Journal of Biomedical Informatics 57 (2015) 278–287

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Structural measures to track the evolution of SNOMED CT hierarchies

Duo Wei [a,*], Huanying (Helen) Gu [b], Yehoshua Perl [c], Michael Halper [d], Christopher Ochs [c], Gai Elhanan [c,e], Yan Chen [f]

[a] Computer Science and Information Systems-BUSN, Stockton University, Galloway, NJ 08205, United States
[b] Computer Science Dept., New York Institute of Technology, New York, NY 10023, United States
[c] Computer Science Dept., New Jersey Institute of Technology, Newark, NJ 07102, United States
[d] Information Technology Dept., New Jersey Institute of Technology, Newark, NJ 07102, United States
[e] Halfpenny Technologies Inc., Blue Bell, PA 19422, United States
[f] Computer Information Systems Dept., BMCC, CUNY, New York, NY 10007, United States

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 1 August 2015
Accepted 1 August 2015
Available online 7 August 2015

Keywords:
Terminology
SNOMED CT
Complexity measure
Abstraction network
Quality assurance

ABSTRACT

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is an extensive reference terminology with an attendant amount of complexity. It has been updated continuously and revisions have been released semi-annually to meet users' needs and to reflect the results of quality assurance (QA) activities. Two measures based on structural features are proposed to track the effects of both natural terminology growth and QA activities based on aspects of the complexity of SNOMED CT. These two measures, called the *structural density measure* and *accumulated structural measure*, are derived based on two abstraction networks, the *area taxonomy* and the *partial-area taxonomy*. The measures derive from attribute relationship distributions and various concept groupings that are associated with the abstraction networks. They are used to track the trends in the complexity of structures as SNOMED CT changes over time. The measures were calculated for consecutive releases of five SNOMED CT hierarchies, including the Specimen hierarchy. The structural density measure shows that natural growth tends to move a hierarchy's structure toward a more complex state, whereas the accumulated structural measure shows that QA processes tend to move a hierarchy's structure toward a less complex state. It is also observed that both the structural density and accumulated structural measures are useful tools to track the evolution of an entire SNOMED CT hierarchy and reveal internal concept migration within it.

1. Introduction

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [1] is a large and complex structure, with its January 2015 release containing about 315,904 concepts organized into 19 hierarchies. Introduced in its original form by the College of American Pathologists (CAP) in 1977, SNOMED CT has been proposed for use as a standard in general encoding in Electronic Health Record (EHR) systems. In 2007, SNOMED CT's ownership was transferred from CAP to the International Health Terminology Standards Development Organization (IHTSDO).

To meet the needs of users around the world, SNOMED CT has been continuously evolving since its creation via the merger of SNOMED RT and CTV3 [2]. New SNOMED CT releases are published twice a year, in January and July, with each release including refinements to descriptions, enhancements of concept definitions, and additions of new concepts. At the same time, SNOMED CT undergoes a clinical and technical quality assurance (QA) process conducted by IHTSDO's Quality Assurance Committee [3]. For a review of SNOMED CT users' views regarding evolution and QA, see [4].

In this paper, we examine the effects of these two kinds of modifications, namely, natural growth and QA, on the complexity of a SNOMED CT hierarchy. Our hypothesis is that, in general, modeling errors (e.g., missing relationships, incorrect parents) contribute to structural disorderliness. The question is: can one expect to see a simplification of the hierarchy structure due to the reduction of such disorderliness after a QA regimen has been carried out? And we would like to ask the same question concerning a natural growth period. Toward this end, we posit a way to assess the complexity of a hierarchy based on previously defined abstraction networks for SNOMED CT. An abstraction network is a framework that, among other things, forms the basis for systematic QA. Specif-

* Corresponding author at: Computer Science and Information Systems, School of Business, Stockton University, Galloway, NJ 08205, United States. Tel.: +1 609 626 3813; fax: +1 609 626 5539.
E-mail addresses: duo.wei@stockton.edu, duobwin@gmail.com (D. Wei).

http://dx.doi.org/10.1016/j.jbi.2015.08.001
1532-0464/

ically, we use the *area taxonomy* and *partial-area taxonomy* that are derived via structural analyses of the underlying SNOMED CT hierarchy. In this context, two derived complexity measures are proposed for quantifying the complexity of a hierarchy. One is called the *structural density measure*; the other is called the *accumulated structural measure*.

As a test-bed, the measures are applied to the Specimen hierarchy in order to track its changing complexity during the years 2004–2013. During that time, we personally carried out two QA processes on the 2004 and 2007 releases. Also, new concepts had been added to the hierarchy due to natural growth in the interim, and their introduction may have indeed led to new errors. Furthermore, both editing and the QA of a hierarchy are difficult tasks, which by themselves are never foolproof. A domain-expert auditor may very well overlook some errors, and the editorial policies may be incomplete or inconsistent. We look for any further impact of this subsequent QA effort on the complexity measures in comparison to the impact of the initial QA audit for the same hierarchy. An initial report of this study appeared in [5]; however, the research further evolved with changes in the definitions of the complexity measures. We also look for the trend of a hierarchy's complexity due to the natural development of SNOMED CT and the trend due to the mixed impact of both kinds of activities. By tracking the structural density measure over multiple years, we are able to identify when intensive QA activities have taken place. While our focus is on the Specimen hierarchy, we also analyze changes in complexities involving four other hierarchies with the use of the structural density measure.

