Allen, R.B. and Wu, Y., Metrics of the Scope of a Collection, JASIST, 55 (10), 2005, 1243-1249.

Metrics for the Scope of a Collection

Robert B. Allen and Yejun Wu
College of Information Studies, University of Maryland, College Park, MD 20742, USA
allen@acm.org and wuyj@glue.umd.edu

Some collections cover many topics, while others are narrowly focused on a limited number of topics. We introduce the concept of the "scope" of a collection of documents and we compare two ways of measuring it. These measures are based on the distances among documents. The first uses the overlap of words between pairs of documents. The second measure uses a novel method which calculates the semantic relatedness to pairs of words from the documents. Those values are combined to obtain an overall distance between the documents. The main validation for the measures compared Web pages categorized by Yahoo. Sets of pages sampled from broad categories were determined to have a higher scope than sets derived from sub-categories. The measure was significant and confirmed the expected difference in scope. Finally, we discuss other measures related to scope.

The Idea of Collection Scope

A collection of documents may consist of several thousand or even several million documents. Some collections, such as collections of dissertations across disciplines from an entire university (e.g., [11]) include many diverse topics while a collection of dissertations on Biochemistry would be relatively narrowly focused. Similarly, one collection might cover the broad topic of Agriculture while another collection might deal only with Agricultural Pests. We define "scope" as the range of topics covered by a collection of documents. One application for such a measure should be to support "collection selection". Users need to know the characteristics of the collections from which they can choose.

Traditionally, a collection is described with a "conspectus" of "front matter". While there has been an initial effort to develop a conspectus suitable for online collections [2], that effort required human specification and, in any event, has not been actively pursued. Other work has examined scope for a book collection by the range of classification categories covered [19]. We prefer statistics that can be calculated automatically and made available upon request in the spirit of Harvest [3] and Open Archives [6]. GlOSS [5] characterizes a collection by its centroid but that is only a single point and it does not measure the collection's scope. The strategy here is to measure "scope" from the aggregate distance between the documents in the collection. Furthermore, we believed that our measure should be able to be interpreted on an absolute scale so that different collections can be readily compared.

Word Overlaps as a Measure of the Distance Between Documents

As a simple statistic of document similarity, we counted the overlap of content words between pairs of documents.

Documents

Ten articles were selected from the online *Encyclopedia Britannica*: *crime and punishment*, *criminal law*, *drug*, *evolution*, *human disease – health versus disease*, *maintenance of health*, *human evolution*, *legal profession*, *history of medicine*, *medicine*, and *nutrition*. Some of these documents were informally judged to be highly similar in content with each other while some were not. The documents were fairly long with an average of 8231 words ($\sigma = 3723$).

The documents were processed to produce word lists. Capitalization was suppressed and all HTML tags, punctuation, sidebars, and images were removed. A stop word list of the top 300 words from the Brown Corpus was applied. The resulting word lists were truncated to the 40 words with the highest term frequency. We decided to use 40 terms since it was enough to give statistical stability but did not include many instances of very low frequency, and possibly a-typical, words from the lists. For tied ranks, the words were taken in alphabetical order.

Human Ratings of Inter-Document Distance

The ten *Britannica* articles were presented to 5 participants who were students at the College of Information Studies. The participants were instructed to judge the semantic similarity between each document pair using a scale between 0 (for least similar pairs) and 10 (for most similar pairs). Each participant was paid $24 for the two-hour study. The average ratings are shown in Table 1.

Documents	1	2	3	4	5	6	7	8	9	10
1. crime/punishment	-	8.0	2.0	1.0	1.0	0.9	1.0	4.4	2.2	0.7
2. criminal law		-	1.4	1.4	1.4	1.3	0.8	5.6	1.0	0.0
3. disease study			-	6.0	4.0	8.2	4.6	0.0	5.0	4.1
4. drugs				-	2.8	4.0	1.9	1.4	5.2	5.2
5. evolution					-	5.4	8.4	0.8	1.3	3.6
6. human disease						-	5.0	0.6	5.8	4.5
7. human evolution							-	0.8	2.4	4.5
8. legal profession								-	2.4	0.2
9. history of medicine									-	4.4
10. medicine/nutrition										-

Table 1: Human similarity ratings between documents.

