MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1075

Information-Content-based Measures for the Structure of Terminological Systems and for Data recorded using these Systems

Ronald Cornet

Department of Medical Informatics, Academic Medical Center – University of Amsterdam, Amsterdam, The Netherlands

Abstract

Terminological systems such as SNOMED CT play an increasingly important role in contemporary record keeping. This drives the need of assessing the content of these systems, as well as the content of medical records captured using these systems. In this paper, the use of information content as a measure for the structure of terminological systems and the content in medical records is explored. Two complementary information-content-based measures for terminological systems are proposed: the proportion of concepts with zero information content, and the average information content. The measures are applied to the latest releases of SNOMED CT. The measures are useful as an indicator of the overall structure of terminological systems or parts thereof. Furthermore, two measures are described which can provide an estimate for the content of medical records that is captured using a terminological system. Information content is shown to provide a useful basis for assessing the structure of terminological systems and the content of medical records.

Keywords:

Terminology, SNOMED CT, Information theory.

Introduction

Medical terminological systems provide a systemized representation of medical knowledge. A large number of terminological systems have been developed over the last decades. Whereas these systems originally were small lists or hierarchical systems, contemporary systems are large and complex.

The increasing size and complexity of terminological systems raises a number of challenges. First, the need arises for automated ways to assess their quality, as the effort of doing this manually becomes too high, and because maintenance of terminological systems increasingly becomes a team effort, which further increases the need for structural and reproducible methods[1]. Recently, a special issue of the Journal of Biomedical Informatics was fully dedicated to the auditing of terminological systems in medicine [2].

Second, the adequate use of terminological systems becomes increasingly intricate. Traditionally, terminological systems focused on the task of classification, i.e., determining the most appropriate category or "label" for a patient. Classification brings the challenge of adequately applying the classification rules to determine which category is the most appropriate. In contemporary compositional terminological systems, where the emphasis shifts from mere classification to structured and detailed coding of information, the user needs guidance not only on determining the most appropriate concept, but also on providing necessary and relevant detail. For example, the 2007 release of the 10th edition of the International Classification of Diseases (ICD-10) distinguished 12 categories of viral meningitis, and 4 residual categories (i.e., "other" or "unspecified"). The July 2009 release of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [1] contains 29 types of viral meningitis and 4 residual categories (which are likely to be removed from SNOMED CT shortly). As another example, ICD-10 contains 4 categories of acute myocardial infarction, and 2 residual categories; SNOMED CT distinguishes 49 types and 3 residual categories.

Using SNOMED CT for capturing patient information enables multiple uses of the data. For this purpose it is required that data is captured with maximal detail. Consequently, rather than resorting to a generic concept such as acute myocardial infarction, the user must be supported to provide any available detail, while having the possibility of being less specific when information is not (yet) available. In order to assess the amount of detail provided, metrics are needed.

In this paper, the use of information content is explored to measure both the structure of terminological systems as well as the content of records captured using these systems. As an example, these measures are applied to SNOMED CT.

Background

Terminological systems have a variety of structural aspects that influence their quality. Generally, metrics such as number of concepts and number of relationships are presented, but these do not necessarily correlate with quality. Other metrics which related more closely to quality are for example: number of superordinate concepts, subordinate concept, and roots, and number and nature of differentiae [3].

[1] http://www.ihtsdo.org/

Information content is applied to biomedical terminological systems, for example to investigate semantic similarity of concepts in the Gene Ontology [4, 5].

A benefit of these measures is that they can be calculated automatically. However, in the case of number of super- or subordinate concepts, averages or frequencies need to be analyzed to summarize the quality of a terminological system. In the case of the semantic similarity measures, a GO-specific measure was developed, which provides concept-based measures and can not be applied to other terminological systems.

