Accepted Manuscript

Evaluating the Granularity Balance of Hierarchical Relationships within Large Biomedical Terminologies towards Quality Improvement

Lingyun Luo, Ling Tong, Xiaoxi Zhou, Jose L.V. Mejino Jr, Chunping Ouyang, Yongbin Liu
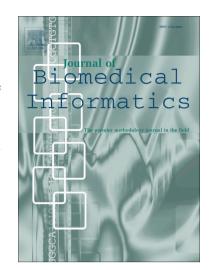
Please cite this article as: Luo, L., Tong, L., Zhou, X., Mejino Jr, J.L.V., Ouyang, C., Liu, Y., Evaluating the Granularity Balance of Hierarchical Relationships within Large Biomedical Terminologies towards Quality Improvement, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.10.001

Evaluating the Granularity Balance of Hierarchical Relationships within Large Biomedical Terminologies towards Quality Improvement

Lingyun Luo[a,*], Ling Tong[a], Xiaoxi Zhou[a], Jose LV Mejino Jr[b], Chunping Ouyang[a], Yongbin Liu[a]

[a]*School of Computer Science, University of South China, Hengyang, Hunan, 421001, China*
[b]*Structural Informatics Group, Department of Biological Structure, University of Washington School of Medicine, Seattle, USA. WA 98195*

Abstract

Organizing the descendants of a concept under a particular semantic relationship may be rather arbitrarily carried out during the manual creation processes of large biomedical terminologies, resulting in imbalances in relationship granularity. This work aims to propose scalable models towards systematically evaluating the granularity balance of semantic relationships. We first utilize "*parallel concepts set (PCS)*" and two features (the length and the strength) of the paths between PCSs to design the general evaluation models, based on which we propose eight concrete evaluation models generated by two specific types of PCSs: single concept set and symmetric concepts set. We then apply those concrete models to the IS-A relationship in FMA and SNOMED CT's Body Structure subset, as well as to the Part-Of relationship in FMA. Moreover, without loss of generality, we conduct two additional rounds of applications on the Part-Of relationship after removing length redundancies and strength redundancies sequentially. At last, we perform automatic evaluation on the imbalances detected after the final round for identifying missing concepts, misaligned relations and inconsistencies. For the IS-A relationship, 34 missing concepts, 80 misalignments and 18 redundancies in FMA as well as 28 missing concepts, 114 misalignments and 1 redundancy in SNOMED CT were uncovered. In addition, 6,801 instances of imbalances for the Part-Of relationship in FMA were also identified, including 3,246 redundancies. After removing those redundancies from FMA, the total number of Part-Of imbalances was dramatically reduced to 327, including 51 missing concepts, 294 misaligned relations, and 36 inconsistencies. Manual curation performed by the FMA project leader confirmed the effectiveness of our method in identifying curation errors. In conclusion, the granularity balance of hierarchical semantic relationship is a valuable property to check for ontology quality assurance, and the scalable evaluation models proposed in this study are effective in fulfilling this task, especially in auditing relationships with sub-hierarchies, such as the seldom evaluated Part-Of relationship.

Keywords: Biomedical Terminology, Quality Assurance, FMA, SNOMED CT, Granularity Balance

*Corresponding author
Email address:* luoly@usc.edu.cn (Lingyun Luo)