Gabel, J., & Smiraglia, R.P. (2009). Visualizing similarity in subject term co-assignment. 20th Annual ASIS SIG/CR Workshop: Bridging Worlds, Connecting People: Classification Transcending Boundaries doi: 10.7152/acro.v20i1.12886

Jeff Gabel
Brooklyn Law School Library
250 Joralemon Street, Brooklyn, New York, 11201
jgabel2617@aol.com
Richard P. Smiraglia
School of Information Studies
University of Wisconsin-Milwaukee, Bolton Hall 5th Floor 3210 N. Maryland Ave. Milwaukee, WI 53211, smiragli@uwm.edu

Visualizing Similarity in Subject Term Co-Assignment

ABSTRACT: The purpose of this research is to improve retrieval performance in systems that use assigned subject descriptors, such as library subject headings. We are looking for wider semantic boundaries surrounding summary headings assigned to documents by providing a means of identifying clustered headings that fall within the indexer's collective common perceptions of relevance. We are here experimenting with two techniques that can help increase both precision and recall. In earlier research citation-chasing was employed to yield a fuller retrieval set than might have been found using subject headings alone. In the present study we are employing multi- dimensional scaling to determine the best fit among works to which subject descriptors have been co- assigned. A term co-occurrence matrix compiled from 19 *LCSH* subject headings assigned to works in the field of "language origin" is used to generate an MDS map of the semantic space. Two clusters emerge: language and languages, and evolution biology, sometimes termed evolingo. Results allow us to visualize how differing perceptions of indexers affect the semantic space surrounding assigned terms. In both cases—citation-chasing and term co-occurrence—and especially when combining the two techniques acting as thresholds for each other, it is possible to overcome the inverse relation between precision and recall.

D-LIST TERMS: citation analysis, indexing, knowledge organization, linguistics

1.0 Introduction

The purpose of this research is to improve retrieval performance in systems that use assigned subject descriptors, such as library subject headings. In this study we are crossing methodological boundaries to reveal the power of perception in term assignment. We begin with the premise that indexers assign headings based on their perception of relevance, and that when two or more headings are assigned all are perceived as relevant. We are looking for wider semantic boundaries surrounding summary headings assigned to documents by providing a means of identifying clustered headings that fall within the indexers' collective common perceptions of relevance.

This research aligns nicely with Weiner (2005) who compared vocabularies generated by subject specialists with those generated by text mining systems. MDS was used to demonstrate convergence of terminology between the two domains. Similarly, Herrero-Solano et al (2006) showed the power of using MDS for generating bibliographic map displays with increased precision in online catalogs. This

ISSN: 2324-9773

work is similar to McCain's approach (2009) to providing a citation-image as context through the use of subject descriptors, and is directly related to studies she describes in that paper (p. 1302), in which term co-occurrence is constrained by a key phrase. Finally, Zhang and Wolfram et al. (2008) found substantial differences between colloquial search vocabulary and formal medical vocabulary; MDS was used to demonstrate term co-occurrences; the promising result relevant to the present study is the breadth of co-occurring neighboring terms.

We are here experimenting with two techniques that can help increase both precision and recall in such systems. In earlier research (Gabel, 2006a and 2006b) citation-chasing was employed to yield a fuller retrieval set than might have been found using subject headings alone. In the present study we are employing multi- dimensional scaling to determine the best fit among works to which subject descriptors have been co- assigned. Specifically, we employ a technique for visualizing how differing perceptions on the part of catalogers affect the semantic space surrounding assigned terms. In both cases—citation-chasing and term co-occurrence—and especially when combining the two techniques acting as thresholds for each other, it is possible to overcome the inverse relation between precision and recall.

2.0 Citation-chasing

In the original study (Gabel, 2006), one subject heading was selected and all titles were retrieved. This was:

> Language and Languages—Origin

We retrieved all titles to which the heading had been assigned from the local online catalog, resulting in a set of 13 monographs after filtering for date and location. All works cited in these monographs were located in OCLC; all *LCSH* on their bibliographic records were recorded. This yielded 2525 subject headings, of which 745 were used more than once. The results were tiered into 4 groups in a Bradford-like distribution in which the numbers of total citations were roughly equal among the tiers, while the numbers of subject headings in each tier increased at a rate between one-third and one-fifth as the frequencies decreased. This straight frequency scale was combined with a scale for the number of the 13 citing sources that produced each subject heading, creating an additional measurement of precision, in this case through restriction. The top citation-frequency tier included 35 subject headings, which were used in 10 to 13 of the original sources. These are the source headings used in the present study.