Documents	1	2	3	4	5	6	7	8	9	10
1. crime/punishment	-	16	2	3	3	2	5	9	4	1
2. criminal law		-	0	2	1	3	1	8	4	0
3. disease study			-	2	9	6	4	1	4	2
4. drugs				-	2	5	1	2	1	2
5. evolution					-	2	5	0	2	5
6. human disease						-	2	2	2	5
7. human evolution							-	1	1	3
8. legal profession								-	8	0
9. history of medicine									-	0
10. medicine/nutrition										-

Table 2: Number of overlapping terms between the document word lists.

Word Overlaps

It is possible to measure the similarity between documents by the number of content-words they have in common, as reported in Table 2. The correlation between human ratings and word overlap measure is $r = 0.60$, $p < 0.01$, $df = 44$.

The simple word-match procedure did not take into account the rank positions of the overlapping words in the word lists. We weighted the overlaps with a score reflecting their relative frequency in the documents. Specifically, we took the ratio of the word frequencies with the total of the frequencies for the 40 words. When terms matched from both lists, the ratios for each of them were summed and the total across all matches was used as a measure of similarity between the documents. These scores improved the correlation with the ratings to $r = 0.69$, $p < 0.01$, $df = 44$. We address the application of these word-overlaps for collection scope after considering an alternative measure of relatedness.

Semantic Similarity and Relatedness Metrics

Although it is also computationally much more expensive, measures of semantic relatedness between words in a document may be a more sensitive method for determining the relatedness of documents than simple term-overlap counts. The semantic content of a document is expressed by the concepts it includes. However, scaling the similarity between concepts is not easy because of the homonymy and polyonymy of words.

Lexical Similarity and Relatedness

Miller and Charles [10] take semantic similarity as a dependent variable, considering semantic similarity as a function of the context in which words are used. "Context" has two senses. The narrow sense includes only linguistic context – the collection of words around the word in question while the broad sense includes both the linguistic context and non-linguistic information which indicates the communicative intentions of the speaker/writer ([10], p.4). "The contextual representation of a word is knowledge of how that word is used and is assumed to include syntactic, semantic, pragmatic, and stylistic conditions governing the use of that word" ([10], p.5). Although the similarity of the contextual representations of two words contributes to the semantic similarity of these words ([10], p.9), we are not going to discuss non-linguistic context due to its complexity, but simply address word similarity under linguistic context.

A distinction may be drawn between semantic similarity and semantic relatedness. Semantic similarity implies that there is an overlap in the attributes of the concepts represented by the words while semantic relatedness suggests that the concepts are similar but the attributes do not directly overlap. Semantic similarity represents a special case of semantic relatedness [13]: for example, cars and gasoline are closely related but not similar, while cars and bicycles are more similar but may not be closely related.

A complete set of human ratings of word similarity would be impractical. Thus, various computerized word (or concept) similarity measures have been explored. Here, we explore a new measure based on word co-occurrence statistics generated by Web search engines and several variant measures. Although this measure does not compete with existing measures in correlation with human judgments, these existing measures have their own shortcomings which are addressed below.

Similarity Metrics Based on Taxonomy

The evaluation of semantic relatedness using network representations has been an active research area. Rada et al. [12] suggest that evaluating similarity in semantic networks can be thought of as involving just taxonomic links, excluding other link types. In this vein, [9, 13, 20], and others have explored the taxonomic approach. Given an IS-A taxonomy (e.g., WordNet [4]), a natural and simple way to measure semantic similarity in a taxonomy is to measure the distance between the nodes corresponding to the words being compared – the shorter the path from one node to another, the more similar they are; given multiple paths, the length of the shortest one is taken [13]. The edge-counting method, though unstable, is conceptually very intuitive – just counting the edges between the nodes in a taxonomy as their distance. For example, in Figure 1, the distance between *dime* and *nickel* in a sub-taxonomy in WordNet is 2; the distance between *dime* and *credit* is 6. "By associating probabilities with concepts in the taxonomy, it is possible to capture the same idea as edge-counting, but avoiding the unreliability of edge distances" ([13], p.97). The similarity of two concepts is considered as the "information" shared by them; this is obtained from the information content of the concepts that subsume them in the taxonomy.

The word-similarity measures based on IS-A taxonomies have inherent shortcomings for establishing the similarity of documents. This is largely because of the distinction between "similarity" and "relatedness". Words that are not included in the taxonomy are not able to be compared. Moreover, the similarity of words in separate sub-taxonomies is always zero. For instance, WordNet has four separate taxonomies: for nouns, verbs, adjectives and adverbs. Thus, the similarity of words from different word classes (e.g., *beautiful* and *flower*) are not able to be computed. Both noun-similarity measures and verb-similarity measures have been proposed [14].

