Augmenting Scientific Creativity with Retrieval across Knowledge Domains

Hyeonsu B. Kang[1*]   **Sheshera Mysore**[2*]   **Kevin Huang**[3*]   **Haw-Shiuan Chang**[2]
Thorben Prein[3]   **Andrew McCallum**[2]   **Aniket Kittur**[1]   **Elsa Olivetti**[3]

[1]Human-Computer Interaction Institute, CMU, PA, USA
[2]Manning College of Information and Computer Sciences, UMASS AMHERST, MA, USA
[3]Department of Materials Science and Engineering, MIT, MA, USA
{hyeonsuk,nkittur}@cs.cmu.edu
{smysore, hschang, mccallum}@cs.umass.edu
{kjhuang, prein, elsao}@mit.edu

Abstract

Exposure to ideas in domains outside a scientist's own may benefit her in reformulating existing research problems in novel ways and discovering new application domains for existing solution ideas. While improved performance in scholarly search engines can help scientists efficiently identify relevant advances in domains they may already be familiar with, it may fall short of helping them explore diverse ideas *outside* such domains. In this paper we explore the design of systems aimed at augmenting the end-user ability in cross-domain exploration with flexible query specification. To this end, we develop an exploratory search system in which end-users can select a portion of text core to their interest from a paper abstract and retrieve papers that have a high similarity to the user-selected core aspect but differ in terms of domains. Furthermore, end-users can 'zoom in' to specific domain clusters to retrieve more papers from them and understand nuanced differences within the clusters. Our case studies with scientists uncover opportunities and design implications for systems aimed at facilitating cross-domain exploration and inspiration.

1 Introduction

Analogies have been a central mechanism for cross-boundary inspirations and innovation throughout the history of science and technology (Gentner et al., 1997; Oppenheimer, 1956). For example, analogical reasoning helped the Greek philosopher Chrysippus (c. 240 B.C.) model sound waves from observations of water waves, two domains that did not interact prior to this insight. However, as knowledge areas deepen in specialization, they interact less with each other (Swanson and Smalheiser, 1996; Chu and Evans, 2021) and analogical reasoning across domains becomes increasingly challenging. Yet, a systematic review of existing literature suggests that enabling connections that

jump between scientific domains may bring significant innovations and scientific progress (Rzhetsky et al., 2015). As such, developing a computational means to augmenting human intelligence with analogical creativity remains a holy-grail challenge in both cognitive science and AI research (Mitchell, 1993; Hesse, 1965; Kittur et al., 2019).

To this end, recent approaches for retrieving analogically similar ideas have demonstrated promising results (Kang et al., 2022a; Hope et al., 2017, 2021; Chan et al., 2018). The central idea of these approaches focuses on retrieving analogs similar in one aspect of a pre-determined schema (e.g., 'problem solved') while differing on the other (e.g., 'method used'), exposure to which has shown to unlock scientists' creativity to break out of fixation and mechanistic conventions (Kang et al., 2022a).

While promising, there are remaining technical and design challenges that prohibit realization of systems that support exploration across boundaries of knowledge domains. The first challenge is developing mechanisms that allow end-users to explore diverse domains and find ones they want to jump to find analogous retrievals. Prior work in cognitive psychology research on human creativity (Chan et al., 2015; Gonçalves et al., 2013) has noted a sweet spot of distance for retrieval, with effective results being neither too different from the query, which may cause early rejection from searchers, nor too similar, which may be insufficient for breaking out of conventions and fixation (Jansson and Smith, 1991). This notion of effective distance is a departure from the commonly used notions of relevance in the literature of IR and NLP that rely on topical overlap (Borlund, 2003) and development of methods often optimized for maximal content-similarity (Carterette, 2011). A notable exception is the prior research in search diversification (see Santos et al. (2015)), which contributes mechanisms for *optimizing* the novelty and coverage of search results given ambiguous queries. However,

---

*Equal contribution.

they often lack the controllability for end [...] who want to specify the kinds of novelty th [...] interested in and the scope of coverage.

