Submitted to the AMIA Annual Symposium 2012.
draft material -- please do not cite

Automatic Classification and Visualization of UMLS Source Vocabularies through Semantic Group Profiles

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

Abstract

The Unified Medical Language System® (UMLS) is a comprehensive terminology integration system designed to support the development of electronic information systems. The UMLS integrates 161 source vocabularies, though for a given purpose, a developer may not need every vocabulary. With the breadth of vocabularies available, there is a need for classifying the UMLS source vocabularies with respect to their content. We describe a technique for automatic content-based classification of source vocabularies using semantic group profiles. We also present different graphical representations of the source vocabularies, in order to facilitate the comparison and selection of sources. Finally, we compare the content-based classification to a manually created usage-based classification. Our classification can be easily produced for upcoming versions of the UMLS and is being integrated as part of the UMLS documentation.

Introduction

The Unified Medical Language System® (UMLS) is a terminology integration system. It integrates over 2.6 million concepts from 161 source vocabularies ranging in function and content[1]. The UMLS provides broad coverage of the biomedical domain, from disorders to procedure to drugs to anatomical structures. While some source vocabularies focus on a subdomain of biomedicine (e.g., RxNorm for drugs), other source vocabularies, such as SNOMED CT and the NCI Thesaurus, are comprehensive vocabularies covering many different subdomains. However, in the absence of a classification of the source vocabularies, their diversity can make it challenging for users to select appropriate sources for use. This is especially true if users are unfamiliar with certain source vocabularies.

Actually, the UMLS now offers a classification of source vocabularies based on usage (see Table 1). The categories are drawn from MeSH Headings or MeSH Entry Terms, in order to support a functional classification of sources. Some categories such as “Nursing” and “Complementary Therapies” reflect usage, whereas the categories “Disease” and “Procedures” emphasize the content. Source vocabularies may be classified into more than one category. This classification has been established manually and was not publicly available in the summer of 2011 at the time we started our work. Moreover, not all source vocabularies are currently classified with these functional categories. For these reasons, we could not use this classification.

The BioPortal also offers a classification of its 300 ontologies into 39 categories (e.g. Anatomy, Cellular anatomy, Subcellular anatomy, see Table 2) for the purpose of selecting ontologies. There is a limited overlap between the UMLS and BioPortal categories. This is not surprising as usage categories are a subjective view on terminologies driven by users’ habits and because a single terminology can be considered from multiple perspectives. These classifications are helpful for the user but have limitations: prior knowledge about the usage of the terminologies is needed, all the existing vocabularies may need to be reclassified when a new category is introduced, and new vocabularies may not fit into existing categories.

The objective of this work is to explore automatic methods for the classification and visualization of UMLS source vocabularies based on their content. More specifically, we create semantic profiles for each UMLS source vocabulary by leveraging the categorization of UMLS concepts to semantic groups. These semantic group profiles form the basis for classifying and comparing the source vocabularies based on their content. Our approach is fully automatic, does not require any additional knowledge about the source vocabularies, and can be easily deployed. We also explore several visualization techniques to render this classification.

Background

UMLS. The Unified Medical Language System® (UMLS) is assembled by integrating 161 source vocabularies. The UMLS Metathesaurus contains about 2.6 million concepts, i.e., clusters of synonymous terms coming from various source vocabularies. The UMLS Semantic Network is a much smaller network of 133 semantic types organized in a tree structure. Each Metathesaurus concept is assigned at least one semantic type. There exists a further categorization of semantic types into fifteen semantic groups, which represent subdomains of biomedicine, such as Anatomy, Chemicals & Drugs, and Disorders. Every semantic type is categorized into only one semantic group. The fifteen semantic groups and the distribution of concepts in each group are displayed in Table 3.

Many concepts have more than one semantic type; however, these multiple semantic types are generally categorized into the same semantic group. Therefore most concepts are categorized by only one semantic group. In fact, only 1,055 concepts have multiple semantic groups. As a result, the fifteen semantic groups form partition for 99.96% of all UMLS concepts, and are thus virtually disjoint. For the purpose of computing the distribution of the concepts from a source vocabulary into semantic groups, the concepts that have multiple semantic groups should logically not be counted more than once. In practice, these concepts are so few in the UMLS that the effect of double-counting them has no significant effect on the frequency distributions.

