Comparison of Ontology-based Semantic-Similarity Measures

Wei-Nchih Lee, MD, MPH[1], Nigam Shah, MD, PhD[1], Karanjot Sundlass[2], Mark Musen, MD, PhD[1]
[1]Center for Biomedical Informatics and Research, Stanford University, Stanford, CA;
[2]Medical College of Wisconsin, Milwaukee, WI

Abstract

Semantic-similarity measures quantify concept similarities in a given ontology. Potential applications for these measures include search, data mining, and knowledge discovery in database or decision-support systems that utilize ontologies. To date, there have not been comparisons of the different semantic-similarity approaches on a single ontology. Such a comparison can offer insight on the validity of different approaches. We compared 3 approaches to semantic similarity-metrics (which rely on expert opinion, ontologies only, and information content) with 4 metrics applied to SNOMED-CT. We found that there was poor agreement among those metrics based on information content with the ontology only metric. The metric based only on the ontology structure correlated most with expert opinion. Our results suggest that metrics based on the ontology only may be preferable to information-content–based metrics, and point to the need for more research on validating the different approaches.

Introduction

Currently, genomics data and data repositories in the public domain are expanding at an explosive pace. The wealth of publicly accessible data are beginning to enable cross-cutting integrative translational bioinformatics studies.[1,2] The use of ontologies has become increasingly prominent in the biomedical domain as the contribution of new knowledge from investigators continues to rise. The mining of public resources is facilitated by the use of standard ontologies to annotate the diagnoses, diseases, and experimental conditions of the data sets stored in the public repositories.[2-4]

For example, a researcher studying the allelic variations in a gene would want to know all the pathways that are affected by that gene, the drugs whose effects could be modulated by the allelic variations in the gene, and any disease that could be caused by the gene, and the clinical trials that have studied drugs or diseases related to that gene. The knowledge needed to study such questions is available in public data sets; the challenge is finding that information.

Repositories such as the BioPortal at the National Center for Biomedical Ontology (NCBO) offer tools to researchers looking to use computational methods for both the annotation and search of public datasets.[5] The NCBO creates tools to enable researchers to annotate their data sets—spanning the biological scales from molecular studies to clinical medicine—and ranging from high-throughput experiments to clinical trials and medical imaging. The NCBO has developed a prototype system for ontology-based annotation of the various resource elements consistently to identify the biomedical concepts to which they relate. These resource elements range from experimental data sets in public repositories, to records of disease associations of gene products in mutation databases, to entries of clinical-trial descriptions. The system processes the metadata of elements in biomedical data resources annotating and indexing them with concepts from appropriate ontologies. Creating ontology-based annotations from the metadata in biomedical resources and identifying diagnoses, pathological states, and experimental agents contained in those resources allows indexing of the resources, enabling end users to formulate flexible searching for biomedical data.[1-5]

Semantic similarity refers to the proximity of two concepts within a given ontology. The distance between two concepts is a numerical representation of how far apart two concepts are from one another in some geometric space, and can be considered the inverse of semantic similarity (i.e. if distance between concepts is '0' then the semantic similarity is '1' and vice versa). If this relationship between distance and semantic similarity holds, having similarity or distance metrics allows the use of the ontology to search efficiently for *related* items, or to identify associations between concepts that may not be immediately obvious to the user.

Similarity metrics, then, play a crucial role in searching within the biomedical domain and ensuring that such searches are accurate and return results superior to key-word search[6-8]. They make it possible for machines to perform tasks such as searching,

ranking, and mining of data efficiently, and to perform these tasks across large and diverse data resources.

The assignment of similarity metrics is the subject of much research. The accepted gold standard of similarity is the **agreement among experts** of a given domain, and most derived metrics have been evaluated using this peer-review standard to assess their performance[9-12]. Use of expert opinion, however, is not considered practical, given the obvious limitations of scale for the numerous ontologies available.

An **information-content strategy** to computing semantic similarity assumes that the frequency which one term will appear with another within a given corpus of knowledge (or information) will be related to similarity of the two terms. This approach takes a statistical view of information to determine the closeness of two terms.

Finally, much research has attempted to derive semantic-similarity measures directly from the **knowledge codified in ontologies**. The advantages of using ontologies include re-usability of the coded knowledge across different domains, tractability, and the ability to utilize the derived measures on a large scale. Measures that have used ontologies to derive semantic similarity measures generally use strategies that either measure the length of the path between two concepts,[11,12] or that weight the edges of the path (once it has been identified) with information content derived from a corpus of textual knowledge or information about the number of child nodes in the ontology hierarchy.[11]

How ontology-derived similarity measures compare with human peer review as well as information-content–derived measures has not been well studied. Agreement among each of these approaches would suggest robustness of our current concepts of disease, while disagreement raises the question of the validity of one approach versus another.

