Joint Content-Context Analysis of Scientific Publications: Identifying Opportunities for Collaboration in Cognitive Science

Anonymous ACL submission

Abstract

This work studies publications in cognitive science and utilizes natural language processing and graph theoretical techniques to connect the analysis of the papers' content (abstracts) to the context (citation, journals). We apply hierarchical topic modeling on the abstracts and community detection algorithms on the citation network, and measure content-context discrepancy to find academic fields that study similar topics but do not cite each other or publish in the same venues. These results show a promising, systemic framework to identify opportunities for scientific collaboration in highly interdisciplinary fields such as cognitive science and machine learning.

1 Introduction

As scientific fields have grown larger and more specialized, researchers may be missing potentially-lucrative avenues of collaboration. For example, researchers may be pursuing similar paths in parallel while lacking a common language and literary academic foundation to connect their works. Uncovering such situations will enable more productive, coordinated research efforts, which is one of the principal goals of science of science.

Science of science, or metascience, is the branch of science that uses quantitative measurements and scientific techniques to understand the interactions between scientific agents with the aim to refine and improve scientific practices and progress (Fortunato et al., 2018). Yet currently, most metascience studies have focused on investigating either the content or context of research in relation to other publications without bridging the gap between them (Evans and Foster, 2011). In this paper, we investigate the field of cognitive science through the twin lenses of content and context; information is extracted from both 1) paper abstracts through natural language processing (NLP) and 2) the citation network via graph community detection techniques. We then propose a simple but effective criteria to determine which subdivisions within cognitive science are similar in content but not in context, and suggest what barriers may lie between them.

We focus on cognitive science, in part because it has been claimed that cognitive science has failed to achieve its intention of integrating the six disciplines of which it was to be comprised (psychology, linguistics, artificial intelligence, anthropology, philosophy and neuroscience) (Núñez et al., 2019). Hence, it will be revealing to discover which interdisciplinary connections are missing in the field and investigate how this gap could be filled. Beyond cognitive science, our approach and methods can provide a framework for the joint study of content and context in other interdisciplinary fields such as applied mathematics and machine learning.

2 Data Acquisition and Preprocessing

A total of 258,039 papers in the field "cognitive science" were obtained from the Microsoft Academic Graph (Sinha et al., 2015), where the field tags of a paper are identified from its text and sometimes citations, and the papers are also given probabilities of being "important" (Shen et al., 2018). In addition, each paper is assigned a unique ID and include metadata such as title, authors, journal and year of publication, abstract, and references.

First, we discard 58,039 papers with the lowest probabilities of being "important" because 1) $\sim 0\%$ of them have abstracts, 2) $\sim 0\%$ have references, 3) none are published in recent years, and 4) the probability is significantly lower than the rest. We then remove papers published prior to 1950 in order to limit the scope to the modern notion of cognitive science from the 1950s (Núñez et al., 2019).

Next, we keep only the papers that contain references, and whose abstracts are between 30 and 500 words long. We found that many exceedingly short abstracts are actually titles and publication information, while exceedingly long abstracts tend to contain extraneous text such as table of contents or the text of the entire first page of the paper. Finally, after removing all papers with duplicate abstracts, we have a dataset of 59,384 papers for analysis.

3 Methods

We introduce NLP and graph methods that were used to conduct content and context analyses on the publications dataset, as well as metrics used to quantify cluster similarities.

3.1 Content Analysis

Bag-of-Words Matrix Construction We first lemmatize the abstracts and remove numbers, punctuations, English stop words, and stop words specific to abstracts (e.g. "et al", "this paper"). We then construct the data matrix using the bag-of-words model and term frequency-inverse document frequency (tf-idf) weighting, including tri-grams and excluding words that appear in more than 80% or less than 0.05% of abstracts. This yields a word-by-abstract matrix $\mathbf{X}$ of size 9,106 $\times$ 59,384.

Non-Negative Matrix Factorization (NMF) We apply NMF (Lee and Seung, 1999) to detect topics and assign papers to topics. NMF approximates $\mathbf{X} \approx \mathbf{WH}$, where the dictionary matrix $\mathbf{W}$ and the coding matrix $\mathbf{H}$ are two low-rank non-negative matrices. The ith column of $\mathbf{W}$ gives the weights of the words in the ith topic, while the jth column of $\mathbf{H}$ gives the weights of the topics in the jth abstract. This allows us to represent a topic as a combination of words, and an abstract as a combination of topics. We describe each topic using its top three weighted words, and assign each paper to its most weighted topic.