2. Background

2.1. Area taxonomy and partial-area taxonomy

The *area taxonomy* and the *partial-area taxonomy* [6,7] of a SNOMED CT hierarchy are derived automatically from the respective lateral (i.e., non-IS-A) relationships exhibited by the concepts. The partial-area taxonomy also relies on local configurations of the IS-A hierarchy itself. Both taxonomies are based on the notion of *area*, a collection of all concepts with the exact same set of relationships. Such a collection is denoted by its respective list of relationships (inside braces). For example, in Fig. 1(a), showing concepts from Specimen, *Lesion sample* and its child *Specimen from ulcer* have only one relationship *morphology* (not displayed). Thus, they are grouped into the area {*morphology*}. *Swab* has only one relationship *procedure* and thus is in the area {*procedure*}. *Skin swab* belongs to the area {*topography*, *procedure*} due to it exhibiting those two relationships.

An *area taxonomy* is a graph structure that consists of only the areas represented (as nodes) and hierarchical *child-of* relationships connecting them. A portion of the area taxonomy for SNOMED CT's Specimen hierarchy corresponding to Fig. 1(a) is shown in Fig. 1(b). The area at the top is on Level 0 (equal to its number of relationships) and is named $\varnothing$ for the empty set of its relationships. It contains all concepts with no relationships. The dashed bubbles in Fig. 1(a), below the level of *Specimen*, denote area membership in Fig. 1(b). The number of concepts in each area appears in parentheses under the name.

A *root* of an area is a concept, of that area, whose parents all reside in other areas. An area may have more than one root. The *child-of* relationships—the arrows in the figure—are derived from the IS-As of the roots as described in [6].

The *partial-area taxonomy* extends the area taxonomy by further refining areas with multiple roots. In addition to areas, the partial-area taxonomy includes *partial-areas*, each being a set of concepts comprising a single root and all its descendants within one area. Fig. 1(c) is the portion of the Specimen hierarchy's partial-area taxonomy refining Fig. 1(b). The nodes representing the partial-areas are embedded in the respective area nodes. A partial-area's label is its constituent root, which hierarchically sits atop (and thus subsumes) all other concepts in the partial-area. For example, partial-area *Swab* has that concept plus its six descendants in the area {*procedure*}. Note that while the root concepts name the partial-areas, the names of the non-root concepts are hidden. We observe that in the area {*procedure*}, eight partial-areas are shown in Fig. 1(c), e.g., *Biopsy sample*, *Smear sample*, and *Swab*. The number in parentheses alongside a partial-area name indicates its number of concepts. For example, in the area {*procedure*}, we see partial-area *Biopsy sample* (4) whose other three non-root (hidden) concepts are *Specimen from unspecified body site obtained by biopsy, Specimen obtained by fine needle aspiration procedure,* and *Specimen from unspecified body site obtained by fine needle aspiration*, which is a child of the previous two children of the root.

The *child-of* relationships in the partial-area taxonomy are defined between partial-areas and are derived from the IS-As directed from the roots, similarly to those in the area taxonomy. To minimize the number of arrows, we use graphical abbreviations described in [6].

In [8], it was shown that concepts residing in more than one partial-area ("overlapping" concepts) have a higher likelihood of being in error than other concepts. Thus, they were chosen as a basis for a QA regimen. Furthermore, in [8], we introduced the *disjoint partial-area taxonomy* in which such overlapping concepts are extracted to form special partial-areas of their own.

2.2. Previous attempts on SNOMED CT complexity measures

The issue we are investigating is how to assess the complexity of a SNOMED CT hierarchy. In particular, we are interested in studying how complexity measures reflect on the evolution of a given hierarchy over multiple releases as a result of QA regimens and natural development of that hierarchy. One natural criterion is a global weighting function for a hierarchy such as size (the number of concepts) or height (number of levels in the longest hierarchical path). Indeed, in a comparison of such measures following our first audit of the Specimen hierarchy in the 2004 SNOMED CT release, the number of concepts was reduced from 1056 to 1044 (July 2005 release), and the height was reduced from 12 to ten. At the same time, SNOMED CT's total concepts went up from 357,134 to 364,461. Furthermore, only two hierarchies of SNOMED CT decreased in size during this period, the second of which was the huge Clinical Finding hierarchy obtained by integrating the two hierarchies Finding and Disorder. We attribute the decrease in the size of the Specimen hierarchy, which went against the general trend of growth in SNOMED CT during the same period, to the correction of duplicate concept errors (such as *Ear sample* and *Specimen from ear*) and the removal of improper concepts due to our QA efforts [6,7]. The former were caused by the failure to identify the synonymy of "sample" and "specimen" when integrating SNOMED RT and CTV3 into SNOMED CT [9]. The errors we found were reported to CAP and were corrected in future releases. The reduction in height can be attributed to finding errors in some of the most complex concepts in the hierarchy, which participated in the longest hierarchical paths.

However, these measures are more magnitude measures than complexity measures. The size measure accounts only for limited QA impacts such as erroneous concepts eliminated from the hierarchy, but not for other errors that were corrected. The size is also influenced by concepts added to the hierarchy as part of normal expansion. The height measure reflects only QA on a few concepts in the longest hierarchical path. Furthermore, such global measures fail to take into account the role of lateral relationships in the complexity of the concepts. For example, a hierarchy may keep