Visualization and cognition. We also present different graphical representations of source vocabularies based on semantic group content. There exists a broad body of literature describing the impact of visual displays on not only the speed of decision making, but also its accuracy[2-5]. Cognitive theories provide context on the processes involved in visualizing information. This can range from the theories of Cleveland and McGill, who propose a set of elementary visual tasks for interpreting displays, to Pinker's models of cognitive processing from raw visual information to encoded visual descriptions[6, 7]. Visualization of our work is an important component of presenting the information in a succinct manner to facilitate use by a broad range of stakeholders within the biomedical community.

Methods

Our method for classifying UMLS source vocabularies based on their content can be summarized as follows. We first select representative source vocabularies from the UMLS. Then we create vectors of semantic groups for each source vocabulary and we group similar vocabularies using hierarchical clustering. Finally, we compare the content-based classification to a usage-based classification we established manually, and we explore visualization techniques for representing the source vocabularies, individually and as groups.

Selecting UMLS source vocabularies for analysis. Our analysis is applicable to all 161 sources in the 2011AB edition of the UMLS Metathesaurus. However, in order to present more meaningful results, we performed our analysis on a limited set of 57 vocabularies. We filtered out non-English vocabularies. While translations of vocabularies contain new labels for concepts, their semantic content is identical to that of their English source. We also filtered vocabularies with fewer than 1,000 concepts since their small size limits their overall significance.

Creating vectors of semantic groups for the UMLS source vocabularies. For each UMLS source vocabulary, we compute the frequency distribution of its concepts among the 15 semantic groups, which we record in a 15 dimensional vector. This is what we call the semantic group profile of a source vocabulary. For example, as shown in Table 4, 99% of the concepts from the Foundational Model of Anatomy (FMA) belong to the semantic group Anatomy (ANAT). Its semantic profile is sparse, with few groups other than Anatomy having a value other than 0. In contrast, concepts from the Thesaurus of Psychological Index Terms (PSY) span a variety of semantic groups, with a concentration in the groups Concepts & Ideas (CONC), Disorders (DISO), Physiology (PHYS), and Procedures (PROC).

We generated a heatmap to visualize the semantic group profiles of the source vocabularies in the UMLS, similar to the heatmaps used for the representation of gene expression data**.** As shown in Figure 1, the heatmap has an axis for source vocabularies and one for semantic groups. The intensity of each cell in the heatmap is proportional to the percentage of concepts in a source that belong to a given semantic group, with red representing the highest percentage and yellow the lowest**.** As a result, a column along the heatmap represents the semantic group profile for a given source vocabulary.

Assessing similarity among UMLS source vocabularies through their semantic group profiles. In order to calculate the distance between two semantic group profiles, we used a Euclidian distance metric, that is, the straight line distance between two vectors. (We also tested other metrics including cosine similarity, Jaccard similarity, and Dice's Coefficient. However, the Euclidian distance provided a range of values more suitable for defining groups of source vocabularies using hierarchical clustering.) We calculated the Euclidian distance between each pair of source vocabularies to generate a distance matrix.

We used an agglomerative method of hierarchical clustering to group together similar semantic group profiles. The agglomerative hierarchical clustering algorithm starts with a distance matrix and identifies the pair of source vocabularies that are the most similar. This forms the first cluster. The distance matrix is then recalculated, with complete linkage defining the distance between clusters as the largest distance between any two of its elements. The elements of the matrix are compared to find the next closest pair between sources or clusters. This is repeated until a single agglomerative cluster of all source vocabularies is formed. Figure 2 shows hierarchical clustering in action. Starting with a 3x3 matrix, FMA and UWDA are identified as being the most closely similar (distance = .236) and aggregated. As a result, at the next cycle, the matrix has become a 2x2 matrix. The distance between the aggregated FMA/UWDA and PSY is 100.2 and these two sources are aggregated, because there is no other source to which they would be most similar.

We generated a dendrogram to visualize the hierarchical clustering of the 57 source vocabularies (Figure 3). Longer branches of the tree represent more dissimilar source vocabularies. The dendrogram was cut to define eight clusters, which correspond to both the optimal threshold of discrimination between clusters and a reasonable number of categories for classifying 57 sources. These clusters are displayed in different colors above the heatmap and represent a content-based classification of source vocabularies using semantic groups.