Hierarchical NMF Let two rank-r matrices $\mathbf{W}$ and $\mathbf{H}$ be the output of performing NMF on $\mathbf{X}$. Once we assign abstracts to topics based on $\mathbf{H}$, we column-wise split $\mathbf{X}$ into r sub-matrices, $\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_1^{(r)}$, such that columns of $\mathbf{X}_1^{(i)}$ correspond to abstracts assigned to the ith topic. Then we perform NMF on each sub-matrix to obtain dictionary and coding matrices for the *sub*topics. This top-down approach (Grotheer et al., 2020) allows us to develop hierarchical topics.

3.2 Context Analysis

Citation Network Construction After assigning papers to nodes and citations between those papers to edges, our citation data yields a graph with 59,384 nodes and 191,871 directed edges. We then isolate the largest weakly-connected component, which leaves us with 41,465 nodes (69.8% of original papers) and 190,997 edges (99.5% of original citations). Symmetrizing our graph allows us to leverage more powerful and trusted algorithms for community detection, so we employ Degree-Discounted Symmetrization (Satuluri and Parthasarathy, 2011).

Modularity and Louvain's Algorithm Modularity is a measure of the quality of a graph partition or community scheme. It records the number of intra-community edges minus how many intra-community edges we would expect to see if the edges were placed at random while following the same degree distribution. We use Louvain's Algorithm (Blondel et al., 2008) to find a community scheme that maximizes the modularity, as the greedy algorithm can be fast, intuitive, and scale to large networks easily.

3.3 Content-Context Discrepancy

Let c_i be the ith largest community of publications in the citation network. We measure topic similarity $T(c_i, c_j)$ and journal similarity $J(c_i, c_j)$ as proxies for content and context similarity, respectively. Then, we calculate the discrepancy $\rho(c_i, c_j)$ and use these metrics to identify communities that are more similar in content than they are in context.

Recall that every paper in c_i is assigned to an NMF topic, and has its journal of publication known. Let $\mathbf{t}_i$ be the frequency distribution of the topics of the papers in c_i. Similarly, $\mathbf{p}_i$ is the frequency distribution of journals that the papers in c_i were published in. Normalize them by $\hat{\mathbf{t}}_i = \mathbf{t}_i / \|\mathbf{t}_i\|_2$, $\hat{\mathbf{p}}_i = \mathbf{p}_i / \|\mathbf{p}_i\|_2$, then define the similarity metrics as their dot product:

$$T(c_i, c_j) = \langle \hat{\mathbf{t}}_i,\ \hat{\mathbf{t}}_j \rangle,\ J(c_i, c_j) = \langle \hat{\mathbf{p}}_i,\ \hat{\mathbf{p}}_j \rangle. \quad (1)$$

Our proposed discrepancy index combines these two metrics by

$$\rho(c_i, c_j) = T(c_i, c_j) - J(c_i, c_j)/2, \quad (2)$$

so that topic similarity is considered more heavily.

4 Results and Discussion

We display topic modeling and community detection results on the publications dataset, and discuss how it may relate to missed opportunities for scientific collaboration in cognitive science.

4.1 Hierarchical Topics in Cognitive Science

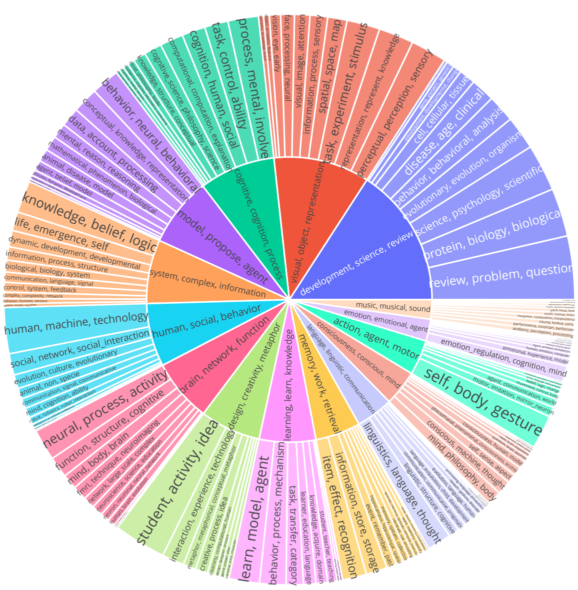

Figure 1: Hierarchical topics of cognitive science according to paper abstracts. Labels are the topics' keywords, and wedge size is proportional to number of papers in the topic.