Methods

We examined four strategies for measuring semantic similarity with the clinical ontology SNOMED-CT and compared the measures to one another. The metrics were chosen to represent the broad classes of approaches that have been used to measure semantic similarity. One metric strictly uses the structure of the SNOMED-CT ontology, while two others combine an information content approach with path finding. The fourth metric involves a set of similarity measures obtained from physician experts.

Selection of Disease Concepts

We selected a set of 225 disease concepts that are the focus of Butte's[13] Genomic Nosology for Medicine (GNOMED) project. An experienced Internal Medicine physician (WL) reviewed the 225 diseases and chose a convenience sample of 20 diseases (Table 1) that spanned a broad range of disease classes and that were familiar to a physician trained in Internal Medicine or Primary Care Medicine. All the diseases chosen had a unique SNOMED-CT code that identified its place within the ontology.

Table 1. Diseases Selected

Autoimmune Hepatitis	Essential Hypertension
Alzheimer's Disease	Atherosclerosis
Parkinson's Disease	COPD[b]
Obstructive Sleep Apnea	Crohn's Disease
H. Pylori Gastritis	Pulmonary Emphysema
Cardiomyopathies	Allergic Asthma
Diabetic Nephropathy	GERD[c]
Type II Diabetes M.	Hepatitis C
LVH[a]	Fatty Liver
Hyperlipidemia	Congestive Heart Failure

[a]Left Ventricular Hypertrophy, [b]Chronic Obstructive Pulmonary Disease, [c]Gastroesophageal Reflux Disease

Survey Design and Implementation

From the 20 diseases, all the possible disease – disease pair-wise combinations were made, for a total of 190 distinct diseases pairs. We designed 19 surveys of 10 pair-wise combinations each, to cover the entire set of disease pairs. Each survey asked a respondent to rate the similarity between the diseases in each disease pair with a 7-point Likert scale of least similar to most similar.[14]

Initial surveys were pilot tested with a convenience sample of 15 physicians, who provided feedback on wording, scaling measures, and overall design. Feedback was reviewed by the investigators and appropriate changes were made to the surveys before data collection.

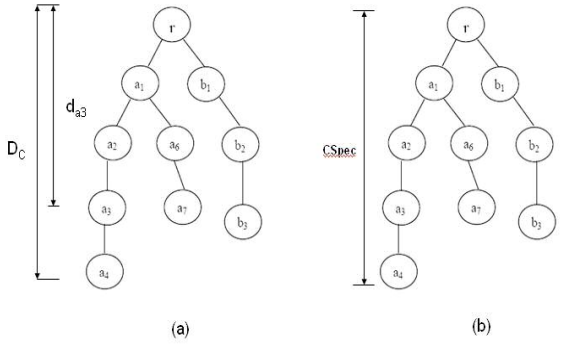
Physicians then were recruited from two sources. The first was from an Internal Medicine residency training program in an academic tertiary care medical center in New York. The second was from Sermo (www.sermo.com), an online physician community that represents more than 40,000 licensed physicians

in the United States. Physicians who were recruited from Sermo were given online versions of the survey, which was created by one of the authors (WL).

Semantic Similarity Measures

We selected 3 semantic similarity metrics for comparison with the similarity scores provided by the physicians. The first, an ontology-only based approach, is represented by Al-Mubaid et al., who used the features of a concept's location within a cluster of nodes to modify the shortest path distance between two concepts[10] (Figure 1). The local granularity of a concept (Fig 1a) is given by the ratio of the depth of that concept within a cluster (d_{a3}) to the depth of the cluster (D_C). The common specificity (CSpec)between two concepts is a measure of the shared information between two concepts – it is the depth of the cluster defined by the least common subsumer of the two concepts (Fig 1b).

Figure 1 Concept clusters within an ontology



The weighted contributions of local granularity and CSpec (α and β respectively), is assumed to be equal (with a value of 1.0), and the distance between two concepts given by:

$$Sem(C_1,C_2) = \log((Path - 1)^{\alpha} \times (Cspec)^{\beta} + k)$$

where *Path* is the shortest path length between the two concepts.[10]