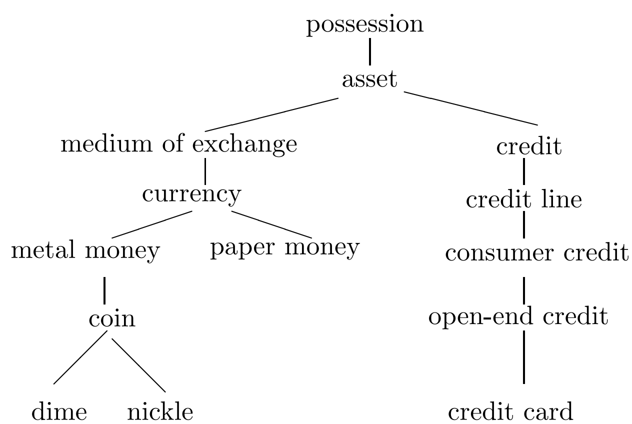

Figure 1: Fragment of a WordNet taxonomy. (Source: WordNet 1.6)

Relatedness and Co-occurrence

In addition to many similarity measures based on taxonomy, there are methods based on Latent Semantic Analysis. Landauer and Dumais [7] used Latent Semantic Indexing (LSI), a high-dimensional linear associative model, to generate a representation that captures the similarity of words and text passages. They assume that two words that appear in the similar document contexts should tend to be proximal in semantic space; then the relative similarity[1] of any pair of words can be analyzed by a statistical technique called singular value decomposition (SVD) which allows words and contexts to be represented as points or vectors in a high dimensional space.

The LSI model is always computable, but its shortcomings are that (1) it needs to be trained by a large corpus and perhaps to adjust the number of dimensions on the basis of trial and error adaptively; (2) whenever new text samples (or documents) are added into the corpus (or document collection) the space needs to be re-built.

Mining the Web

We explored a simple word-relatedness measure based on word co-occurrence statistics generated by Web search engines. We made three assumptions: The first assumption is that the Web itself approximates a complete, broad word space and real distribution of the words. That is, essentially all English words are used on the Web and their usage on the Web is the same as their usage averaged across other linguistic domains. Steinberg [17] helps to make this assumption. "A year and a half ago (in early 1995), the content of the Web was

[1] Following the distinction between similarity and relatedness described above, we call this "relatedness".

heavily tilted toward a few niches: there was a lot about Unix and UFOs, not much about real estate or poetry. But today (May 1996) the breadth of the Web comes close to covering all major subjects. Indeed, at its current growth rate, the Web will contain more words than the giant Lexis-Nexis database by this summer (1996), and more than today's Library of Congress by the end of 1998" [17].

The second assumption is that the Web pages indexed by the large search engines approximate a complete word space (i.e., all the English words are indexed) and a real distribution of the words. Sometimes when we query search engines, they may generate some "not found" pages or duplicate pages. Lawrence and Giles [8] compared six Web search engines during December 1997, and found the percentage of invalid links were less than 5%. Thus, the third assumption is that the disadvantageous impact any missing pages on the calculation of the query word's frequency is tiny.

Relatedness Metric

In search engines, "pages found" or "results found" provides a binary indication of whether the word is present or not. A relatedness measure can compare the number of pages returned to single-word queries with the number of pages returned for Boolean ANDs of word pairs (Equation 1.[2])

$$Relatedness_{Asymmetric} = \frac{(word1 \cap word2)_{count}}{word1_{count}} \quad (1)$$

It is more convenient to deal with symmetric relatedness measures. So, we define A as the number of hits for word1, B for word2, and C for word1 AND word2. The relatedness between word1 and word2 can be easily calculated as $\frac{C}{A}$ or $\frac{C}{B}$ or other variant expressions. To simplify the calculation, we calculate symmetric relatedness by taking the mean (see Equation 2). A joint entropy measure can also be developed as Equation 3.

$$Relatedness_{Mean} = \frac{\frac{C}{A} + \frac{C}{B}}{2} \quad (2)$$

$$Relatedness_{Joint\ Entropy} = -(\frac{C}{A}Log\frac{C}{A}) - (\frac{C}{B}Log\frac{C}{B}) \quad (3)$$

Validating the Co-occurrence Measures

An evaluation is conducted using the correlation against a benchmark set of human similarity judgment. "Semantic similarity is easily estimated by asking people to rate pairs of words with respect to their likeness of meaning" [10]. In Miller and Charles's [10] experiment, 38 university students (all native speakers of English) were asked to judge a subset of 30 noun pairs (see Table 4) from the original list of 65 studied by Rubenstein and Goodenough [15]. The study showed high inter-observer reliability with r=0.97, *df*=37 [10]. Resnik [13] repeated Miller and Charles's test, giving 10 graduate students the same 30 noun pairs. The result indicated high replicability, $r = 0.96$.