Figure 1 shows the hierarchical topics extracted from abstracts. The inner circle contains 15 topics, and each topic is further split into 8 or 10 subtopics in the outer circle. Some keywords suggest connections to known fields of cognitive science:

- language, linguistic, communication→ linguistics
- human, social, behavior→ anthropology,
- consciousness, conscious, mind→ philosophy.

It is notable that neither "computer science" nor "psychology" seem to exist as keywords to a main topic even though they are claimed to dominate the field of cognitive science in (Núñez et al., 2019). A hypothesis is that as those fields have become so broad and popular, researchers avoid those terms and instead use specific subtopics or methods under the field to describe their work. Alternatively, these fields could be so prevalent and diffused within cognitive science that they would not appear as a distinct topic.

4.2 Content-Context Discrepancy Criteria

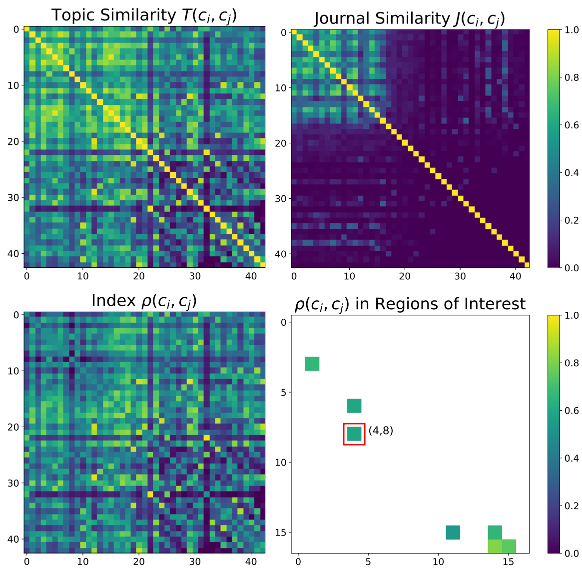

Figure 2: Heatmaps of metrics T, J and ρ. Axes are community indices i, j.

After uncovering 15 topics in the abstracts and 43 communities in the citation network, we examined and visualized in Figure 2 the metrics $T(c_i, c_j)$ (top left), $J(c_i, c_j)$ (top right), and $\rho(c_i, c_j)$ (bottom left). The color of each pixel represents the metric value for the pair of publication clusters. Note that $J(c_i, c_j)$ drops significantly at $i, j = 17$. The sample space of journal distribution in this dataset is large, but many communities are very small, often with merely tens of papers. This means the journal distribution vectors are necessarily sparse, leading to a flawed comparison between smaller communities. Therefore, we limit our analysis to the 17 largest communities and compare only those close to each other in size to minimize other size effects.

We use the following criteria to identify regions of interest, i.e. communities in cognitive science that may discuss similar themes but do not cite each