Establishing the usage-based classification of the UMLS source vocabularies. As mentioned earlier, the functional classification of the UMLS source vocabularies available on the UMLS website was not available at the time we started our work, which is the reason why we created our own functional classification. Moreover, the classification provided by the UMLS does not cover all source vocabularies and the number of categories (19) is not suitable for the classification of the 57 source vocabularies under investigation in this study. Using the "purpose" section of the description of the source vocabularies, we identified 8 categories (Patient Care, Health Services Billing, Public Health Statistics, Indexing and Cataloguing Biomedical Literature, Basic Research, Clinical Research, Health Services Research, and Nursing) and manually assigned each source vocabulary to one of the categories.

Comparing the usage-based and content-based classifications. For each source vocabulary, we have the functional classification performed manually (8 categories) and the content-based classification derived from the semantic group profiles through hierarchical clustering (8 clusters). In order to compare the usage-based and content-based classifications, we simply created a contingency table and recorded the count of source vocabularies categorized into each combination of usage- and content-based groups (see Table 5).

Visualizing UMLS source vocabularies through their semantic group profiles. In order to facilitate the exploration and selection of source vocabularies, we provide two main types of visual representations. On the one hand, we created individual semantic group profiles to display the content of a given source with respect to semantic groups. On the other hand, we used network visualization to provide an overview of all the sources from the perspective of their relation to semantic groups.

Individual semantic group profiles for a given source vocabulary are visualized using star diagrams. The source vocabulary is at the center of the diagram, with fifteen evenly spaced semantic groups on the periphery, separated by an equal radial distance from the center. Arrows emanate from the source vocabulary to each of the semantic groups, whose length is proportional to the percentage distribution of concepts from the source vocabulary that belong to this semantic group.

We applied a *network visualization* of semantic group profiles for visual display of content across multiple source vocabularies. This was implemented as a bipartite graph (with two classes of nodes) taken from classical social network analysis. One class of nodes represents the set of all fifteen semantic groups, while the other class of nodes represents the set of all source vocabularies. In addition, the size of the nodes reflects, at a logarithmic scale, the number of concepts in each source vocabulary, or the number of concepts that belong to each semantic group. (We

used a logarithmic scale to accommodate wide ranges. The number of concepts in source vocabulary ranges from two to 630,000 concepts, while the number of concepts in each semantic group can range from 1,200 to 660,000.) A relationship (i.e., an edge in the graph) is defined between source vocabularies and semantic groups. An edge from source S to semantic group G exists if the source vocabulary contains at least some percentage of concepts in G. Several percentage thresholds were explored at 1%, 5%, 10%, and 15%. We used a Fruchterman-Reingold layout algorithm to arrange nodes and edges for the network.

All statistical analyses and visualizations were performed using the R statistical software with the "plotrix", "gplots", "ca", "igraph", and "calibrate" library packages[8-12].

Results

Similarity among UMLS source vocabularies through their semantic group profiles. The *dendrogram* shown in Figure 3 represents the hierarchical clustering of the 57 source vocabularies, i.e., reflects the similarity among the source vocabularies. For example, two anatomy vocabularies, FMA and UWDA, are grouped together, as are two vocabularies having to do with genes and genetic diseases, HUGO and OMIM. Moreover, the branches for the FMA and UWDA group are very short, denoting high similarity (in content). In contrast, the branches for HUGO (representing genes) and OMIM (representing both genes and genetic disorders) are longer. The dendrogram was cut to define eight clusters, which correspond to both the optimal threshold of discrimination a between clusters and a reasonable number of categories for classifying 57 sources. These clusters are displayed in different colors above the heatmap and represent a content-based classification of source vocabularies using semantic groups. For example, the cluster in light green on the right hand side groups 15 source vocabularies whose content is mostly disorders, including the International Classification of Diseases (ICD10CM, ICD9CM) and MedDRA, the Medical Dictionary for Regulatory Activities, which is a medical terminology used to classify adverse events (MDR). Another large cluster, in dark blue, groups drug vocabularies, including RxNorm, the Multum Medisource Lexicon (MMSL), and First Databank's National Drug Data File Plus (NDDF). Finally, the last large cluster in orange groups comprehensive source vocabularies, whose content spans several semantic groups, including SNOMED CT, the NCI Thesaurus (NCI) and, somewhat surprisingly, the Veterans Health Administration's National Drug File-Reference Terminology (NDFRT). In fact, although primarily a drug vocabulary, NDF-RT also contains disorders (to express the therapeutic intent), as well as other types of entities.