The similarity measures based on IS-A taxonomy were evaluated against human similarity judgments. As Table 3 shows, almost all the measures perform better than the traditional edge-counting measure [9, 13]. There are two pairs of words whose similarity cannot be calculated because the word "woodland" is not included in WordNet [13].

Relatedness Metric	Correlation
Resnik's Info Content	0.79
Sim Edge	0.67
Resnik's SimP	0.67
Wu & Palmer	0.80
Lin	0.83

Table 3: Correlations of several taxonomic-similarity measures with human similarity ratings [10].

For each of the 30 pairs of terms, we manually submitted the individual terms to the Advanced Search Interface of AltaVista with language set to English. We collected the frequency count for each term and applied that to Equations 2 and 3.

Results and Implications

Table 4 summarizes the results. The correlations for both of the new measures are significant, $p < 0.01$, $df = 29$ ([16], Table A11). While our measures appear to be robust and easy to calculate, they do not compete with existing measures based on IS-A taxonomy in correlation with human judgments (Table 3).

We tested Landauer and Dumais's assumption that first-order word co-occurrence statistics captures psychological word similarity to some degree. Some similar words do not co-occur frequently. Word co-occurrence statistics also capture word relatedness which helps to reflect real-word usage in language although it hurts its correlation with human similarity judgments.

While this word-relatedness statistic is promising, it has limitations. The first is that it depends on the search engines over which the research community has little control. The way the documents are selected and the stability of the collections are out of our control. While

[2] We found that the relatedness metric has been applied independently to Web word frequencies by [18]. However, we validate this measure, extend it, and apply it to a different domain.

Word Pair		Ratings	Rel_{Mean}	$Rel_{JointEntropy}$
car	automobile	3.92	0.293	0.238
gem	jewel	3.84	0.052	0.132
journey	voyage	3.84	0.092	0.176
boy	lad	3.76	0.161	0.187
coast	shore	3.70	0.153	0.237
asylum	madhouse	3.61	0.044	0.099
magician	wizard	3.50	0.036	0.093
midday	noon	3.42	0.046	0.108
furnace	stove	3.41	0.025	0.081
food	fruit	3.08	0.268	0.243
bird	cock	3.05	0.065	0.124
bird	crane	2.97	0.048	0.117
tool	implement	2.95	0.198	0.215
brother	monk	2.82	0.074	0.147
crane	implement	1.68	0.002	0.013
lad	brother	1.66	0.081	0.145
journey	car	1.16	0.088	0.168
monk	oracle	1.10	0.009	0.035
cemetery	woodland	0.95	0.010	0.038
food	rooster	0.89	0.037	0.086
coast	hill	0.87	0.002	0.008
forest	graveyard	0.84	0.024	0.068
shore	woodland	0.63	0.044	0.115
monk	slave	0.55	0.016	0.058
coast	forest	0.42	0.124	0.224
lad	wizard	0.42	0.009	0.034
chord	smile	0.13	0.025	0.076
glass	magician	0.11	0.023	0.067
rooster	voyage	0.08	0.004	0.017
noon	string	0.08	0.022	0.071
Correlation with Ratings			0.52	0.57

Table 4: Correlation of the Web Co-occurrence Measures with Human Ratings from [10].

the Web generally reflects the distribution of the words that people use, this is not uniformly true. For instance, the word "page" is very common because of its special meaning for the Web. Moreover, apparently some products and discussions are disproportionately represented.

Our semantic relatedness measure, defined as a statistical word-similarity measure in actual language usage, is independent of the size of the document collection. Because there is no perfect way of computing word similarity, our method is useful, at least to measure the scope of document collections. However, those previous measures based on the IS-A taxonomy work only under certain environments; for example, words in comparison must be included by the taxonomy, and in WordNet only those words in the same word class can be computed for their similarities. Our measures, however, are always computable.

In order to demonstrate that these results were general across search engines, we compared scores obtained by AltaVista with those from Google. For the Miller and Charles data [10], the correlation of the joint entropy measures for the two search engines was very high, $r = 0.94$, $p < 0.01$, $df = 29$.

