Evidence of Term-Structure Differences among Folksonomies and Controlled Indexing Languages

Benjamin M. Good[1] and Joseph T. Tennis[2]
[1] Bioinformatics Graduate Program, University of British Columbia
[2] The Information School, University of Washington

Introduction

With the advent of Internet-based technologies for information organization, many groups have constructed their own indexing languages. Biologists, Library and Information Science practitioners, and now social taggers have worked together to create large and many times complex indexing languages. In this environment of diversity, two questions surface: (1) what are the measurable characteristics of these indexing languages, and (2) do measurements of these indexing languages *speciate* along these characteristics? This poster presents data from this exploratory work.

Literature

A growing number of theoretical works have compared folksonomies to controlled indexing languages (Beghtol, 2003; Feinberg, 2006; Tennis, 2006; and Voss, 2006). However, empirical work on the comparative anatomy of these languages has lagged (an exception is Milne, Medelyan, and Witten (2006).

Methodology

Sample

Working primarily in the semantic Web environment, we harvested 25 indexing languages that met three criteria: (1) were freely available, (2) were considered valuable by a body of designers/users (were always used and most often maintained by and given imprimatur of a respected organization), and (3) were used to index or curate documents. These languages spanned what we commonly understand to be thesauri, ontologies, and folksonomies. The domain of the indexing language was of secondary concern for this research question. We accounted for that in our analysis, though it did not shape our sampling procedure. The 25 indexing languages that met these criteria are listed online (Good and Tennis, 2008).

Term Normalization

Here, when discussing indexing languages, we refer only to the set of terms that compose them - the tags in folksonomies, and the concept labels in thesauri and ontologies. Each of the terms in these indexing languages was subjected to a normalization process meant to help consistently delineate the boundaries of compound words. We carried out four procedures of normalization: (1) all non-word characters (e.g. ; , _ , -) were mapped to spaces using a regular expression (\\W), so the term "automatic-ontology_evaluation" would become "automatic ontology evaluation," (2) case-delineated compound words were mapped to space separated words (e.g., "camelCase" becomes "camel case"), (3) all words were made all lower case, and (4) any redundant terms were removed.

Metrics

For the primary analysis, a variety of different measurements were recorded for each term set. These metrics, summarized in Table 1., included indicators of the size of the term sets, the lengths of the terms, and the apparent levels of modularity within the sets. Measures of modularity expose the structure of the term set based on the proportions of multi-word terms and the degrees of sub-term re-use. These measures include two main categories, Observed Linguistic Precoordination (OLP) and Compositionality.

OLP indicates whether a term appears to be a union of multiple terms based on syntactic separators. For example, the MeSH term ‘Fibroblast Growth Factor” would be observed to be a linguistic precoordination of the terms ‘Fibroblast’, ‘Growth’, and ‘Factor’ based on the presence of spaces between the terms. We categorize terms as uniterms (one term), duplets (combinations of two terms), triplets (combinations of three terms) or quadruplets or higher (combinations of four or more terms). Using these categorizations, we also record the ‘flexibility’ of a term set as the fraction of sub-terms (the terms that are used to compose duplets, triplets, and quadplus terms) that also appear as uniterms.

The OLP measurements were adapted from characteristics of indexing languages introduced by Van Slype, who, in the process of comparing thesauri to the ISO Standard, identified a number of simple measures for gauging the extent of a thesaurus (Bureau-Marcel-Van-Dijk, 1976). His measures provide numbers that give the basic extent of these indexing languages. These were proposed as benchmarks for standards revision. They outlined the anatomy of the sample of thesauri in English, French, and German. Our intent in using a subset of these metrics here is to provide a means to generate such an anatomical description of any indexing language.

The OLP measures were extended with related measures of ‘compositionality’ (Ogren, Cohen, Acquaah-Mensah, Eberlein, & Hunter, 2004). Compositionality measures include a) the number of terms that contain another complete term as a proper substring, b) the number of terms that are contained by another term as a proper substring, c) the number of different *complements* used in these compositions, and d) the number of different *compositions* created with each contained term. A *complement* is a subterm that is not itself an independent member of the set of terms. For example, the term set containing the two terms “macrophage” and “derived from macrophage” contains one complement – “derived from”. A *composition* is a combination of one term from the term set with another set of terms (forming the suffix and/or the prefix to this term) to form another term in the set. For example, in the Academic Computing Machinery (ACM) subject listing, the term “software program verification” contains three subterms that are also independent terms (“software”, “program”, and “verification”). According to our definition, this term would be counted as three compositions – “software”+suffix, prefix+”program”+suffix, prefix+”verification”. As another example, the term “denotational semantics” would only result in one composition because “semantics” is an independent term while “denotational” is not (and thus is a *complement* as defined above).