Visualizing UMLS source vocabularies through their semantic group profiles.

Individual semantic group profiles for a given source vocabulary, represented as star diagrams, make it possible to easily compare the distribution of the content of two source vocabularies with respect to semantic groups. As shown in Figure 4, the following three source vocabularies have different profiles. RxNorm is a drug vocabulary and contains only drugs. In contrast, ICNP, a nursing vocabulary, and SNOMED CT, a clinical vocabulary, have closer profiles dominated by disorders and procedures. However, SNOMED CT is more comprehensive as it also includes organisms, anatomical structures and drugs.

The *network analysis* provides a different perspective. Instead of primarily grouping source vocabularies, it relates them through the semantic groups that are predominant for these source vocabularies. In other words, the network graph indicates the main (semantic group) components of the sources. As mentioned earlier, different thresholds can be used for the minimal percentage of concepts in a source vocabulary from a given semantic group required to draw a link between this source vocabulary and the semantic group. In the graph shown in Figure 5, the threshold is 5% (i.e., only semantic groups accounting for at least 5% of the concepts in a source vocabulary will be shown as linked to this source vocabulary). The size of the nodes is proportional to the logarithm of the number of concepts to the source vocabularies. For example, the central group of source vocabularies (e.g., SNOMED CT) corresponds to comprehensive sources linked to groups such as disorders (DISO), procedures (PROC), drugs (CHEM), organisms (LIVB) and anatomical structures (ANAT).

Comparing the usage-based and content-based classifications. The 8x8 contingency table for these two classifications – 8 categories for the functional classification and 8 clusters for the content-based classification – is shown in Table 5. For example, the 15 source vocabularies containing mostly disorders (content-based classification) are associated with the following functional categories: Clinical Research (3), Health Services Billing (2), Health Services Research (1), Indexing and Cataloguing (1), Nursing (2), Patient Care (4), and Public Health

Statistics (2). Conversely, the 15 source vocabularies containing classified under "Clinical Research" are associated with the following content-based categories: drug vocabularies (10), disorders vocabularies (3), Procedures vocabularies (1) and General vocabularies (1).

Discussion

Significance

Our approach to classifying UMLS source vocabularies based on their semantic group profile is fully automated and does not require any other information than what is present in the UMLS. Although we applied it to a subset of 57 UMLS source vocabularies (in order to provide more meaningful results), our method is applicable to all source vocabularies in the UMLS, as it relies on the semantic group categorization, which is provided for all concepts regardless of their origin. In contrast, the usage-based classification (in 19 categories) made available as part of the UMLS documentation is developed manually and is provided for only 46 of the 161 source vocabularies. Our method can be easily implemented and the graphical representations can be generated completely automatically for each new version of the UMLS.

Moreover, unlike the usage-based classification, our content-based classification does not require a human judgment and provides an objective perspective on the content of the source vocabularies. Finally, another advantage of our method is that, because it is a vector-based representation of the source vocabularies, it lends itself nicely to visual representation. Arguably, the star diagrams representing individual semantic group profiles allow for an easy exploration of the sources and for easy discrimination between sources. The network representation complements the visualization by providing an overview of the sources and groupings thereof.

Limitations and Future Work

The assignment of a semantic type to a UMLS concept is sometimes subjective and can be arguable. Many concepts are categorized with multiple semantic types. In contrast, all UMLS concepts are categorized in 15 disjoint semantic groups. (Less than 0.05% of the concepts of the UMLS are assigned to more than one semantic group.) Because the semantic groups are broader, the assignment of concept to a group is less likely to be arguable. However, some groups can be viewed as too general for this application. For example, the semantic group "Chemicals" contains both drugs and other chemicals. A user could be interested in retrieving drug vocabularies, rather all chemical vocabularies. As suggested in[13], the grouping of semantic types into semantic groups could be modified to fit the requirements of a particular application.

Generalization

Our approach could conceivably be used for other ontology repositories than the UMLS (e.g., BioPortal). The requirement that the concepts be categorized with semantic groups can somewhat be alleviated. First we could use the mappings found in the BioPortal to propagate the UMLS semantic group assignment to concepts from other ontologies. Second, since our approach is based on the overall proportion of each semantic type, it is sufficient that a representative sample of the concepts in an ontology have semantic group assignment.