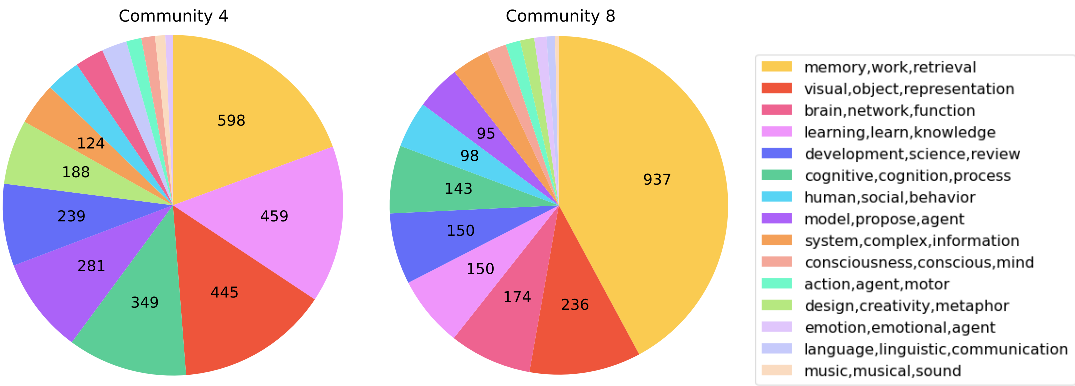

Figure 3: Topic distributions in communities 4 and 8. Wedge labels are numbers of papers in the topic. Legend shows keywords.

other or publish in the same venues:

- Similar topics: $T(c_i, c_j) > 0.75$,
- Dis-similar journals: $J(c_i, c_j) < 0.5$,
- High discrepancy: $\rho(c_i, c_j) > 0.5$,
- Similar size: $|i - j| \leq 5$,
- Large enough size: $i, j \leq 16$.

The bottom right of Figure 2 shows the 7 identified pairs, which we can then examine in greater detail.

4.3 Case Study on Communities 4 & 8

Communities 4 and 8 (boxed in red in Figure 2 bottom right) yielded $T(c_4, c_8) = 0.826$, $J(c_4, c_8) = 0.479$, and $\rho(c_4, c_8) = 0.586$. According to the pie charts in Figure 3, the two communities have a very similar topic composition—both are a mix of “memory” + “visual” + “learning”. At the same time, the fact that they are split into two graph communities indicates that they are not very connected in the citation network. In fact, there are approximately 15,000 intra-community edges in these two communities, and only 800 inter-community edges. Furthermore, we find very little overlap in the respective journal sets of these two communities. See below for their top 10 published-in journals:

Community 4	
Advances in Psychology	78
Memory & Cognition	66
Journal of Experimental Psychology	63
Applied Cognitive Psychology	61
Educational Psychologist	52
Educational Psychology Review	44
Psychology of Learning and Motivation	43
Journal of Educational Psychology	35
Psychonomic Bulletin & Review	34
Memory	32

Community 8	
Trends in Cognitive Sciences	84
Behavioral and Brain Sciences	50
BiorXiv	47
Frontiers in Human Neuroscience	35
Neuropsychologia	34
Journal of Cognitive Neuroscience	33
Neuron	30
Current Biology	30
Neuroscience & Biobehavioral Reviews	30
Memory	29

Community 4 is mostly published in (educational) psychology journals, whereas community 8 is associated with neuroscience journals. Clearly, there is a citational and academic disconnect between them, even though they share similar topic distributions. Initiating conversation between them could help further our understanding of complex subjects like memory, as it can provide a more holistic view of the theme, and even inspire fresh research questions and methods.

5 Conclusions and Future Work

We outlined an application of NLP on science of science— a method that connects the analysis of the content and context of scientific papers. We extracted topics from abstracts using hierarchical NMF, detected communities in the citation network, and analyzed their journal publication distributions. These approaches allowed us to find groups that are close in content but not in context, which indicate potential opportunities for collaboration.

In the future, we wish to add a temporal dimension to our analysis. For example, can we recognize changes in citation network and prominent topics over time? Can we detect shifts in rhetoric and composition? We plan to apply this framework to particularly entangled fields such as artificial intelligence and machine learning.

References

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

James Evans and Jacob Foster. 2011. Metaknowledge. *Science (New York, N.Y.)*, 331:721–5.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379):eaao0185.

Rachel Grotheer, Yihuan Huang, Pengyu Li, Elizaveta Rebrova, Deanna Needell, Longxiu Huang, Alona Kryshchenko, Xia Li, Kyung Ha, and Oleksandr Kryshchenko. 2020. Covid-19 literature topic-based search via hierarchical nmf. *arXiv preprint arXiv:2009.09074*.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Rafael Núñez, Michael Allen, Richard Gao, Carson Miller Rigoli, Josephine Relaford-Doyle, and Arturs Semenuks. 2019. What happened to cognitive science? *Nature Human Behaviour*, 3(8):782–791.

Venu Satuluri and Srinivasan Parthasarathy. 2011. Symmetrizations for clustering directed graphs. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 343–354.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.