Modularity, though not indicative of conceptual structure or meaning, is indicative of the factors that go into the semantics of an indexing language, and shape its use. Here we are guided by Soergel’s rubric from concept description and semantic factoring. He tells us “we may note that often conceptual structure is reflected in linguistic structure; often multi-word terms do designate a compound concept, and the single terms designate or very nearly designate the semantic factors. Example: Steel pipes = steel:pipes [demonstrating the factoring]. This fact can be used in thesaurus building,” p. 75 (Soergel, 1974). The combinations of terms or the factoring out of semantics is theoretically important for another reason. It shapes the result of indexing, what we call indexes here.

Together, these measurements combine to begin to form a descriptive picture of the anatomy of the many diverse term sets used for indexing. Table 1 lists and provides brief definitions for all of the term set measurements taken.

Table 1. Parameters of Indexing Languages

Metric	Definition
Number distinct terms	The number of syntactically unique terms in the set.

Term Length	The length of the terms in the set. We report the mean, minimum, maximum, median, standard deviation, skewness, and coefficient of variation for the term lengths in a term set.
OLP uniterms, duplets, triplets, quadplus	We report both the total number and the fraction of each of these categories in the whole term set.
OLP flexibility	The fraction of OLP sub-terms (the independent terms that are used to compose precoordinated terms) that also appear as uniterms
OLP number subterms per term	The number of subterms per term is zero for a uniterm ("gene"), two for a duplet ("gene ontology"), three for a triplet ("cell biology class"), and so on. We report the mean, max, minimum, and median number of subterms per term in a term set.
contains another	The terms that contain another term from the same set. Both the total and the proportion of terms that contain another are reported
contained by another	The terms that are contained by another term from the same set. Both the total and the proportion of terms that are contained by another are reported
complements	A *complement* is a subterm that is not itself an independent member of the set of terms. The total number of distinct complements is reported
compositions	A *composition* is a combination of one term from the term set with another set of terms (forming the suffix and/or the prefix to this term) to form another term in the set. The total number of compositions is reported.

Analysis

To enable visualizations of the relationships between data types of varying dimensions, the non-ratio data, such as the size of the term sets, was log-transformed and then mapped to a 0-1 scale by dividing each value by the largest number in the set. The variables were then plotted on radar graphs to provide a clear, visual representation of the distinct *shapes* of these indexing languages.

In addition, we applied cluster analysis to the normalized data using seven variables that were, upon earlier inspection, highly variable in the sample. Those variables were: % of uniterms, % of duplets, flexibility, % contained by another, standard deviation of term length, skewness of term length, and number of complements. We used Ward's[i] method of cluster analysis using SPSS. This allowed us to create a dendrogram illustrating the clusters of normalized indexing languages.

Findings

Examining the radar graphs of the normalized indexing languages we can see that the different groups of indexing languages investigated here display particular shapes that correspond to varying extents to the kinds of systems they arose from. For example, as Figure 1 illustrates, we observe distinct shapes for folksonomies and for subsets of the controlled languages. Interestingly, though maintaining the basics of the folksonomy shape, the Connotea folksonomy appears much more similar to the controlled vocabularies then the other folksonomies do. In addition, the term sets gathered from the Open Biomedical Ontologies Foundry display a quite unique shape in comparison to the others (Smith et al., 2007).