The second challenge is the lack of labele [...] for scaling a schematic representation of [...] ments. Barriers to domain knowledge ma [...] curate labeling of research papers using [...] determined schema across diverse domair [...] hibitively costly. While Chan et al. (2018) c [...] strated a potential pipeline for crowdsourci [...] labels with non-expert crowdworkers, a subs [...] paper in which a supervised model was t [...] based on the crowd-labeled data showed [...] tions on the model accuracy and downstre [...] trieval usefulness (Kang et al., 2022a). Further- more, emerging scientific domains present an addi- tional challenge with updating data.

Finally, studying how cross-domain explorations occur in authentic use scenarios requires a) develop- ment of an interactive search system with a usable performance and b) evaluating it with scientists seeking inspirations on real-world research prob- lems. Insights from the study also contribute to the growing literature that studies open questions around how informational and exploration needs in creative work change over time and during the exploration (Li et al., 2022).

In this work we develop an interactive prototype system which we use to probe these challenges and uncover design implications for future analog- ical search systems aimed at supporting retrieval across knowledge domains[1]. Our system leverages the recently introduced *faceted Query-by-Example (QbE)* paradigm to increase the controllability of search novelty and scope, by allowing searchers to query using a paper abstract and a sentence[2] in the abstract that indicates personally interesting core aspects of the paper (Mysore et al., 2021). To find matching abstracts based on the query aspect, our system first indexes documents using each of its sentences and retrieves similar sentences using a strong sentence encoding model (sbert.net, 2021) at runtime. To go beyond the high-demand for labeled data in a challenging domain, our approach relaxes the mechanism for retrieval from relying on a pre- determined schematic representation of abstracts to using a readily available, intuitive sentence-level representation. To support an increased diversity
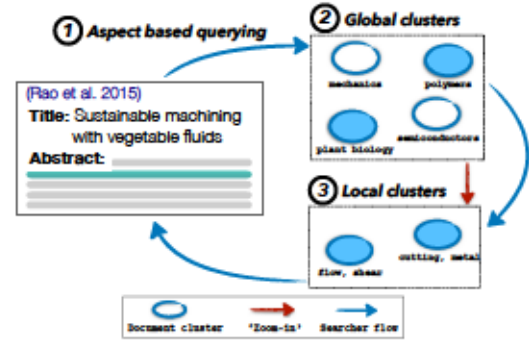
[1]https://github.com/olivettigroup/cross-domain-exploration

[2]Future systems may extend this to support sub-sentence spans or multiple-sentence queries.

Figure 1: We leverage free-form sentence level query- ing ① to retrieve documents across automatically built domain clusters ②. These are expanded based on searcher selection and re-clustered to allow navigation of nuanced differences within selected global clusters ③. Searchers can move between global clusters, 'zoom in' to examine global clusters, and reformulate their queries for additional alternative exploration.

of matching over prior work that mapped each ab- stract into a single 'best' schematic representation, we enable users to select any sentence in an ab- stract as interesting for querying. The system then retrieves papers similar to the selected aspect but different in terms of their domains. Domain clus- ters in our system were constructed using a pre- trained scientific document representation model trained on citation data (Cohan et al., 2020) captur- ing a global domain knowledge structure, and also do not require explicit labeled data in the process.

In order to center authentic user needs in the de- velopment of our prototype system, we adopted an iterative design process and sought feedback from a group of materials science researchers studying cli- mate change mitigation strategies. Due to the com- plex systems nature of climate change, develop- ment of mitigation strategies such as clean energy technologies and more sustainable building materi- als requires a cross-disciplinary approach (Xu et al., 2016) and related closely to the needs we set out to support in this research. We conducted case stud- ies with our system to gain deeper insights into whether the system retrieves results that represent meaningful 'jumps' between knowledge domains, whether scientists can make sense of them, and the design challenges for future exploratory search systems that facilitate cross-boundary inspirations.