Conclusion

The growth of the UMLS makes it difficult for users to select appropriate source vocabularies for a given purpose. In this article, we present a new method to classify biomedical source vocabularies based on their content. We leverage the high level semantic categorization of concepts in semantic groups to create a profile for each source vocabulary. Our approach is completely automated and can easily be applied to all source vocabularies in the UMLS, including upcoming versions of the UMLS.

To assist the user in the exploration of available source vocabularies, we propose to use several graphical representations illustrating the individual content of source vocabularies (star diagrams, heatmaps), as well as the relations among source vocabularies (dendrogram, network). We are currently collaborating with the UMLS team to

add the graphical representations to the UMLS documentation, as a complement to the classification they already provide.

Acknowledgments
This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). Funding for TL was provided through the NLM Training Grant T15 LM007442-07.

References

[1] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res l2004;32: D267-70.
[2] Elting LS, Martin CG, Cantor SB, Rubenstein EB. Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures. BMJ l1999;318: 1527-31.
[3] Feldman-Stewart D, Brundage MD, Zotov V. Further insight into the perception of quantitative information: Judgments of gist in treatment decisions. Medical Decision Making l2007;27: 34-43.
[4] Hoeke JO, Bonke B, van Strik R, Gelsema ES. Evaluation of techniques for the presentation of laboratory data: support of pattern recognition. Methods Inf Med l2000;39: 88-92.
[5] Morrow DG, Hier CM, Menard WE, Leirer VO. Icons improve older and younger adults' comprehension of medication information. J Gerontol B Psychol Sci Soc Sci l1998;53: P240-54.
[6] Cleveland WS, Mcgill R. Graphical Perception - Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association l1984;79: 531-554.
[7] Pinker S. A Theory of Graph Comprehension: Erlbaum; 1990.
[8] Csardi G NT. *InterJournal* l2006;Complex Systems 1695.
[9] Greenacre M NO. Simple, Multiple and Joint Correspondence Analysis. In; 2010.
[10] J G. The calibrate package. l2009.
[11] Lemons JA. Plotrix: a package in the red light district of R. *R-News* l2006;6: 8-12.
[12] Warnes G BB, Lumley T. *gplots: Various R programming tools for plotting data*. In: *R package version 2.6.0*.
[13] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform l 2001;84: 216-20.

Usage categories in the UMLS
Drugs
Procedures
Disease
Insurance Claim Reporting
Diagnosis
Genetics
Nursing
Laboratory Techniques and Procedures
Medical Devices
Adverse Drug Reaction Reporting Systems
Anatomy
Complementary Therapies
Consumer Health Information
Disabled Persons
Subject Headings
Dentistry
Phylogeny

Table 1: Vocabularies categories in the UMLS

Ontology categories in Bioportal			
Anatomy	Dysfunction	Imaging	Plant
Animal Development	Ethology	Immunology	Plant Anatomy
Animal Gross Anatomy	Experimental Conditions	Microbial Anatomy	Plant Development
Arabadopsis	Fish Anatomy	Molecule	Protein
Biological Process	Gene Product	Mouse Anatomy	Subcellular
Biomedical Resources	Genomic and Proteomic	Neurologic Disease	Subcellular anatomy
Cell	Gross Anatomy	Neurological Disorder	Taxonomic Classification
Cellular anatomy	Health	Other	Vocabularies
Chemical	Human	Phenotype	Yeast
Development	Human Developmental Anatomy	Physicochemical	

Table 2: Vocabularies categories in BioPortal

Semantic Group	Abbreviation	No. Concepts	% Concepts
Activities & Behavior	ACTI	5,067	0.19%
Anatomy	ANAT	119,899	4.59%
Chemicals & Drugs	CHEM	593,264	22.71%
Concepts & Ideas	CONC	49,682	1.90%
Devices	DEVI	60,022	2.30%
Disorders	DISO	535,271	20.49%
Genes & Molecular Sequences	GENE	56,286	2.15%
Geographic Areas	GEOG	1,213	0.05%
Living Beings	LIVB	661,612	25.33%
Objects	OBJC	16,245	0.62%
Occupations	OCCU	1,549	0.06%
Organizations	ORGA	2,988	0.11%
Phenomena	PHEN	12,641	0.48%
Physiology	PHYS	139,413	5.34%
Procedures	PROC	357,927	13.70%
Total		2,613,079	100.04%

Table 3: Distribution of concepts within the 15 semantic groups.