Materials and Methods

SNOMED CT

Today, SNOMED CT is among the largest clinical healthcare terminological systems. The most recent release, of July 2009, contains about 290,000 active concepts. SNOMED CT provides formal definitions for these concepts using about 430,000 IS_A relationships and some 700,000 attribute relationships such as finding site, method, and associated morphology. These relationships serve three purposes: making semantics explicit; automated classification; and allowing post-coordination. SNOMED CT content is organized in a number of categories, such as clinical finding, body structure, and procedure.

Information content

The information content of a concept is a numerical measure of the information that is represented by the concept [6]. The information content of a concept c can be quantified as negative the log likelihood, -log p(c). Generally any base of the logarithm can be used (e.g., 2, e, or 10). In this paper, the $\log_2$ will be used.

For example, when tossing a fair coin, the probability of coming up heads is 0.5, and the information content thereof is 1. Likewise, when throwing a fair die, the probability of throwing 2 is 1/6, and the information content thereof is 2.58. So, a lower probability corresponds to higher information content.

As the above examples show, information content is determined based on actual probabilities. Suppose one has a manipulated die that always results in throwing 6, the information content of a throw is 0, as the probability of throwing 6 is 1.

The fact that information content depends on actual probabilities is a drawback when attempting to determine the information value of concepts in a terminological system. As terminological systems can be used in a broad range of situations, the actual information content may differ. For example, the probability of a person to be of female gender may be about 0.5 in the general population, 0.1 in the army, and 1 in a gynecology department, in which cases the respective information content is 1, 3.3, and 0. To get round this, the probability of coordinate concepts will be regarded as equal, i.e., when a concept has 4 subordinate concepts, each is regarded as having a probability of 0.25, and an information content of 2.

The information content of concepts is used in various ways to provide measures for a terminological system. First, two de-

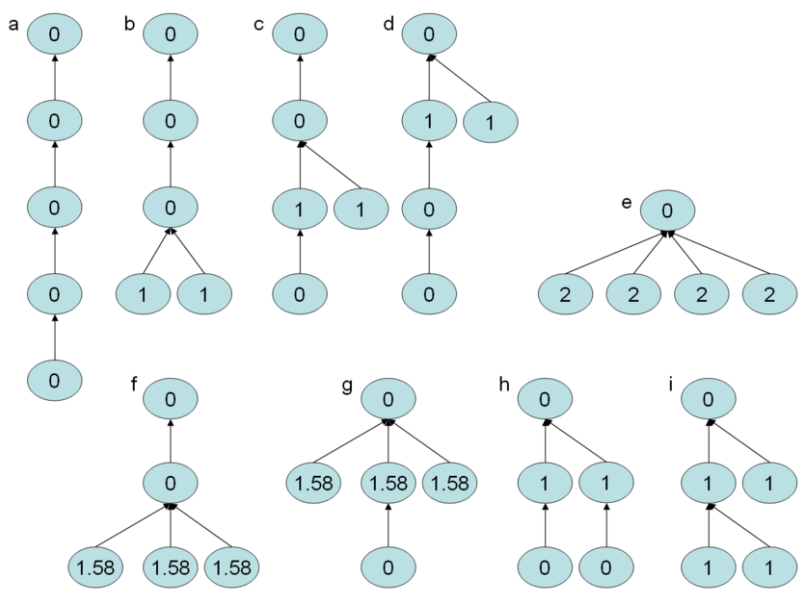

Figure 1- Possible configurations of a rooted mono-hierarchical terminological system with 5 concepts. Arrows denote Is_A relations, ovals denote concepts, and the numbers indicate the information content of the individual concepts.

rived measures for terminological systems are defined which are illustrated by a small example, consisting of 5 concepts, as shown in Figure 1. Then, the use of information content for data in a medical record will be addressed. Finally, the measures will be applied to SNOMED CT. Calculation of these measures was done by importing the text files of the latest releases of SNOMED CT into a database. The basic relational structure that is reflected in the SNOMED CT text files were extended, so that for every content its number of superordinate, subordinate and coordinate concepts could be recorded, as well as the resulting information content.