The original study ended here, with frequency analysis of the 35 subjects. The suggestion was implicit, that if a search engine were programmed to use this technique, a larger and more informative retrieval set with essentially fuzzy boundaries would be the result. Relevance in such a retrieval set would be dependent on the assignment of headings to different works.

3.0 Term Co-Assignment

The present study uses the *LCSH* headings compiled through one round of citation-chasing to analyze the perceptions of the indexers who assigned the headings. MDS is used to plot the headings in 2 dimensions according to the perceived proximity of semantic relevance. The hypothesis would be that use of MDS to weed extraneous subject assignment from the plot would help to constitute a citation-chased retrieval set with boundaries that lead to increased precision. Headings in the source list that had low incidence of co-occurrence were removed, which resulted in a core co-occurrence list of 19 headings. Table 1 (below) shows the headings that were used in the present analysis.

LCSH	**Abbreviation in Figure 1**
Behavior evolution	Bee
Biolinguistics	Bling
Brain - Evolution	Bre
Cognition and culture	Cac
Evolution	E
Evolution (Biology)	Eb
Human beings - Origin	Hbo
Human evolution	He
Linguistics	L
Language acquisition	La
Language and languages	Lal
Language and languages – Origin	Lalo
Language and languages – Philosophy	Lalp

Neuropsychology	N
Natural selection	Ns
Psychology	P
Psychology, Comparative	Pc
Psycholinguistics	Pling
Sign language	Sl

Table 1. ***LCSH*** **derived by citation-chasing from "Language and languages—Origin"**

All bibliographic records from the citation-chasing study were consulted to locate all instances of co-occurrence among the terms in Table 1. A co-occurrence matrix was analyzed using SPSS and the MDS map that appears in Figure 1 below was produced. Stress is .05 and R^2 is .99. McCain (1990, 438) suggests goodness of fit is indicated by low stress and high R^2, which we have in this case. This tells us the plot fits the data well. We have linguistics at the left tending toward the upper quadrant, and evolution square in the middle on the right, which suggests the domain is predominated by the linguistics headings. After some manipulation the MDS plot appears in Figure 1.