To represent the information content strategies, we used two distance metrics defined by Melton – Descendant Distance (DD) and Term Frequency (TF).[11] These measured the clinical distance between two patients based on the diseases present in each, and are analagous to distances between a pair of diseases. The generalization of their distance metric is:

$$D_{Pa,Pb} = \frac{\Sigma\, w \min(distance(v_i,v_b))}{\Sigma\, w}$$

where *w* is a weight applied to each link in the path between patient A and patient B, and the term *distance(v_i,v_b)* is the distance between each of the disease nodes assigned as a diagnosis to patient A and a disease node assigned to patient B. For the Descendant Distance (DD) metric, the descendants of a given concept are used to weight the path links from one concept to another. Thus, the weight of each link $w = -\log P(v_i)$, and the distance between nodes[11]:

$$Distance(C_1,C_2) = \Sigma P(v_y)$$

where $P(V_y) = \frac{\text{No. descendants of each path node}}{\text{Total No. of Terms}}$

For our study, each "patient" is one disease concept, so that the term *min(distance(v_i,v_b))* became the shortest path between two disease concepts in the SNOMED-CT ontology. Each link in the path, then, was weighted with the number of descendants for each node in the path, and summed together to obtain the distance metric for DD.

For the Term Frequency (TF) metric, Melton used a concept's term frequency in a corpus of text to weight the path edges within an ontology.[11] The term frequency came from a clinical data repository of more than 14,000 patients to derive a set of all SNOMED-CT concepts used[11]. We modified the weighting scheme of the DD metric, so that

$w = -\log P(v_i)$ where

$$P(V_y) = \frac{\text{Term Frequency of disease concept}}{\text{Max. Term Frequency in Corpus}}$$

Since we did not have a clinical data repository available to us, we used the UMLS Medline index to obtain the requisite term frequencies to weight each of the path links between the disease concepts of interest. The UMLS Medline index has an added advantage over a clinical corpus of text in that the term frequencies within the index are not institution dependent, and are a de facto gold standard of medical terminology.

Analysis

We applied each of the approaches (Expert assessments, Cluster-Based metric, Term-Frequency (TF) Metric, and Descendant-Distance (DD) Metric) to a square symmetric matrix of the 20 diseases in Table 1. Each matrix had 190 distinct pair-wise combinations, and we obtained similarity (or distance) metrics for each cell within the matrices. Physician scores were collated and averaged together for each of the pair-wise combinations of diseases.

We derived a set of scores for each metric for each of the 190 possible pair-wise disease combinations. Because the distributions of the results for the Cluster Based, TF and DD metrics are unknown, we used a

non-parametric rank-correlation test to compare the scores among each set of metrics. The advantage of a rank-correlation test is that it obviates the need for normalization of the data, which would have been difficult without prior knowledge of each metric's distribution.

For our rank-correlation test, we chose Kendall's *tau* to compare the different distance or similarity metrics. Kendall's *tau* provides a correlation coefficient similar to Spearman's rank coefficient, but has the added advantage of being better suited when rank ties occur in two sets of scores. We used a type I error rate of 0.05 when comparing the distances based on one metric to another, with the null hypothesis of no differences between the scores of any two sets of metrics. Statistical tests were performed with the R statistical package.

Table 2. Correlations between each metric approach

	Expert	Cluster	TF	DD
Expert	X	-0.11[a]	0.005[b]	0.009[b]
Cluster	X	X	-0.048[b]	0.016[b]
TF	X	X	X	0.22[c]
DD	X	X	X	X

[a] p value = 0.032; [b] p value non-significant; [c] p value < 0.001

Results

We surveyed a convenience sample of 25 physicians.. With the exception of one surgeon who participated, all physicians were in a primary-care specialty (Internal Medicine, Pediatrics, or Family Practice). The mean years in practice for the physicians was 3.78 (s.d. 5.34) years. 14 (56%) of the physicians were based in academic institutions, 7 (28%) in hospital based practices, and 4 (16%) were in office based practices. 11 (44%) of the physicians practiced > 80% of the time, while 12 (48%) practiced between 20 and 80% of the time, and 2 (8%) practiced less than 20% of the time.

Table 2 shows the results of the comparisons for agreement between the different metrics. As shown, the only significant, albeit weak, correlation to Expert evaluations occurred with the Cluster-Based Ontology-Only approach. A somewhat stronger correlation was found between the TF and DD approaches (which is expected, given that both make similar use of information content in their calculations).

Discussion

Since Rada first described the shortest-path approach to semantic similarity and conceptualized the similarity as a distance,[15] a number of investigators have put forth variations of this theme. The performance of most of these metrics has been assessed against an expert peer-review gold standard. Expert assessments, however, may vary depending on the expert, leading to variability in the performance of the distance metrics. Agreement of the different metric approaches to one another would suggest a robustness of the semantic distance approaches thus far.