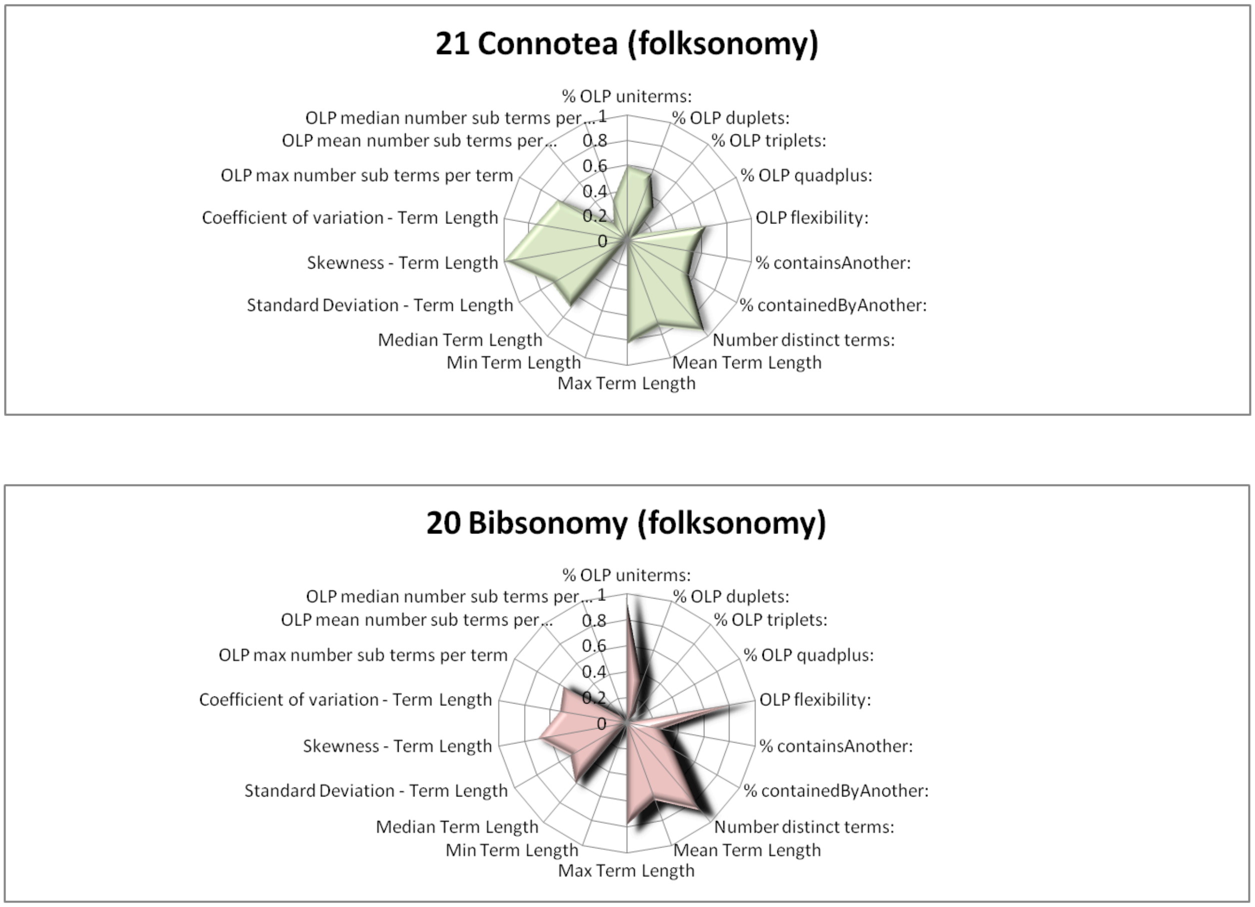

Figure 1. Radar graphs of normalized indexing languages

Figure 2 shows how the clusters formed from the seven selected variables and the 25 normalized indexing languages indicate, as expected, that the folksonomies form a separate group from the controlled indexing languages. In addition, they show that two distinct subsets exist within the group of controlled languages.

```
Dendrogram using Ward Method

                      Rescaled Distance Cluster Combine

CASE     0         5        10        15        20        25
  Num   +---------+---------+---------+---------+---------+
```

