shows the important columns from the two examples of i2b2 metadata. The first examples is transitive closure metadata, and the final column shows the traditional path based meta data. Note that we have created a concept_dimension_2 table analogous to concept_dimension that we use to isolate our testing with the traditional i2b2 style of metadata.

![Image](image2.png)

To compare accuracy and efficiency of the transitive closure methods, we developed a set of queries to be run using PATH-based logic and again using TC-based logic. The order in which the query logic was employed for each testing pair was randomized using a flip of the coin for each concept queried. The scope and complexity of each concept query varied from high-level concept such as 118234003|Findings by site(finding)| to mid-level concepts (128238001|Chronic heart disease (disorder)|) and low-level concepts (93030006|Congenital absence of spleen (disorder)|). We tested complex query performance by running queries that involved set calculations and disjoinedness, such as “count all patients with a chronic disease but not chronic cardiovascular disease”. We obtained database query resource requirements assessing query economy for each query using the “explain plan” SQL tool which provides an Oracle predicted relative resource value.

We further constructed an independent graph database and instantiated it with the fully classified SNCT model prepared from the same release files. We developed software to create the transitive closure table from the graph
database for validation of the i2b2 build. We compared the graph database transitive closure table to an instance developed with IHTSDO documented methods and validated the content of SNOMED CT. We copied all observation fact data from i2b2 to the graph database and reproduced the anonymized patient and SNOMED CT problem coding as an independent validation set. We then used the graph database to reproduce the concept queries that we employed for testing of TC and PATH query function and compared patient query sets for accuracy and reproducibility.

**Results**

The metadata implementation we developed employs the 20140901 RF2 snapshot release of SNOMED CT which is 503 MB of uncompressed terminology. The Python code produced a transitive closure table of 181 MB and a metadata set of 7.45 million rows at 4.984 GB. The metadata is complex due to SNOMED CT polyhierarchy and has from 1 to 7802 paths per concept. 16865 concepts of the 302,000 in the release have >=100 paths, most from 404684003|Clinical finding (finding)| hierarchy.

Table 1 summarizes the results of our controlled experiments with query execution. All queries retrieved identical patient sets regardless of query method, PATH-based or TC-based. We validated patient counts independently with queries of the graph database. We observed that subsumption query times for high-level concepts frequently ranged from 50% to 200% longer using the PATH-based approach. Query A in table 1 summarizes results for 404684003|Clinical finding (finding)| and query B 118234003|Finding by site (finding)| which are representative of these differences. Oracle relative resource estimates differed by up to three orders of magnitude in favor of the TC-based approach for these queries. We observed that query times involving concepts deeper in the SNOMED CT hierarchy were comparable between methods and differences between Oracle relative resource differences were negligible. Examples of these observations are reported in cases C, D and F. Queries incorporating disjunction for mid-level concepts also ran equally efficiently between methods (Table 1, query E). Surprisingly, we noted run time variability in PATH-based queries depending on browsing path chosen by the user. For example, see queries G and H in table 1. In no case did PATH-based queries run faster that TC-based queries and summary statistics favored transitive closure procedures in an unpredictable manner that made non-biased quantitative assessment impractical.

Figure 4 recounts SQL generated by the i2b2 query engine to retrieve observation facts for query instances G and H from table 1. All three examples are searching for subsumed concepts of 13645005|Chronic obstructive lung disease (disorder)|. Queries 4a and 4b are both from path-based queries for this concept set but were generated when the user chose different navigation paths from the root to the concept during browsing for the concept in the SNOMED CT hierarchy. Query 4c is the SQL generated by the i2b2 query engine when employing metadata prepared for transitive closure searches as in figure 3. Note that SQL generated by the two PATH-based queries differs in the PATH segments chosen for searching.

![Figure 4. i2b2 SQL code fragments querying for 13645005|Chronic obstructive lung disease (disorder)| and 3b are path-based subsumption queries; 3c transitive closure-based subsumption query](image)
**Discussion**

We have documented our SNCT metadata build program for i2b2 and have published the code, transitive closure table and SNCT snapshot for PCORI sites which are using i2b2. Due to the relative simplicity of metadata creation from a transitive closure table for the use case of subsumption queries, we have discussed the publication of a transitive closure table for US Edition of SNOMED CT with NLM staff.

Deployment of the SNOMEDCT transitive closure tables allows construction of subsumption-complete subsets of SNOMEDCT. That is, with the subsumption logic embodied in the transitive closure table instead of explicit in enumeration of concept paths, subsets both in breadth and depth of the SNOMED CT ontology can be built. The build will find all subtype concept instances in the database, regardless of path build in the browser metadata set.

Requirements of the i2b2 metadata build are a real impediment to deployment of advanced ontologies such as SNOMED CT due to complexity and file size. Searches employing PATH-based metadata do not generate unique SQL statements in the setting of polyhierarchies and are a risk to the interoperability for queries to be shared within
a research network. In order to limit the size of metadata files, we had performed earlier experiments building only a limited number of paths for each SNCT concept. We verified that the strategy was untenable and produced erroneous responses to queries as well as dysfunction of the browsing environment. Therefore management of interoperability risk constrains i2b2 metadata since the metadata also supports the data browsing environment. By eliminating the need for a Concept_dimension table in the CRC cell and removing PATH data from the SQL query, a transitive closure metadata build is a step towards a flexible browsing interface, reduced i2b2 maintenance for terminology updates and improved interoperability of network query management.

We expect, but have not yet tested our results for consistency of i2b2 function on other SQL supported databases; namely SQL Server and PostgreSQL. We will be deploying our SNOMEDCT metadata and transitive closure table on each platform.

We conclude that the set of cases returned by the PATH and TC-based approaches were identical for each query. We confirmed the accuracy of these query results using an independent graph database representation of the data and different programmatic search engine. We included a complete and comprehensive enumeration of IS_A pathways in our build of i2b2 metadata because of the requirements of the i2b2 browsing software. Regardless of the specific path chosen by the user from the top-level concept, the query responses were identical for PATH and TC queries.

We consistently observed that speed and programmatic efficiency was as good or better with TC-based queries compared to conventional i2b2 PATH-based searches. As a puzzling observation, we found PATH-based queries of polyhierarchal concept data introduced unanticipated variation in query time response. While ASCII string length and the number of paths involving the target concept seemed to influence efficiency, we were unable to develop a quantitative model of query performance.

Conclusions

Deploying complex polyhierarchical ontologies in i2b2 can be a daunting task. We have documented and published one approach which has the virtues of employment of a US Meaningful Use² standard ontology, modest file size and freely available software. Metadata maintenance workload is diminished since metadata tables are more compact. Extending the work of Harris¹¹, we have confirmed that modifying i2b2 metadata build to employ transitive closure tables for support of subsumption queries leads to accurate results with retrieval efficiency as good or better than the current convention. SQL queries built from these modifications are devoid of browsing ontology references and are semantically parsimonious, promoting query interoperability.

References