Source	CONC	PHEN	CHEM	LIVB	ACTI	DISO	GEOG	ANAT	GENE	OCCU	OBJ	DEVI	ORGA	PHYS	PROC
FMA	.38	0	.55	0	0	.03	0	99	.01	0	.02	0	0	.02	0
PSY	18.23	1.74	8.27	8.89	10.74	14.43	.17	4.72	0.06	2.71	2.63	.51	1.51	13.04	12.36

$\sqrt{(18.23-.38)^2 + (0-1.74)^2 + (8.27-.55)^2 + \ldots\ldots + (13.04-0)^2}$

Euclidian distance = 28.65

Table 4: Example of the semantic group profile for two source vocabularies (Foundational Model of Anatomy and Thesaurus of Psychological Index Terms. A calculation of the Euclidian distance metric between these two profiles is also shown.

	Basic Research	Clinical Research	Health Services Billing	Health Services Research	Indexing and Cataloguing	Nursing	Patient Care	Public Health Statistics	Total
Anatomy (Turquoise)	2								2
Chemicals (Blue)		10	1		1				12
Disorders (Light green)		3	2	1	1	2	4	2	15
Genes (Purple)	2								2
General (Orange)	1	1	1	5	4		3		15
Living Beings (Pink)	1								1
Physiology (Light blue)	1			1			1		3
Procedures (Green)		1	2			1	2		6
Total	7	15	6	7	6	3	10	2	**56**

Table 5: A contingency table providing the count of source vocabularies found in each of the eight usage-based and content-based categorization techniques.

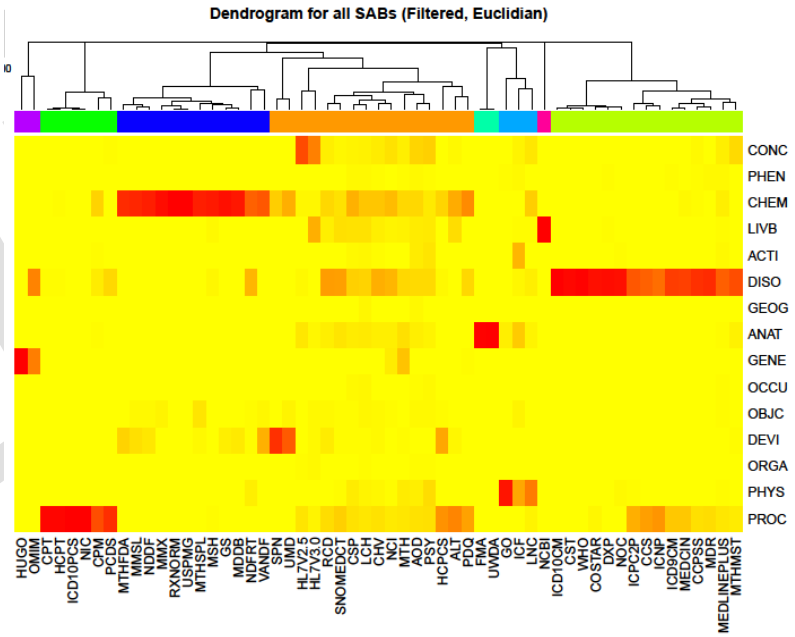


Figure 1: Heatmap of semantic group profiles across a set of filtered source vocabularies[1]. The sources are also hierarchically clustered into eight groups.

[1] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

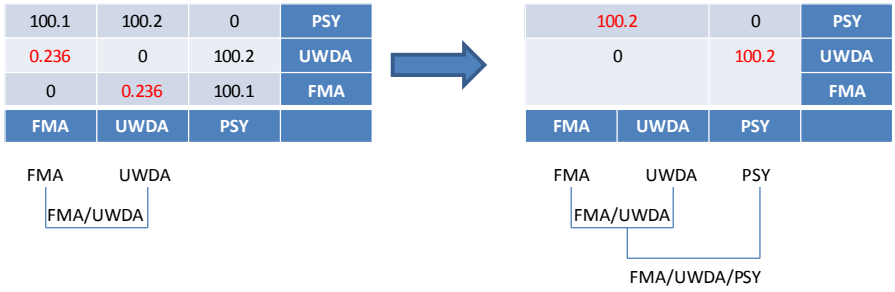


Figure 2: An example calculation shown for hierarchical clustering of the source vocabularies.