Measures

Information content as a measure of the structure of terminological systems

Figure 1 shows the possible configurations of a terminological system consisting of 5 concepts. In this figure, all concepts are subordinate to exactly one superordinate concept, apart from the root concept, which has no superordinate concept.

From these straightforward examples, a number of measures can be determined. First, the total and average information content of the terminological system can be calculated. In this example, the total differs between 0 for configuration (a) and 8 for configuration (e). The average (excluding the root concept, which is 0 by default) differs likewise between 0 and 2. As shown in Figure 1, both of the configurations (a) and (e) have hardly any structure, from which it can be concluded that neither a large nor a small value for the total or average information content is preferred. In the examples, concepts with zero information content contribute significantly to the small total values. These concepts are subordinate concepts without co-ordinate concepts (i.e., they are the single child of their parent concept). As stated in [3] "the presence of such cases is reason to suspect the presence of error." In configuration (e), all concepts are co-ordinate. According to [3], when there are a large number of co-ordinate concepts, these concepts "may point to issues such as a lack of organization or incomplete descriptions."

These examples show that information content can provide a measure for the organization of the terminological system. A terminological system which is organized as a full binary tree (in which every node other than the leaves has two children) has an average information content of exactly 1, and can be regarded as a maximally organized system.

The configurations from Figure 1 are all mono-hierarchies, whereas terminological systems such as SNOMED CT are poly-hierarchies, in which concepts can be subordinate to more than one superordinate concept.

Examples thereof are shown in Figure 2. In these cases the information content is calculated from a likelihood which is the division of the total number of superordinate concepts by the total count of co-ordinate concepts, in which any concept that is a co-ordinate concept for multiple superordinate concept count multiple times.

In configuration (a) of Figure 2, one concept has 2 superordinate concepts, which have 1 and 2 subordinates respectively. So the information content is $-\log(2/3) = 0.58$. In configuration (b), the two "leaf" concepts both have two superordinate concepts, each of which has 2 subordinates. So for the leaf concepts the information content is $-\log(2/4) = 1$.

Using these measures, a terminological system can be characterized by the proportion of concepts with zero information content, and the average information content of the other concepts. These measures can also be calculated for parts of a terminological system, thereby providing insight in the structure of sub-hierarchies of the system.

Information content as a measure of the content of medical records

Information content can not only play a role in evaluating the structure of a terminological system, but also in estimating the information content of terminological-system-based data entered into a medical record. As terminological systems such as SNOMED CT should enable multiple use of patient information, data is preferably captured with maximal detail.

To determine the amount of detail a concept provides, rather than the individual information content of a concept, cumulative information content is used. A distinction is made between total and relative information content.

Total information content is the sum of the information content of superordinate concepts up to the root. In configuration (a) from Figure 2, the leaf nodes will have a total information content of 2 and 1.58 respectively.

Relative information content is the sum of information content of superordinate concepts up to a non-root superordinate concept. For example, if one records gender, the information content relative to gender (e.g., "Finding related to biological sex" in SNOMED CT) can be calculated.

Application to SNOMED CT

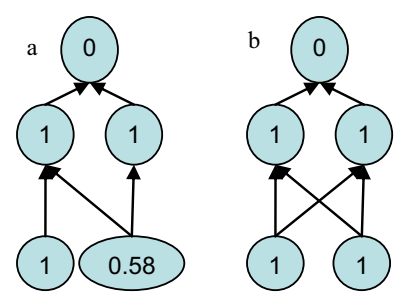

Figure 2- Example configuration of two rooted poly-hierarchical terminological systems with 5 concepts.