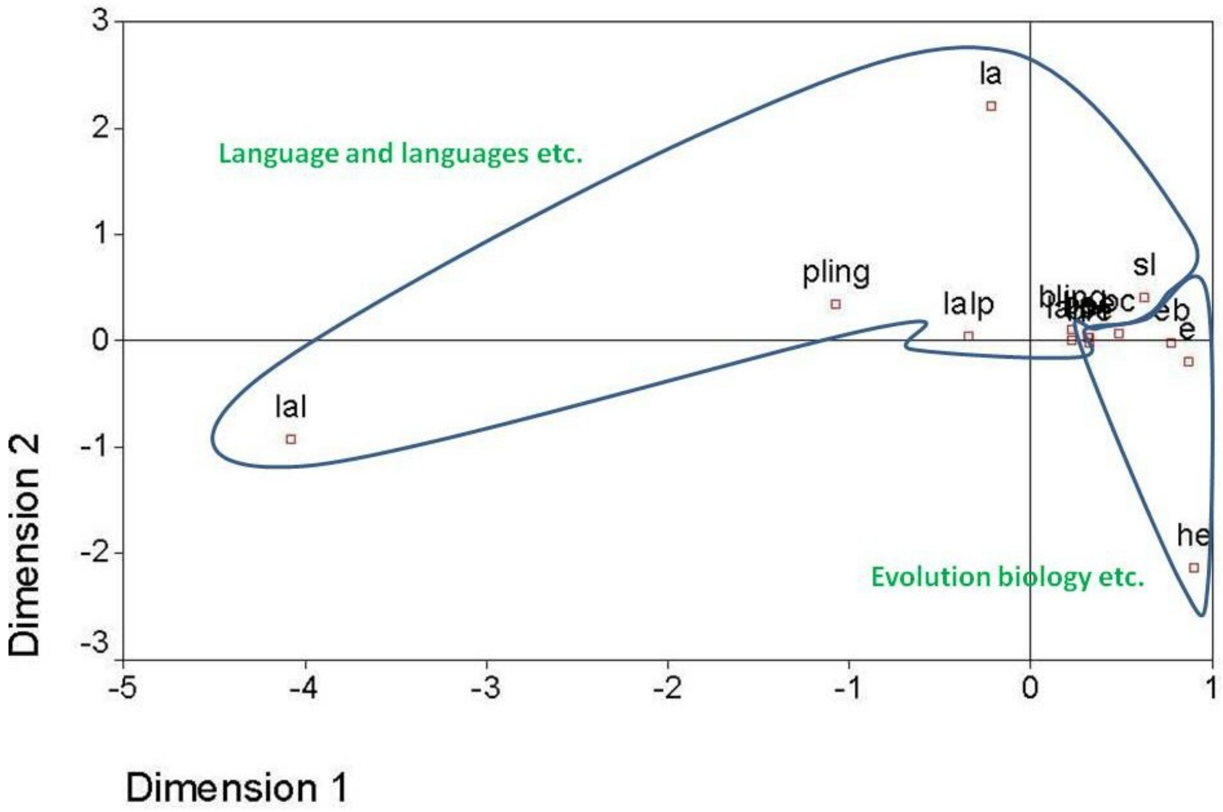

Figure 1. MDS plot of co-occurring *LCSH*

Two large clusters predominate: (1) Language-and-languages/Linguistics (including Biolinguistics, Language acquisition, Language and languages—Philosophy and Psycholinguistics) and Sign language; and (2) Human evolution/Human beings—Origin (including Brain evolution, Language and languages—Origin, Behavior Evolution, Cognition and Culture) and Evolution/Natural Selection (including Evolution (Biology), Neuropsychology, Psychology, and Psychology, Comparative). In semantic terms, we have two essential clusters: linguistics, and evolution. The two clusters are tightly associated around concepts of evolution and behavior, but the clusters do not overlap, and there is substantial distance from language to evolution (left to right). These linkages illustrate the perceived relevance among these headings based on the perceptions of the indexers who assigned them.

Looking more closely at the two clusters in semantic terms, one sees that each is composed of semantically representative subject headings, including both ‘pure’ headings and headings representing a combination of the namesake term with a term from another field (Biolinguistics and Brain—Evolution, for an example from each cluster). In addition, both clusters contain a secondary semantic representation composed of psychology-related headings, a trait which is more evident in the Evolution cluster than in the Linguistics one. Methodologically speaking (because it is the supplied topic of interest for the study), Language and languages--Origin belongs in both clusters. Semantically speaking, It also belongs in both clusters, but more so in Linguistics. However, it resides in the Evolution cluster, and is hooked to the semantically precise evolution topics (according to the Dendrogram, which is not reproduced here). In this 2-cluster group, Cognition and Culture alone is a semantic oddball among the 2 main cluster topics Evolution and Linguistics, and the notable supporting topic, Psychology.

4. Results of this research and evolingo

The *LCSH* “Language and languages—Origin” (or informally, the term “language origin”), semantically speaking, implies a broader concept than that of evolutionary linguistics (commonly referred to as evolingo in the research), one in which evolingo should represent but one among various sub-disciplines, or theories, within the discipline of language origin. However, in the current state of research in language origin, evolingo dominates the field to the extent that the concept often seems to be synonymous with language origin. Whatever the case, for the purposes of this study, the concepts evolingo or evolutionary linguistics are considered to be synonymous with that of language origin. Indeed, this is further demonstrated by the semantic properties of the clusters in this study’s results.

An *LCSH* terminology clarification: In the scientific literature, “Evolutionary linguistics” is often used to represent the study of the evolutionary development of language (hence the new ‘evolingo’). However, the *LCSH* “Evolutionary linguistics” in not considered a valid heading, but rather is a SEE reference directing users to “Historical linguistics”, a term which, semantically speaking, ought to be both more narrow than and parallel to its referent. Though the subject heading and concept “Historical linguistics” did not appear with sufficient frequency in Gabel (2006) to effect the outcome of the present study, it is important to clarify that the *LCSH* term “Evolutionary linguistics” is not semantically compatible with its counterpart in the research.