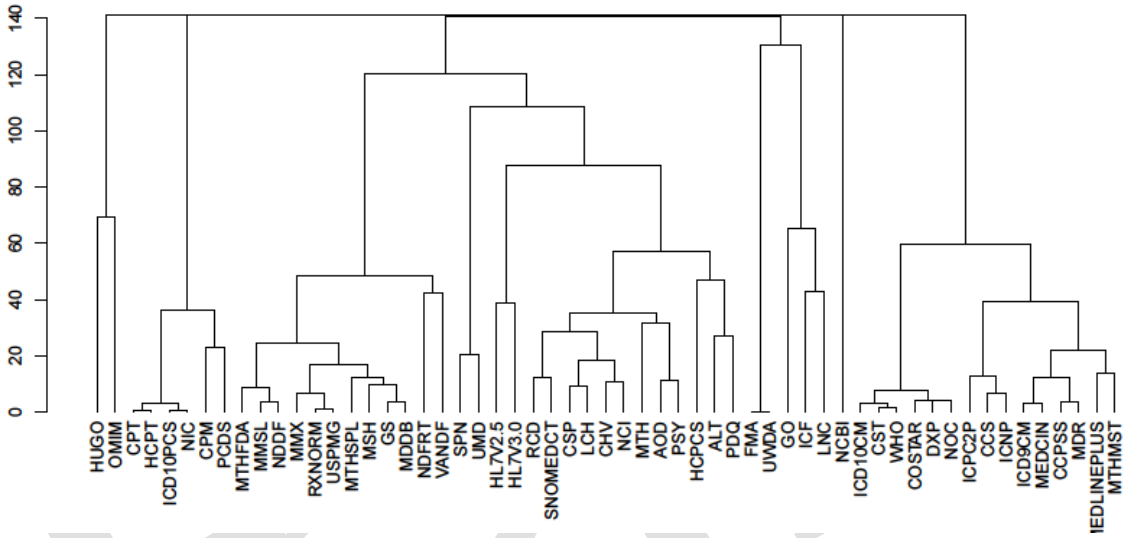
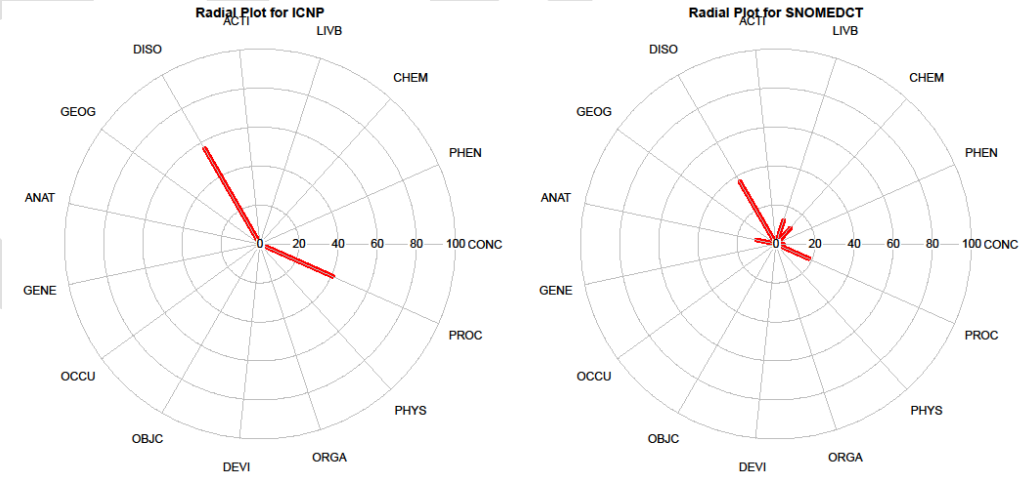


Figure 3: A hierarchical clustering of the source vocabularies[2] based on the Euclidian distance metric.