2 System Description

Our system allows searchers to explore retrievals from global clusters and select interesting ones. The user may zoom into the clusters by retrieving additional results from them, which may reveal nuanced differences via local clustering. This in- teraction is supported by the 'Zoom in' button in
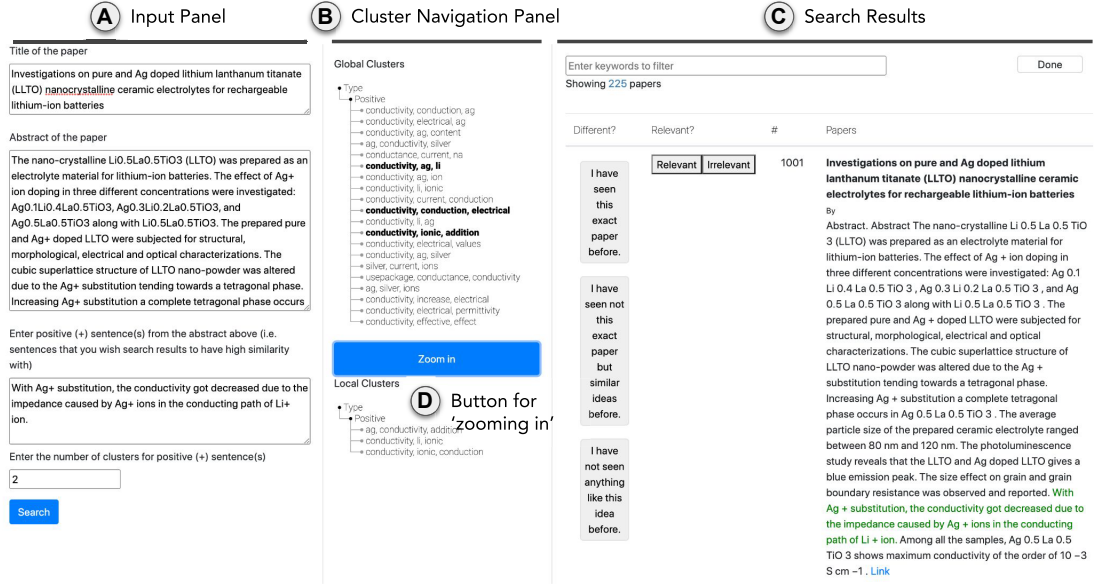
Figure 2: Front-end interface. (A) In the input panel, searchers enter the query paper abstract and a specific sentence aspect they want to align the retrieval results on. (B) Once searchers click on 'Search' at the bottom left, the system retrieves papers along with global domain clusters and displays them at the top of the navigation panel. Searchers can switch between different clusters to examine the results. Clicking on a cluster label highlights it and filters the papers that belong to the cluster in the table (clicking again on a selected cluster label de-selects it and refreshes the filtered table view). (C) The papers are displayed in the table on the right, with a green sentence highlight showing the closest aspect to the query. Searchers can type keywords in the filter at the top to find papers that contain them (matching keywords in the abstract are highlighted). Optional control buttons for our case studies for collecting feedback data are also shown in the table. (D) Clicking the 'Zoom in' button retrieves more papers from the selected clusters, allowing searchers to examine nuanced differences within the selected domain clusters.

the interface (Fig. 2). There are three main components of the system (Fig. 1): *Global clusters*. A set of clusters partitioning the entire corpus of papers into domains, built from paper embeddings computed with Cohan et al. (2020)'s transformer-based model; *Sentential representations*. Each sentence in an abstract is encoded using a state-of-the-art sentence encoder (sbert.net, 2021) to enable faceted retrieval at that level; *Local clusters*. A set of nested clusters built from query-specific retrievals within user-specified global clusters to reveal variations within the clusters. For the text corpus of our system, we used 3.2M Materials Science abstracts in the S2ORC corpus (Lo et al., 2020).