The above measures were calculated for the last three releases of SNOMED CT (July 2008, January 2009, and July 2009). To this end, all active non-limited concepts (i.e., those where concept status is 0) were taken into account, and all Is_A relationships between them. Table 1 shows the total number of concepts in each release, the proportion of concepts with zero information content, and the average information content of the concepts that have non-zero information content. Table 2 shows the measures for the 15 largest categories in the July 2009 release of SNOMED CT (based on the category mentioned in the fully specified name). The highest and lowest values for the proportion of concepts without information content and the average information content of the remaining concepts are shown in bold. These figures indicate that the subhierarchy of body structures overall contains relatively small numbers of subordinates per concept. This is indicated by the high proportion of concepts without information content (i.e., which are to only subsumer of a concept) on the one hand, and by the low average information content on the other hand. Conversely, the category "pharmaceutical/biological product" shows have relatively little organization, having many subordinates per concept. It turns out that the 4500 concepts with the highest information content in SNOMED CT are all products, with "Saliva stimulating tablet" having the highest information content of 11.33 (as it has 2580 co-ordinate concepts and 1 superordinate concept).

The "event" hierarchy combines a low proportion of concepts without information content with an adequate structure.

Application to data collected using SNOMED CT

At the department of Intensive Care of the AMC, SNOMED CT is used in a pilot to record reasons for admission to intensive care. This pilot started mid-December 2008. Figure 3 shows for the reasons for admission that were recorded at the Intensive Care of the AMC in the period January-June 2009 their average total information content and their total number of recorded concepts. This figure does not take into account a small number of concepts that were post-coordinated. Figure 3 suggests that the level of detail in which users are recording was stable for the first four months, and dropped thereafter. More data will be needed to see if this reduction is permanent, and further analysis of the data is needed to explain any loss of information content.

Table 1 – Measures for the latest releases of SNOMED CT

Release	# concepts	no content	avg content
July 2008	289028	5.14%	3.79
January 2009	292104	5.05%	3.94
July 2009	289897	5.11%	3.89

Discussion

Information content as a measure of the structure of terminological systems

Two measures were introduced regarding the structure of terminological systems: the proportion of concepts with zero information content, and the average information content of concepts with non-zero information content. These measures provide insight into two complementing structural aspects of terminological systems: concepts without co-ordinate concepts and concepts with large numbers of co-ordinate concepts. These types of concepts are identified as possibly erroneous or ill-defined.

The benefit of these measures is that they can be applied to either a complete terminological system or any part thereof. A drawback of the measure of average information content may be that it is strongly influenced by concepts with large numbers of co-ordinate concepts. As all of the co-ordinate concepts have a large number of co-ordinate concepts, the higher information content is multiplied by the large number of concepts, and outweighs the concepts which have a small number of co-ordinate concepts and hence a lower information content. Further research is needed to determine whether this is actually a drawback, or whether this helps in pointing out areas that need review.

Table 2 – Measures for the 15 largest categories in SNOMED CT, July 2009 release

category	# concepts	no content	avg content
Disorder	63841	3.31%	3.79
Procedure	47880	3.86%	4.08
Clinical finding	32836	3.90%	3.43
Organism	31857	8.39%	4.18
Body Structure	26144	**10.27%**	**2.79**
Substance	23621	5.57%	4.61
Pharma/Biol. Product	16879	5.25%	**5.28**
Qualifier Value	8902	3.02%	3.96
Observable entity	7945	6.46%	2.85
Physical object	4420	5.45%	3.19
Morphologic abnormality	4307	3.34%	4.09
Occupation	3842	3.18%	2.88
Event	3579	**2.49%**	2.89
Situation	3090	6.83%	4.23
Regime/therapy	2875	5.36%	3.54

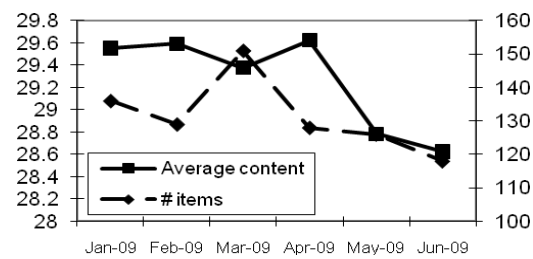

Figure 3 – Average of total information content (left axis) and number (right axis) of reasons for admission recorded at the intensive care unit of the AMC in the first six months of 2009.