Using Semantic Relatedness To Measure Distances Between Documents

We used the semantic distances to calculate pairs of document distances. Specifically, we used the 40 words selected from each document in the *Encyclopedia Britannica* test collection to calculate all 1600 pair-wise word distances for the two documents. Because there were 55 pairs of documents, approximately 88000 semantic distances were calculated. Naturally, obtaining these values required extended interaction with the search engines. We developed programs for automatically collecting, checking, and analyzing the data. We were also careful to complete the runs at off-peak hours and after notification of the search engine companies.

Table 5 shows the document distances obtained with the co-occurrence statistic. The correlation between human ratings and inter-document similarity measure is $r = 0.46$, $p < 0.01$, $df = 44$. Note that the diagonals were not included in the calculation; they represented within-document similarity.

Comparing Overlap and Semantic Similarity Measures for Scope

The simplest measure of the "scope" of the collection would be the greatest distance between any pair of documents. For the word-overlap measure (Table 2), this is the difference between 16 (for "crime" with "criminal law") and 0 (for several document pairs). Hence, the scope might be said to be 16.

A second measure is affected by the distribution of documents within the topics. It would be the mean value of the distances between the documents. To obtain distances for the data in Table 2 we subtract the overlap (a similarity score) from the largest possible overlap (40). Across all documents, the mean of the distance scores is 36.82. A large distance mean score corresponds to a large scope. We composed a small pseudo-collection composed of the three related encyclopedia articles on "crime", "criminal law", and "legal profession", and obtained a scope statistic of 29.00. We might express these values as a ratio of the number of terms which would give a scope of 0.92 for the entire set of documents and 0.73 for the "legal collection" of documents. Similarly, we could calculate values for the subset of documents dealing with disease.

Validation with Yahoo Science Categories and Sub-Categories

To test our scope metric with natural data, we examined the scope of categories in the Yahoo classification scheme. A category such as "Biology" should have a larger scope than one of its sub-categories such as "Marine Biology". We conducted a systematic test of the hypothesis that higher-level categories would have a greater scope than lower-level categories by considering the categories under the "Science" heading in Yahoo. A program was developed to crawl the Yahoo "Science" tree and extract all Web pages appearing under it. Not all of the categories and sub-categories were employed. In particular, categories that were links to other parts of the classification hierarchy were dropped, as were categories that were primarily disjoint lists only loosely about the category topic (e.g., "employment"). These Web pages were stripped of HTML and Javascript tags and then processed into word lists using the same procedure described above. Short documents (those with less than 100 words) were dropped. After processing, some of the categories did not have a sufficient number of sub-categories or documents for an adequate test of the scope metrics. We retained only those categories in which there was at least one sub-category with at least 10 documents. The result was a set of 16 categories; these categories had up to 15 sub-categories and up to 722 documents. For example, under Agriculture, we used the sub-categories: Animal Science, Aquaculture, Biotechnology, Crops and Soil, and Factory Farming Issues. In order to eliminate the possible effect of collection size on collection scope, we randomly sampled 10 documents from each category or sub-category as its test collection. For each of the 10 documents, we took the 40 words with the highest *tf* as its surrogate.

Scope tests were conducted for each category/sub-category pair. This means that some of the categories had several scope comparisons while other categories had only one test. Overall, 74 pairs in the 16 categories were tested. The word overlap measure was used to compute

Documents	1	2	3	4	5	6	7	8	9	10
1. crime/punishment	0.168	0.180	0.145	0.126	0.128	0.147	0.139	0.166	0.165	0.094
2. criminal law		0.189	0.152	0.132	0.132	0.157	0.143	0.178	0.176	0.129
3. disease study			0.163	0.148	0.148	0.170	0.146	0.138	0.170	0.143
4. drug				0.154	0.133	0.160	0.130	0.119	0.152	0.130
5. evolution					0.146	0.137	0.140	0.111	0.126	0.115
6. human disease						0.172	0.139	0.134	0.159	0.136
7. human evolution							0.149	0.122	0.133	0.113
8. legal profession								0.162	0.149	0.102
9. history of medicine									0.191	0.122
10. medicine/nutrition										0.142

Table 5: Inter-document relatedness based on the Joint Entropy Measure.

the similarity of any document pair.

The result overwhelmingly confirmed the prediction that category scopes would be larger than sub-category scopes. This simultaneously validated our scope metric. Specifically, the ratio of sub-category scopes that were smaller than the categories that contained them were calculated for each category. A paired-sample t-test on the 74 pairs showed that the category scopes were significantly different from the sub-category scopes, $t = 8.60, p < 0.01, df = 73$.