Evolingo has recently become a rapidly expanding and evolving field. Kenneally’s work (2007) is a well representative survey of the history and current state of evolingo research (Hoff, 2008). The position she assumes is probably the most widely-held view in the field. A confluence of new discoveries, re-examination of previous research and its corresponding assumptions, and new combinations of disciplines are replacing a long-standing situation where multidisciplinary work had been stifled due to strong dogmatism in the respective fields of linguistics and evolution, as well as considerable lack of physical evidence (Kenneally, 2007). Most of the science mentioned in her work has been accomplished in the last 30 years. Kenneally (2007) calls evolingo the most difficult problem in science today.

The survey lines up nicely with the results of this study. According to Kenneally, two scientific developments in the last half of the 20th century set the stage for evolingo. In short, the recognition of language as a property of the human mind led to the interdisciplinary study of psychology and linguistics. Later, the assertion that the mind is a result of natural selection as well as the body led to the interdisciplinary study of evolution and psychology. Together, these two multidisciplinary fields further prompted the question of how language evolved (Hoff 2008). Corroboratively, we have seen in the mapped results above that clusters were formed that are best represented by the terms Linguistics and Evolution. Further, psychology-related terms form a secondary topic across the clusters. Kenneally devotes 5 chapters to the processes of biological and cultural evolution. Once again, this lines up nicely with the results above. Cognition and culture is, semantically speaking, the least compatible subject heading in the results above. In conjunction with Kenneally’s “biological and cultural evolution”, the subject occupies an appropriate spot in the Evolution cluster.

5.0 Conclusion

Indexers assign headings based on their perception of relevance. When two headings are assigned to the same document there is evidence that both headings, then, are perceived as relevant. The co-occurrence technique in this study allows us to visualize the proximity of co-assigned terms within this commonly held perception of relevance. It is important to bear in mind that we are not looking at a hierarchical array of terms with semantic similarity to “Language origin” in *LCSH*. Rather, what we now can visualize is a map of the concept-space created by the perceptions, which are held in common by the assigning indexers. Additionally, within the clusters we have proximity maps of additional terms, which can be used to enrich retrieval. Because the semantic context is preserved, the enriched retrieval set does not come at the expense of precision. Clearly more research is required to hone these techniques. For instance, as we see in the case of evolingo, specific vocabularies may constrain perceptions of relevance. Similarly, we clearly are constrained by the specific semantic parameters of this single case. However, we see promise in the combination of citation-chasing and co-occurrence visualization, and hope to encourage further study of these techniques.

References

Gabel, Jeff. (2006a). Improving information retrieval of subjects through citation-analysis: a study. In Budin, Gerhard, Swertz, Christian and Mitgutsch, Konstantin eds., *Knowledge organization for a global learning society: Proceedings of the Ninth International ISKO Conference 4-7 July 2006 Vienna Austria*. Würzburg: Ergon-Verlag, pp. 19-26.

Gabel, Jeff. (2006b). Improving information retrieval of subjects through citation-analysis. *Knowledge organization* 33: 86-95.

Herrero-Solana, Victor, Moya-Anegon, Felix, Guerrero-Bote, Victor and Zapico-Alonso, Felipe. (2006). Graphical table of contents for library collections: the application of *Universal Decimal Classification* codes to subject maps. *Information technology and libraries* 43-47.

Hoff, Erika, (2008). Evolingo, or, evolutionary psychology meets linguistics: a review of Christine Kenneally, 'The First Word: The Search for the Origins of Language'*. Evolutionary psychology* 6: 213-16.

Kenneally, Christine. (2007). *The first word: The search for the origins of language*. New York: Penguin Group.

McCain, Katherine W. (2009). Using tricitation to dissect the citation image: Conrad Hal Waddington and the rise of evolutionary developmental biology. *Journal of the American Society for Information Science and Technology* 60:1301-19.

Weiner, John M. (2005). Differences in indexing term vocabularies and agreement with subject specialists. *Electronic journal of academic and special librarianship* 6n1/2.

Zhang, Jin and Wolfram, Dietmar, Wang, Peiling, Hong, Yi, and Gillis, Rick. (2008). Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science & Technology* 59: 1933-47.