Information content as a measure of the content of medical records

Measuring total or relative information content is especially useful for record items that can be captured with a varying level of detail. This is the case for example when recording findings or procedures, which can be recorded with more or less detail. For example, with increasing level of detail, one can use SNOMED CT to record infective meningitis, bacterial meningitis, Gram-negative bacterial meningitis, Haemophilus meningitis, and thromboembolic meningoencephalitis. As not all detail will be available in any situation, the possibility of recording information with less detail must exist. However, it is useful to analyze whether users generally resort to generic concepts, or try to provide maximal detail. The information content can provide insight in this recording behavior, for example over time, or depending on the way in which information is recorded by users. In a previous study, an analysis was performed on the level of detail in which information was recorded and a comparison was made between free-text recording and terminology-based recording [7]. In that research it turned out that it was generally hard to determine whether one or the other provided more detail due to lack of an adequate measure for that. Figure 3 provides an example of how information content can be used for such purposes.

Further work

This paper applies the measures only to SNOMED CT, which shows that the three most recent releases are relatively constant regarding these measures. Applying the measures to other terminological systems will provide more insight into the results, enabling comparison between terminological systems rather than between versions or specific parts thereof.

Ideally, measures like these are not only calculated for evaluation purposes, but also for providing guidance or prioritization for future maintenance and improvement of these systems. It needs to be determined whether this is practically feasible.

Contemporary terminological systems such as SNOMED CT include other than Is_A relationships. Currently these relationships are not explicitly addressed in the measures presented, but only implicitly, as they are essential for the way in which the hierarchical structure in SNOMED CT is realized, namely by automated classification. However, as these attribute relationships are also important for supporting post-coordination, it would be important to address them. As post-coordination can play an important role when capturing information in a medical record, further research on the measure of information content of recorded data is necessary.

Conclusion

In this paper 2 measures are presented that are based on information content: the proportion of concepts with zero information content, and the average information content of a terminological system. Two other measures are described for the content of medical records that is captured using a terminological system: total and relative information content. Benefits of these measures are that they can be relatively easily calculated, and are grounded in information theory.

The measures are useful as an indicator of the overall structure of terminological systems or parts thereof. Information content is shown to provide a useful basis for assessing the structure of terminological systems and the content of medical records.

Quantifying the level of detail in which information is captured in a medical record is a first step towards assessing the quality of recorded data.

Acknowledgements

The author likes to thank Olivier Dameron for valuable discussions and feedback on the manuscript.

References

[1] Bakhshi-Raiez, F., R. Cornet, and N.F. de Keizer, Development and application of a framework for maintenance of medical terminological systems. J Am Med Inform Assoc, 2008. 15(5): p. 687-700.

[2] Geller, J., et al., Special issue on auditing of terminologies. J Biomed Inform, 2009. 42(3): p. 407-11.

[3] Bodenreider, O., et al. Investigating Subsumption in DL-Based Terminologies: A Case Study in SNOMED CT. in KR 2004 Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004). 2004. Whistler, BC, Canada: AMIA.

[4] Pesquita, C., et al., Semantic similarity in biomedical ontologies. PLoS Comput Biol, 2009. 5(7): p. e1000443.

[5] Lord, P.W., et al., Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 2003. 19(10): p. 1275-83.

[6] Ross, S., A first course in probability. 1998, Upper Saddle River, NJ: Prentice-Hall. 514.

[7] de Keizer, N.F., et al., Post-coordination in practice: Evaluating compositional terminological system-based registration of ICU reasons for admission. Int J Med Inform, 2008. 77: p. 828-835.

Address for correspondence

Ronald Cornet, AMC, dept. of Medical Informatics, J1B-115
P.O. Box 22700 1100 DE Amsterdam, The Netherlands
r.cornet@amc.uva.nl