Detailed analysis suggested that several of the sub-categories whose scope exceeded the category scope were faceted. That is, they included a second dimension, such as geography, which led them to have a broad scope. We believe that future work may demonstrate that the scope measure could be useful in detecting such faceted document sets.

Discussion

Scope

We selected measures of scope that do not depend on the properties of the collection from which they are drawn. LSI measure might work. Our current scope measure is sensitive to the density of coverage of topics within the collection. Given that we have what appears to be reasonable measures of scope, we can evaluate it for a wider variety of collections. Indeed, we might also investigate how a collection's scope changes across time.

We explored the exact word match measure as a simple way to obtain the scope of collections. We also explored word-relatedness measures based on word co-occurrence statistics generated by Web search engines and its application in measuring the scope of collections. Both techniques work well, but the simple word-match measure gives a higher correlation in our experiments.

Related Measures

In addition to "scope" we have been exploring other related measures such as generality and coverage using text data mining. By "generality" we mean that a document "addresses general things or concepts." In [1], we explored predictors for the generality of six encyclopedia texts and had human subjects rank-order the generality of the texts. The generality of a text was computed as the relatedness of the words in the text with a collection of reference words using a joint entropy measure taking word co-occurrence statistics in Google as input. We found a statistically significant relationship between the human ratings of text generality and our automatic measure.

Acknowledgments

We thank AltaVista, Encyclopedia Britannica, Google, and Yahoo for the use of their resources. We also thank the reviewers for useful suggestions.

References

1. Allen, R. B., and Wu, Y. Generality of texts. In *International Conference on Asian Digital Libraries* (Dec. 2002), pp. 111–116.

2. Atkins, D. E. The University of Michigan Digital Library Project: The testbed. *D-Lib Magazine* (July/August 1996). http://-www.dlib.org/.

3. Bowman, C. M., Danzig, P. B., Hardy, D. R., Manber, W., and Schwartz, M. The Harvest information discovery and access system. In *World-Wide Web Conference* (Oct. 1994), pp. 762–771.

4. Fellbaum, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

5. Gravano, L., Garcia-Molina, H., and Tomasic, A. GlOSS: Text-source discovery over the Internet. *ACM Transactions on Information Systems 17* (1999), 229–264.

6. Lagoze, C., and de Sompel, H. V. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings Joint*

ACM/IEEE Digital Libraries Conference (2001), pp. 54–62.

7. Landauer, T. K., and Dumais, S. T. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*, 2 (1997), 211–240.

8. Lawrence, S., and Giles, C. L. Searching the World Wide Web. *Science*, 280 (1998), 98–100.

9. Lin, D. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML-98)* (Madison, Wisconsin). http://www.cs.ualberta.ca/~lindek/papers.htm.

10. Miller, G. A., and Charles, W. G. Contextual correlates of semantic similarity. *Language and Cognitive Process 6*, 1 (1991), 1–28.

11. Phanouriou, C., Kipp, N., Sornil, O., Mather, P., and Fox, E. A. A digital library for authors: Recent progress of the Networked Digital Library of Theses and Dissertations. In *Proceedings ACM Digital Libraries Conference* (Aug. 1999: Berkeley).

12. Rada, R., Mili, H., Bicknell, E., and Blettner, M. Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics 19*, 1 (1989), 17–30.

13. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligent Research 11* (1999), 95–130. http://www.umiacs.umd.edu/ resnik/pubs.html.

14. Resnik, P., and Diab, M. Measuring verb similarity. In *Proceedings Cognitive Science Society (CogSci)* (Aug. 2000), pp. 399–404.

15. Rubenstein, H., and Goodenough, J. B. Contextual correlates of synonymy. *Computational Linguistics*, 8 (1965), 627–633.

16. Snedcor, G. W., and Cochran, W. G. *Statistical Methods*. Iowa State University, Ames, IA, 1967.

17. Steinberg, S. Seek and ye shall find maybe. *WIRED Archive 4.05*, 5 (1996).

18. Turney, P. D. Mining the Web for synonyms PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning* (Sep. 2001), pp. 491–502.

19. White, H. Computer techniques for studying coverage, overlap, and gaps in collections. *Journal of Academic Librarianship 12* (1987), 365–371.

20. Wu, Z., and Palmer, M. Verb semantics and lexical selection. In *Association for Computational Linguistics* (Las Cruces, NM), pp. 133–138.