[2] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

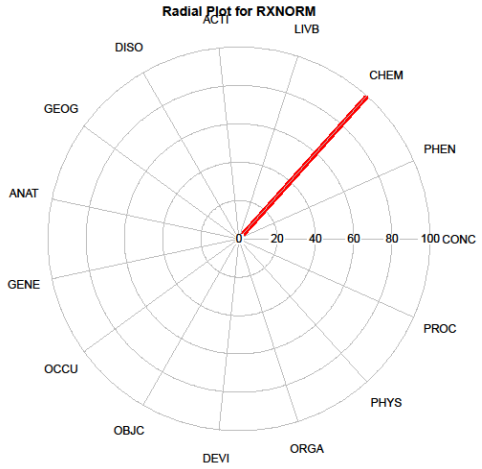


Figure 4: Star diagram plots representing the semantic group content for three different source vocabularies.

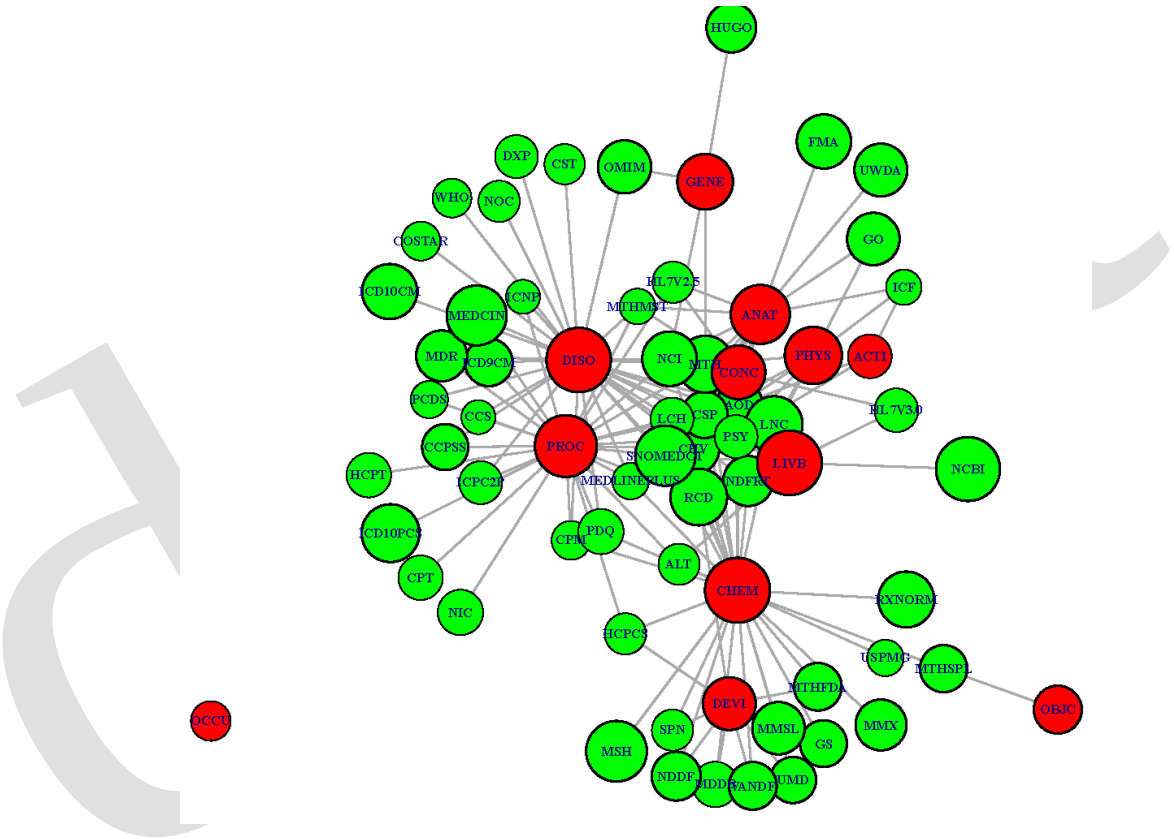


Figure 5: Social network diagram displaying the relationship between source vocabularies[3] and semantic groups. Each node represents either a source vocabulary (in green) or a semantic group (in red). The size of nodes represents the number of concepts within each node on a logarithmic scale. An edge from source vocabulary to semantic group represents the existence of at least 5% of concepts within the source vocabulary belonging in the semantic group.

[3] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html