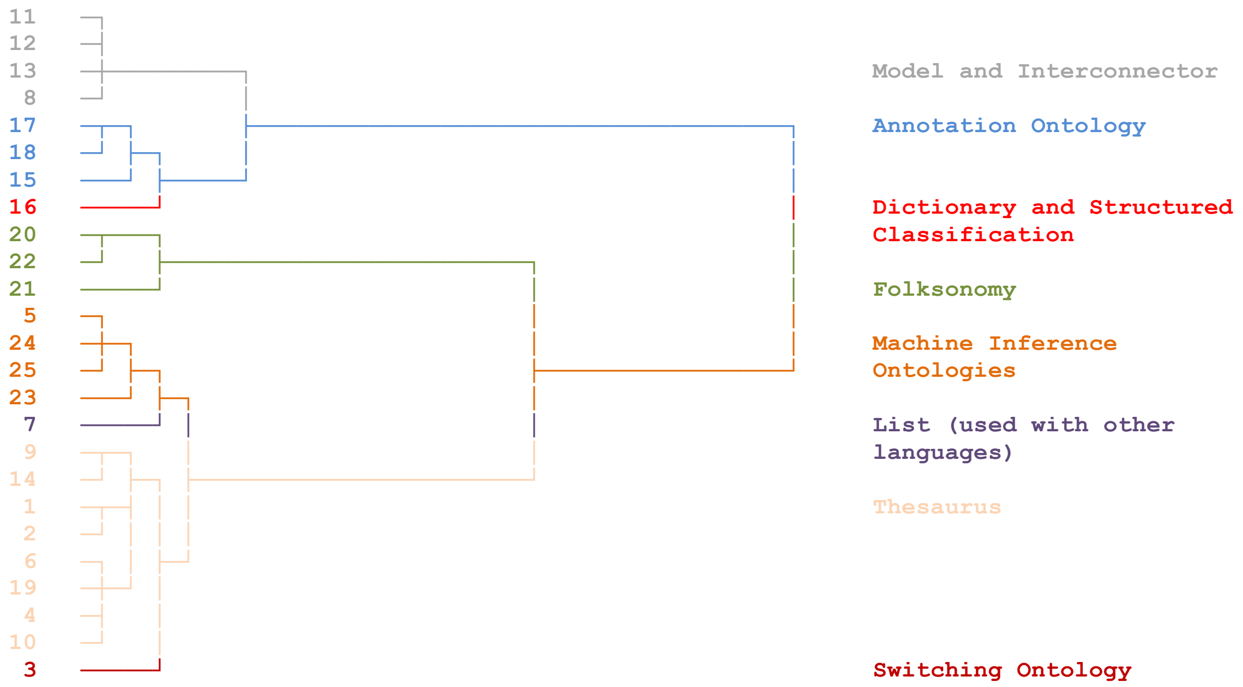

Figure 2. Dendrogram 1 of indexing language clusters

Conclusion

The primary contribution of this work is the development of a framework for the empirical comparison of the terms from different indexing languages. This framework provides a conceptual foundation for data-based investigations of the tools and products of emerging indexing practices that complements ongoing theoretical work. In the exploratory results presented here, we use the lens provided by this framework to visualize the shapes of the different species of indexing languages. Through this lens, we see precisely how the tags of folksonomies differ from the terms of controlled indexing languages as well as how different subsets of indexing languages differ from each other. By defining and visualizing the axes of comparison discussed above and to be displayed in the poster, we provide a solid foundation for understanding how different indexing languages relate to one another and thus a) how they might best be used in combination and b) how the work practices, purpose, software etc. used to create them effect their final form.

References

Beghtol, C. (2003). Classification for information retrieval and classification for knowledge discovery: Relationships between professional and naive classifications. Knowledge Organization*,* 30(2), 64-73.

Bureau-Marcel-Van-Dijk. (1976). Definition of Thesauri Essential Characteristics. (study carried out by G. Van Slype) (Vol. 2). Brussels.

Feinberg, M. (2006, January 01). An Examination of Authority in Social Classification Systems. Paper presented at the Proceedings 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research.

Good, B.M., & Tennis, J.T. (2008). Term Set Sources Retrieved May 29, 2008, from http://bioinfo.icapture.ubc.ca/bgood/comparisons/termsetlist.html

Milne, D., Medelyan, O., & Witten, I. (2006). Mining Domain-Specific Thesauri from Wikipedia: A case study (pp. 442-448). Paper presented at the EEE/WIC/ACM International Conference on Web Intelligence.

Ogren, P.V., Cohen, K.B., Acquaah-Mensah, G.K., Eberlein, J., & Hunter, L. (2004). In The compositional structure of Gene Ontology terms (pp. 214-225). Paper presented at the Pacific Symposium on Biocomputing.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol*,* 25(11), 1251-1255.

Soergel, D. (1974). Indexing languages and thesauri: construction and maintenance. Los Angeles: Melville Pub. Co.

Tennis, J.T. (2006). In J.T.T. Jonathan Furner (Ed.), Social Tagging and the Next Steps for Indexing. Paper presented at the 17th ASIS&T SIG/CR Classification Research Workshop, Austin, Texas.

Voss, J. (2006). Collaborative Thesaurus Tagging the Wikipedia Way. Retrieved March 1, 2007, from http://arxiv.org/abs/cs.IR/0604036

Notes

[i] Ward's method is a minimum distance hierarchical method which calculates the sum of squared Euclidean distances from each case in a cluster to the mean of all variables. The cluster to be merged is the one which will increase the sum the least. That is, this method minimizes the sum of squares of any pair of clusters to be formed at a given step. As such it is an ANOVA-type approach which maximizes between group differences and minimizes within-group distances, optimizing the F statistic. This method tends to create clusters of small size.