Global domain clusters. We represent broad subdisciplines of materials science with a set of global clusters, $G \in \mathcal{G}$. These serve as the primary means for searchers to control the distance of retrievals to a query. The clusters were obtained via citation-based similarity clustering using SPECTER embeddings of abstracts d's in the corpus (Cohan et al., 2020). We employ a K-Means clustering of the embeddings with $|\mathcal{G}| = 20$. This use follows from results indicating that pre-trained representations often capture domain structures well (Aharoni and Goldberg, 2020; Peng et al., 2021). We further validate these clusters in Appendix A.

Following the creation of $\mathcal{G}$, we index the sentences s, of documents in each cluster, $d_g \in G$, with a cluster specific approximate nearest neighbour search index, I_G, in preparation for search. We use the HNSW index of Faiss (Johnson et al., 2019) and a pre-trained sentence encoder, BERT_s to encode the sentences (sbert.net, 2021).

Sentence-level retrieval. We leverage the recently proposed faceted QbE search paradigm (Mysore et al., 2021) for querying, shown to support exploratory search tasks well (Lissandrini et al., 2019). Here, searchers query the system with an abstract q and a sentence in the abstract $s_q \in q$ indicative of the aspect of their interest.

We embed s_q with BERT_s, and retrieve the T nearest sentences from each global cluster index I_G. Users can then browse the results by switching between domains in the front-end interface, and additionally indicate specific global clusters $\mathcal{G}_s \subset \mathcal{G}$ they want to explore further. These serve as input to constructing query-specific local clusters.

Zooming in on global clusters with local clustering. Given a set of global clusters $\mathcal{G}_s$, and the query s_q, we retrieve the L nearest sentences to s_q from selected indices $I_{G \in \mathcal{G}_s}$. To support retrieval of more results from the same domain(s), we set $L > T$. We then re-cluster the similar document

sentences into a set of M clusters to show variations within the selected global clusters.

Finally, our prototype used a preliminary TFIDF scoring of unigrams in a cluster to extract descriptors for global and local clusters – the informativeness of the descriptors was however limited.

Interface used in the case studies. The main components of the front-end interface of our system are (See Fig. 2) Ⓐ: Query input panel, Ⓑ: Cluster navigation panel, Ⓒ: Retrieval table view with filtering, and Ⓓ: An optional 'Zoom in' button. Clicking this button retrieves more papers from the domain clusters searchers selected which allows them to explore and make sense of nuanced details among the results within the selected clusters.

3 Case Studies

We conducted think-aloud interviews (Fonteyn et al., 1993) with three participants who engage in Materials Science and Engineering research to probe 1) what potentially *distinct* values our system may bring to scientists exploring their own research questions compared to baseline systems, and 2) design challenges for realizing the full potential of search systems aimed at cross-domain exploration. We detail the study procedure in Appendix B.

4 Findings

Diverse results & complementary value. All participants commented on seeing more diverse results using our system over the baselines (Table 3). P1 compared his experiences as follows:

> "*On ConnectedPapers (P1's baseline system), I can see the overall relations between works such as who's citing who, and identifying what's well-cited...but focused in the closely related fields like geochemistry and dissolution of cement...In [our system], you can see that **it's picking up things that I quite frankly had no idea why it picked up at first.***" – P1

This quote demonstrates that while the prototype system did seem to retrieve more diverse results compared to the baseline system, the difference in results may require a closer look for apt engagement. Participants thought the value of retrieving such diverse results in the prototype system was complementary to that of the baseline systems. Both P2 and P3, who chose keyword query-based search engines as baselines, felt the baseline systems were useful for finding things in the domain they already knew: "*In regular search engines like SCIFINDER I can put in the exact query to find*

precisely that kind of papers" (P2). In comparison, "*[our system] is fantastic for learning more about a topic and kind of exploring...there are a lot of things that I found that I hadn't seen anywhere before.*" (P2). However, this value also seemed to depend on how focused the search goal was: "*It depends on what I'm trying to do...If I had a specific experiment in mind, for example tracking the activity of magnesium...I probably wouldn't be interested in looking at even iron or non-manganese materials...but in other times, I'll be open to seeing even non-battery domains so long as similar challenges are addressed.*" (P3).

