Examining Communities in the Transdisciplinary Area of Cognitive Science: Automatic Classification for Examining Communities in the Web of Science Using Unsupervised Clustering Methods

Maxime Sainte-Marie
Université de Montréal
maxime.sainte-marie@umontreal.ca

Vincent Larivière
Université de Montréal
vincente.lariviere@umontreal.ca

Laura Ridenour
University of Wisconsin Milwaukee
ridenour@uwm.edu

ABSTRACT

We propose methodology for examining classification to identify and make explicit community perspectives that are neglected by traditional journal-subject classification in order to provide a more flexible and customizable classification system. Our method is based on keyword matches, and is applied to the broad transdisciplinary area of cognitive science. In the Web of Science (WoS), Scopus, and the National Science Foundation (NSF) classification, the classification of journals places each journal into a silo based on pre-determined categories deemed appropriate to demonstrate the relatedness of journals. Classification at the journal level does not necessarily represent the perspectives of a community, as a community in both membership and topical scope may transcend the bounds of a single journal classification. Our approach is novel because we examine topics within the transdisciplinary domain of cognitive science, and within that domain, we identify community perspectives on the conceptual contents as found in the titles of publications in the WoS.

Keywords

Automatic classification, community-based classification, unsupervised clustering methods, knowledge organization.

INTRODUCTION

Knowledge organization and classification research seek to identify and make explicit relationships between concepts within and between domains. We find it useful to frame our approach using Star's notion of the boundary object (Star & Griesemer, 1989), or, an entity of shared interest to multiple communities. In this case we find the boundary object as a conceptual entity naturally extends to the realm of Popper's third world (Popper, 1979), which is a useful way of framing the part of reality where ideas interact. Concepts as boundary objects in business have been examined by Langenohl (2008), and this this type of theoretical approach for examining classification of conceptual entities was investigated by Ridenour (2016) for the topic of network theory. Titles have been used to automatically detect facets in knowledge organization (KO) (Green, 2014), fitting in with an increasing trend in KO and other disciplines to incorporate automatic classification methods. We see this method as being able to identify and make explicit community-based perspectives through the identification of core concepts.

METHODOLOGY

In order to extract as most cognitive-relevant articles as possible, all WoS entries whose title attribute contains the substring 'cogni' were extracted. Regular expressions used in the extraction process were specifically designed to avoid substring matching based on the sole basis of recognition-related expressions (for example, based on words like 'recognition' or 'recognize', resulting in 105,226 articles).

Disciplines considered foundational to cognitive sciences were then identified on the basis of the NSF three-tiered field classification of journals (higher level: *Grand discipline*, medium level: *Discipline*, lower level: *Specialty*). In accordance with both prior surveys and the broad consensus in the relevant scientific community (Miller, 2003), six different disciplines were identified and used in this study: Anthropology, Computer Science, Linguistics, Neuroscience, Philosophy, and Psychology. Table 1 shows the different NSF field categories chosen for each partaking discipline.

Discipline	NSF Specialty	NSF Discipline	NSF Grand Discipline
Anthropology	*Anthropology and Archaeology*	*Social Sciences*	SSH
Computer Science	*Computers*	*Eng. & Tech.*	NSE
Linguistics	*Language & Linguistics*	*Humanities*	SSH
Neuroscience	*Neurology & Neurosurgery*	*Med. Science*	NSE
Philosophy	*Philosophy*	*Humanities*	SSH
Psychology	--	*Psychology*	SSH

Key: NSF = National Science Foundation, Eng. & Tech = Engineering and Technology, SSH = Social Sciences and Humanities, NSE = Natural Sciences and Engineering

Table 1. NSF Field Classification of Foundational Disciplines of Cognitive Science.

