AMIA Annu Symp Proc 2013:90.

Network Visualization of UMLS Source Vocabularies using Semantic Groups

Thai Le[1,2], Bastien Rance, PhD[2], Olivier Bodenreider, MD, PhD[2]
[1] Biomedical and Health Informatics, University of Washington, Seattle, WA, USA;
[2] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract

Exploring UMLS source vocabularies can be challenging for non-specialists. In this poster, we generated network graphs to visualize the content of medical terminologies. We built a graph containing UMLS source vocabularies and semantic groups, where edges are created between a source and its significant semantic group(s). Significance was calculated using the frequency of the semantic groups in the source vocabulary and may be adjusted.

Description

The Unified Medical Language System (UMLS) contains over 160 source vocabularies covering the broad range of terminologies in biomedical information systems. We present a content-based visualization of these source vocabularies through bipartite network graphs consisting of source vocabularies and semantic groups as nodes. An edge is created between a source and a semantic group if the proportion of concepts from this semantic group in the source exceeds a given threshold. The figures below show parts of two networks, for two distinct thresholds. In these graphs, the size of a node is proportional to the logarithm of the number of concepts of the source (or semantic group). Dark red nodes correspond to semantic groups and yellow nodes to UMLS source terminologies. Only sources larger than 1000 concepts are displayed for simplicity.

In figure 1a, the threshold for the creation of an edge is 25%. For example, NDF-RT (circled in green) contains at least 25% of concepts from the semantic groups *Chemicals* (CHEM) and *Disorders* (DISO). Most source vocabularies are linked to only one or two main semantic groups (e.g. *Disorders* for MEDLINEPLUS and OMIM; *Procedures* (GENE) for ICD10-PCS and CPT). In fact, this threshold provides a high-level overview of the sources, but does not reflect well the difference between multi-domain vocabularies (e.g. SNOMED CT), and specialized ones (e.g. ICD).

In figure 1b, the threshold is set at 5%. Expectedly, graph 1b contains noticeably more edges than graph 1a. Nodes highlighted in light pink show the semantic groups linked to a particular source, here CSP (circle in blue), i.e., the semantic groups from which the proportion of CSP concepts is at least 5%. The network obtained with this threshold paints a more detailed picture of the content of terminologies, but is more complex and harder to read.

The network visualization helps UMLS users in identifying the content of the sources. By varying the threshold, the user can either focus on the main semantics of the UMLS sources, or obtain a more detailed picture of the sources.

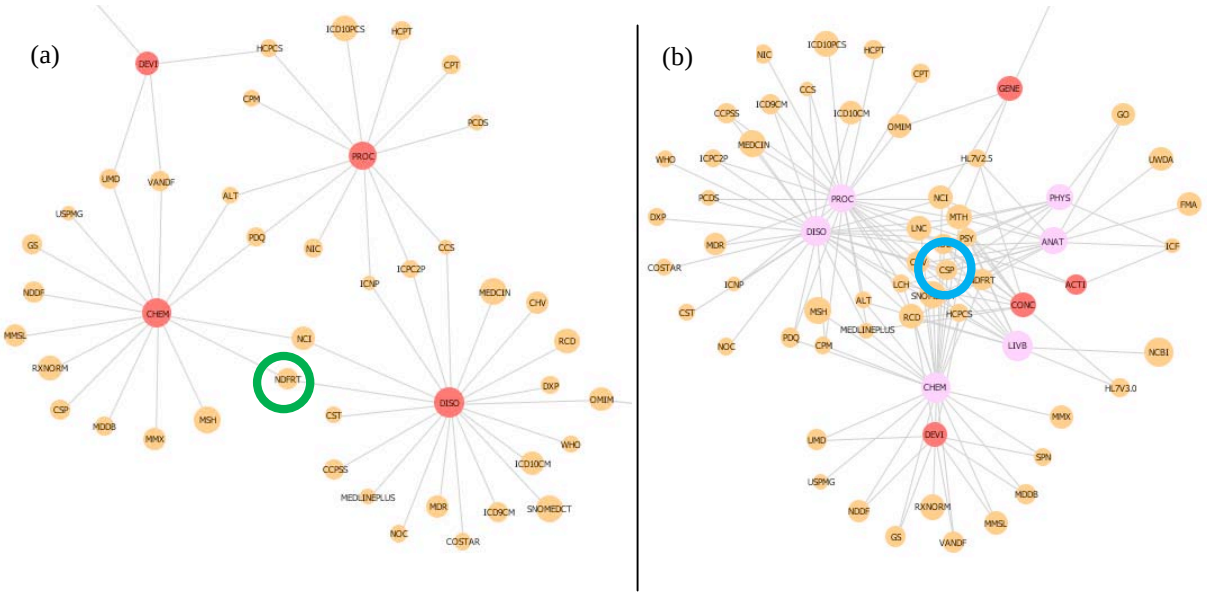

Figure 1. Network visualization of UMLS source vocabularies and associated semantic groups (a) with a threshold of 25%, (b) with a threshold of 5%. (Refer to the text for details)

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.