Sourcing diversity from global domain clusters. This diversity seemed to originate from the differences in domains that participants perceived while interacting with the system. Specifically, participants thought the global clusters mapped to different application domains of similar mechanisms captured in their query sentences. For P1, this included clusters of papers focused on 'mixture strategies to create cement with radiation shielding to use in military bases' (as opposed to recapturing materials from industrial wastes to create mixtures for buildings for non-military use) or 'mixtures for bone-like scaffolds and implants in mice' (as opposed to buildings). In addition, P1 perceived the application domain of military bases as closer than Orthopedics in mice. An example P2 found interesting was a paper describing solid-state electrolytes for use in electrochromic mirrors, as opposed to the domain of battery technologies he was familiar with. This and other papers in the cluster showed P2 a whole new set of application domains related to optical properties of electrolytes that P2 did not know of. P3 found the results describing 'discharging' mechanisms interesting and useful for learning new contexts of research that explore similar mechanisms, e.g., papers on electrostatic discharging observed in clouds or the effects of mechanical properties in anode coatings on discharging.

Zooming in & realizing hidden assumptions. While reviewing the variations among the results, participants realized hidden assumptions that guided them to seek inspirations in 'typical' domains. For example, in seeing varying uses of a similar mechanism (e.g., discharging) in different contexts (e.g., electrostatic vs. electrochemical), P3 learned about other domains which she did not tap into for inspirations before: "*Now I can see that discharging happens everywhere, even outside the bat-*

tery domains. What I assumed was that discharging is only related to rate-related challenges in (an) electrochemical (context)." – P3. Unlike similarity-maximizing search, learning about various domains helped P3 realize her own pre-conceived notion of typicality of a search domain, broaden potential domains of inspiration, and connect previously disjoint clusters (e.g., [mechanical strategies], such as nanoindentations in anode coating materials, to [electrochemical] discharging efficiency).

Retrieving more results in selected domain clusters and surfacing local clusters not only assisted participants in making sense of the high-level similarity within the clusters but also helped them see nuanced differences. For P2, seeing results on [doping] and [sintering temperature] were on-point with regards to the core mechanistic relevance for his experiments (high-level similarity), while zooming in the clusters revealed more nuanced trade-off relationships such as how doping has been used as a mitigation strategy against loss of lithium at high temperatures. Examining the additional retrievals also led P2 to notice diverse sub-fields such as [powder metallurgy] and [laser deposition] that the nuanced mechanistic differences occurred.

Sometimes recognizing such nuanced differences helped participants make sense of how the search system worked: *"So these papers lack (connection to) dissolution and are mostly about production of cement which makes sense because my query sentence only mentions producing aggregates. But I thought because the rest of the abstract talks about dissolution of aggregates, the system would know that's the context I'm focusing on."* – P1. Interpreting how the system worked guided participants' query reformulation to counter their hidden assumptions or broaden the scope of search.

5 Implications for Design

These results confirm earlier findings that emphasize support for an iterative discovery of important and generative *mis*-alignments during analogical search (Kang et al., 2022a), while also contributing new challenges specific to cross-domain exploration. Compared to similarity-maximizing search that optimizes for alignment of search results, cross-domain inspiration requires identification of aspects in analogs that may be diversified for end-user inspiration versus those that need to be preserved for retaining relevance.