Of all articles extracted from the WoS dataset, only those pertaining to one of the above disciplines were kept. We created word-based disciplinary vector space models by removing stop words from article titles within each discipline, and then converting the remaining text into tf-idf vectors. Then, mean disciplinary vectors were created by averaging the number of occurrences by title of each word present in each disciplinary matrix. These 'stereotypical titles' for each discipline represent the 'center of mass' or centroid of their respective disciplinary matrix, that is, the arithmetic mean of each dimension (each distinct word) of all article vectors in the corresponding matrix. These centroids were then used for two different classification tasks. First, all articles pertaining to the six disciplines were reclassified using the classic k-means clustering algorithm, with the above-mentioned disciplinary vectors as initial centroids. Second, a new reclassification was attempted by computing distances between each article title vector and the six disciplinary vectors, and then reassigning each article to the discipline whose stereotypical vector was the closest. The second classification, here called 'nearest centroid procedure' can be seen as a simpler and shorter version of the first, as it is formally equivalent to an "uniterative" k-means algorithm. Finally, intra-class distance and inter-class distance of all three classifications (original, k-means, nearest centroid) were computed.

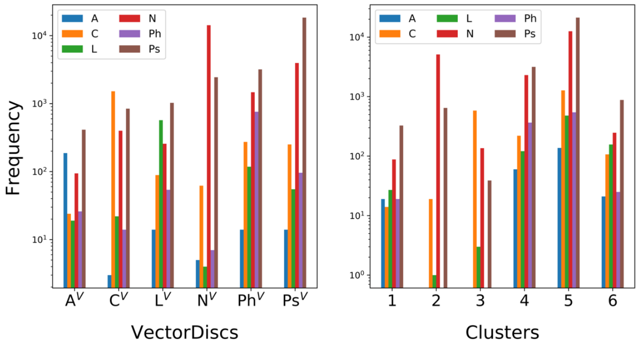
RESULTS AND DISCUSSION

Table 2 shows the intra- and inter-class distances for the original, nearest-centroid, and k-means classifications, along with their respective means.

	Original		Nearest centroid			K-Means	
Field	Inter	Intra	Inter	Intra	Cluster	Inter	Intra
Anthropology	.57	.99	.69	.97	1	.85	.92
Computer Science	.64	.98	.73	.98	2	.83	.93
Linguistics	.54	.99	.68	.98	3	.89	.90
Neurology	.61	.98	.72	.98	4	.84	.96
Philosophy	.57	.99	.73	.97	5	.71	.99
Psychology	.48	.99	.64	.99	6	.81	.95
Average	.57	.99	.69	.98	**Average**	.82	.94

Table 2. Intra- and Inter-distance Scores for all Top-level Classifications.

At first glance, both reclassification procedures attempted in this paper clearly outperform the original, NSF- and journal-based, classification: disciplinary matrices are more cohesive and further apart from each other than in the original classification. While the K-Means results are clearly optimal in terms of both cohesion (intra-distance) and distinctiveness (inter-distance), explaining what the K-Means clusters stand for might prove extremely difficult. As shown on the right-hand subplot of Figure 1, the K-Means reclassification procedure has shuffled articles to the point where the resulting clusters bear little resemblance to the original NSF disciplines. In a way, by blurring the original disciplinary landscape, the K-Means procedure solves a classification problem by creating a new one, that is, the problem of making sense of the new article clusters.

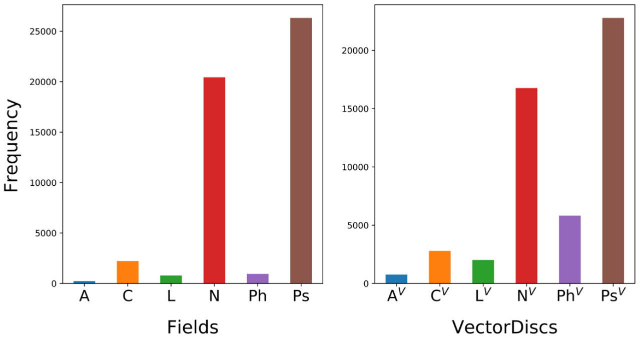


Key: A = Anthropology, C = Computer Science, L = Linguistics, N = Neuroscience, Ph = Philosophy, Ps = Psychology, A^V = Anthropology (vectorized), C^V = Computer Science (vectorized), L^V = Linguistics (vectorized), N^V = Neuroscience (vectorized), Ph^V = Philosophy (vectorized), Ps^V = Psychology (vectorized)

Figure 1. Frequency Distribution by Discipline of Vector and Clustered Classifications.