In this paper, we sought to compare approaches to semantic similarity (1) that are based on only the ontology structure, (2) that incorporate information content of some corpus of text, and 3) that are based on domain experts. A similar study by Pederson, et al. examined different metric approaches applied to SNOMED-CT.[9] In his study, however, a limited number of disease pairs (30) were assessed, and these pairs were pre-selected on the basis of high agreement among medical coders. As a result, the performance of the distance measures in Pederson's study may be biased toward a stronger correlation effect. Also, like many other authors, Pederson only compared the performance of the distance metrics to the assessments of domain experts – thus lacking a global perspective of how the different semantic distance approaches might compare with one another. Our study improves on Pederson by evaluating a six times larger set (190) of disease-pair combinations. Our study also utilizes a larger group of domain experts than has been used in prior studies, with the advantage that the experts do not come from an informatics background – potentially limiting bias in the similarity assessments.

In our study, we found that the strict ontology-based distance metric (Cluster-Based analysis as by Al-Mubaid) correlated most closely with the domain-expert assessments. This metric, however, did not correlate significantly with the measures that used information content (TF and DD). As expected, the two information-content approaches (TF and DD) did correlate with each other, but not strongly, with a ranked coefficient of 0.22.

The magnitude of the correlation coefficients in our study is less than what has been reported previously. This difference is likely an effect of the sample of the physicians chosen as the domain experts. By choosing physicians from a variety of clinical settings, what we gained in a more representative sample came at the cost of more variability in physician responses, thus blunting the strength of

correlation that we might have found. The use of only primary care diseases in assessing distance metrics may also have biased the results of our study towards Mubaid's metric, which favors the concepts that are already clustered.

Conclusions

The semantic-distance measures (cluster, TF, DD) do not seem to correlate strongly with one another. This result is a bit of a surprise, in that one would expect at least a moderate correlation among the three. TF and DD, which are similar to each other in formulation, share only a modest, albeit statistically significant, correlation with one another. Such poor agreement between two metrics that aim to quantify the same thing – semantic similarity – is of concern, and indicates that more work is needed before notions of similarity based on *distances* can be interpreted as valid surrogates of semantic similarity.

The purely ontology-based measure had a weak, but statistically significant, correlation to expert opinion. Our results suggest that ontology-only approaches to semantic distance may be preferable to information content approaches to semantic distance. The use of an approach based entirely on an ontology structure has the added advantage in that it is not dependent on the size, quality, and availability of a corpus of text, as is the case with information-content approaches.

Finally, until further research sorts out the most appropriate notion of distance to use for semantic similarity, a purely ontology-based approach like Al-Mubaid's will probably be the metric of choice – in searching for *related* items in the tools the NCBO is developing for searching open biomedical resources – because of their simplicity, scalability, highest agreement with expert opinion and independence from needing a text corpus.

Acknowledgments

The author WL is supported by the National Library of Medicine Training Grant (NLM training grant 07033). The study is also supported by the NIH Grant U54 HG004028. We would like to thank Dr. Richard Olshen for statistical guidance.

References

1. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annu Symp Proc. 2006;:106-10.
2. Shah NH, Chiang AP, Butte AJ, et al. Ontology-driven Indexing of Public datasets for Translational Bioinformatics. AMIA Summit on Translational Bioinformatics; 2008; San Franscisco; 2008.
3. Marinelli RJ, Montgomery K, Liu CL, et al. The Stanford Tissue Microarray Database. Nucleic Acids Res. 2008 Jan;36(Database issue):D871-7.
4. Shah NH, Rubin DL, Espinosa I, et al. Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics. 2007 Aug 8;8:296.
5. Rubin DL, Lewis SE, Mungall CJ, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006 Summer; 10(2):185-98.
6. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc. 2007 May-Jun;14(3):253-63.
7. Sneiderman CA, et al. Knowledge-based Methods to Help Clinicians Find Answers in MEDLINE. J Am Med Inform Assoc. 2007;14:772-780.
8. Moskovitch R, Martins SB, Behiri E, et al. A comparative evaluation of full-text, concept-based, and context-sensitive search. J Am Med Inform Assoc. 2007 Mar-Apr;14(2):164-74.
9. Pedersen T, Pakhomov SV, Patwardhan S, et al.. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform. 2007 Jun;40(3):288-99.
10. Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. Conf Proc IEEE Eng Med Biol Soc. 2006;1:2713-7.
11. Melton GB, Parsons S, Morrison FP, et al. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform. 2006 Dec;39(6):697-705.
12. Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform. 2004 Apr;37(2):77-85.
13. http://bmir.stanford.edu/projects/view.php/genomic_nosology_for_medicine_gnomed
14. Weisberg HF, Krosnick JA, Bowen BD. An Introduction to Survey Research, Polling, and Data Analysis 3rd Edition. Sage Publications. London. 1996. pp. 77-102.
15. Rada R, Hafedh M, Bicknell E, Blettner M. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man. and Cybernetics. 1989, 19(1). 17-30.