By contrast, the nearest centroid procedure does not take article shuffling as far as the K-means-based procedure: while refining the silhouette of each disciplinary matrix, this procedure generates a classification that is still reminiscent of the original one. Indeed, despite the fact that thousands of articles of each discipline are reassigned to other classes, disciplinary specializations still remain: while all new classes include articles from every original category, each one distinguishes itself by having more articles of one and only one given discipline than the other ones. For example, the class A^V has the most Anthropology articles, the class L^V has more Linguistics articles than the other ones, and so on. By contrast, cluster 5 of the K-Means classification has more articles from the original disciplines than any other cluster; this not only complicates the interpretation of that cluster, but also makes it hard to make sense of the other ones. In sum, while the nearest centroid classification significantly improves the original NSF Classification on various ground, the article reshuffling is not so drastic as to lose all connection with the imperfect, yet intuitive and understandable original disciplinary schema, as is the case with the K-means classification.

However, the main interest of the nearest-neighbor procedure is that it enables a smooth transition from a journal-based classification to an article-based one. At first glance, the disciplinary portrait offered by the original NSF classification of cognition-related articles is both highly specialized and diversified: on the one hand, with the exception of Psychology, which is a medium-level NSF discipline, all relevant fields pertain to the lower-level 'NSF Specialty' category; on the other hand, these fields cover a wide spectrum of the disciplinary landscape, from Social Sciences and Humanities to Medical Sciences and Engineering. However, since the NSF Field Classification is journal-based, not article based, and given that each journal is assigned one and only one field, the number of misclassified articles must be non-negligible: surely, not all articles included in journals classified in one given field actually pertain to that field, and not all articles pertaining to a given field are actually published in journals classified in that same field. In the case of cognitive science, these problems are further exacerbated by the fact that the most important and relevant journals of cognitive science are either classified as ‘Psychology’ or ‘Neuroscience’ journals by the NSF, which skews the distribution at the expense of the other disciplines, as in the left-hand subplot of Figure 2. Examples will be discussed in the workshop.



Key: A = Anthropology, C = Computer Science, L = Linguistics, N = Neuroscience, Ph = Philosophy, Ps = Psychology, A^V = Anthropology (vectorized), C^V = Computer Science (vectorized), L^V = Linguistics (vectorized), N^V = Neuroscience (vectorized), Ph^V = Philosophy (vectorized), Ps^V = Psychology (vectorized)

Figure 2. Frequency Distribution of Field and Vector Classifications.

By contrast, by reclassifying NSF-classified articles based on their individual attributes (in the present case, titles), the nearest centroid reclassification algorithm presented here allows for a more fine-grained and flexible partitioning. As shown on the right-hand plot of Figure 2, the disciplinary distribution of article is more homogeneous than the original one, as the article reshuffling done by the algorithm allows for an enhanced and arguably more realistic representation of cognitive disciplines other than Psychology or Neuroscience. In this sense, the nearest centroid algorithm presented allows for a seamless journal- to article-based field classification, one that not only results in a cleaner partition, but also maintains and even helps emphasize original disciplinary identities. In our view, these results alone are sufficient evidence of the usefulness and reliability of quantitative approaches to classification.

CONCLUSION

Community perspectives can be roughly derived from textual data, as is demonstrated by the clustering found for each sub-discipline in cognitive science. This method moves bibliographic database classification from the current paradigm of journal-based classification to an article-based classification that can be used to honor community perspectives. The method we have implemented would be useful for the examination of topics core to inchoate interdisciplines, marginalized, and other communities as they exist in a state of flux and lack a consistent and cohesive set of disciplinary traditions as their boundaries are fuzzily defined and topics key to each document and community do not reside in a neat classificatory space.

REFERENCES

Green, R. (2014). Facet Detection Using WorldCat and WordNet. In W. Babik (Ed.), *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects: Proceedings of the Thirteenth International ISKO Conference* (pp. 168–175). Krakow, Poland: Ergon-Verlag.

Langenohl, A. (2008). How to change other people’s institutions: Discursive entrepreneurship and the boundary object of competition/competitiveness in the German banking sector. *Economy and Society 37*, 68–93.

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences 7*, 141–144.

Popper, K. (1979). *Three worlds*. Ann Arbor, Michigan.

Ridenour, L. (2016). Boundary Objects: Measuring Gaps and Overlap Between Research Areas. *Knowledge Organization 43*, 44–55.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, “translations” and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science 19*, 387–420.