Specifically, how to identify clusters that lack important connections to query problems is challeng-ing because a cluster can be centered around a specific mechanism relevant to the query while lacking connections to other contextual mechanisms (e.g., a cluster of papers that focuses on production mechanisms of aggregate mixture but does not study the dissolution process of waste materials, which is a related mechanism for providing the ingredients to the production pipeline), or may not preserve important directionality of the cause-and-effect relationships (e.g., papers that demonstrate techniques to make electrolytes more inflammable may be useful in certain application domains but not for making battery technologies safer).

Given the importance of participants' reflection on their own assumptions and how the system performed search, future designs may focus on supporting *stateful* explorations and aim to help searchers keep track of which clusters they have visited, what they found as useful (and not useful) in them, and how they might reformulate subsequent search queries based on the previously hidden but salient aspects of their search intent. This information may provide valuable feedback to the system for tuning its behavior over time.

Lastly, unfamiliar domains may require explanations of relevance for searchers to scaffold engagement at a deeper level and to prevent early rejection based on snap judgment. We observed that organizing papers in domain clusters based on their high-level similarity to the query aspect serves as an important anchor for sensemaking. Furthermore, encouraging end-users to zoom into local clusters that exhibit nuanced differences may have led our expert users to *slow down* their judgment (Carolanne et al., 2007), allowing a deeper processing of results. However, open questions remain for future work to effectively communicate papers' relevance to scientists (e.g., Kang et al. (2022b)).

6 Conclusion

Our work presents an novel approach for exploring domain clusters to broaden sources of scientific inspiration and probe design implications for future interactive systems. Our case studies suggest that systems leveraging the global structure of knowledge domains and enabling flexible aspect-based retrieval help scientists find meaningful inspirations outside the domains they are familiar with. Additional work is needed to realize the full potential of interactive systems that augment scientific creativity via cross-domain inspirations.

7 Acknowledgments

References

KP Abhilash, P Christopher Selvin, B Nalini, P Nithyadharseni, and BC Pillai. 2013. Investigations on pure and Ag doped lithium lanthanum titanate (LLTO) nanocrystalline ceramic electrolytes for rechargeable lithium-ion batteries. *Ceramics International*, 39(2):947–952.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Pia Borlund. 2003. The concept of relevance in IR. *Journal of the American Society for information Science and Technology*, 54(10):913–925.

E Moulton Carol-anne, Glenn Regehr, Maria Mylopoulos, and Helen M MacRae. 2007. Slowing down when you should: a new model of expert judgment. *Academic Medicine*, 82(10):S109–S116.

Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*, pages 903–912.

Tanmoy Chakraborty, Amrith Krishna, Mayank Singh, Niloy Ganguly, Pawan Goyal, and Animesh Mukher-

jee. 2016. Ferosa: A faceted recommendation system for scientific articles. In *Proceedings, Part II, of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume 9652*, PAKDD 2016, page 528–541, Berlin, Heidelberg. Springer-Verlag.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Joel Chan, Steven P. Dow, and Christian D. Schunn. 2015. Do The Best Design Ideas (Really) Come From Conceptually Distant Sources Of Inspiration? *Design Studies*, 36:31–58.

Johan SG Chu and James A Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W Bruce Croft. 2018. Cross domain regularization for neural ranking models using adversarial learning. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1025–1028.

Khalid El-Arini and Carlos Guestrin. 2011. Beyond keyword search: Discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 439–447, New York, NY, USA. Association for Computing Machinery.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. 1993. A description of think aloud method and protocol analysis. *Qualitative health research*, 3(4):430–441.

D. Gentner, S. Brem, R. W. Ferguson, P. Wolff, A. B. Markman, and K. D. Forbus. 1997. Analogy and Creativity in the Works of Johannes Kepler. In T. B. Ward, J. Vaid, and S. M. Smith, editors, *Creative thought: An investigation of conceptual structures and processes*, pages 403–459. American Psychological Association, Washington D.C.

Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. 2013. Inspiration peak: exploring the semantic distance between design problem and textual

inspirational stimuli. *International Journal of Design Creativity and Innovation*, 1(4):215–232.

Kevin Heffernan and Simone Teufel. 2018. Identifying problems and solutions in scientific text. *Scientometrics*, 116(2):1367–1382.

Mary Hesse. 1965. Models and analogies in science. *British Journal for the Philosophy of Science*, 16(62).

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 235–243, New York, NY, USA. ACM.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel Weld, Marti Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 135–143, Online. Association for Computational Linguistics.

Tom Hope, Ronen Tamari, Hyeonsu Kang, Daniel Hershcovich, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2021. Scaling creative inspiration with fine-grained functional facets of product ideas.

Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J Marshall, and Byron C Wallace. 2018. Learning disentangled representations of texts with application to biomedical abstracts. In *EMNLP*, volume 2018, page 4683.

David G Jansson and Steven M Smith. 1991. Design fixation. *Design studies*, 12(1):3–11.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–42.

Hyeonsu Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022a. Augmenting scientific creativity with an analogical search engine. *Transactions on Computer-Human Interaction*.

Hyeonsu B Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S Weld, Doug Downey, and Jonathan Bragg. 2022b. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Aniket Kittur, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E Kraut, and Dafna Shahaf. 2019. Scaling up analogical innovation with crowds and ai. *Proceedings of the National Academy of Sciences*, 116(6):1870–1877.

Alex Ksikes. 2014. *Towards exploratory faceted search systems*. Ph.D. thesis, University of Cambridge.

Or Levi, Ido Guy, Fiana Raiber, and Oren Kurland. 2018. Selective cluster presentation on the search results page. *ACM Trans. Inf. Syst.*, 36(3).

Yuan Li, Yinglong Zhang, and Robert Capra. 2022. Analyzing information resources that support the creative process. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 180–190.

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2019. Example-based search: a new frontier for exploratory search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1411–1412.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Melanie Mitchell. 1993. *Analogy-making as perception: A computer model*. Mit Press.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. *NAACL*.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy. Association for Computational Linguistics.

R. Oppenheimer. 1956. Analogy in science. *American Psychologist*, 11(3):127–135.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liang Pang, Qingyao Ai, and Jun Xu. 2021. Beyond probability ranking principle: Modeling the dependencies among documents. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 1137–1140, New York, NY, USA. Association for Computing Machinery.

Hao Peng, Qing Ke, Ceren Budak, Daniel M Romero, and Yong-Yeol Ahn. 2021. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, 7(17):eabb9004.

Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. 2015. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574.

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search result diversification. *Found. Trends Inf. Retr.*, 9(1):1–90.

sbert.net. 2021. `sentence-transformers all-mpnet-base-v2`.

Soumya Shukla and Orland Hoeber. 2021. Visually linked keywords to support exploratory browsing. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 273–277.

Amin Sorkhei, Kalle Ilves, and Dorota Glowacka. 2017. Exploring scientific literature search through topic models. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics*, pages 65–68.

Nicole Sultanum, Christine Murad, and Daniel Wigdor. 2020. Understanding and supporting academic literature review workflows with litsense. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–5.

Don R Swanson and Neil R Smalheiser. 1996. Undiscovered public knowledge: A ten-year update. In *KDD*, pages 295–298.

Brandon Tran, Maryam Karimzadehgan, Rama Kumar Pasumarthi, Michael Bendersky, and Donald Metzler. 2019. Domain adaptation for enterprise email

search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–34.

Brian Traynor, Ciara Mulcahy, Hugo Uvegi, Tunahan Aytas, Nicolas Chanut, and Elsa A Olivetti. 2020. Dissolution of olivines from steel and copper slags in basic solution. *Cement and Concrete Research*, 133:106065.

Ryen W. White and Resa A. Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1:98.

Yi Xie, Yuqing Sun, and Elisa Bertino. 2021. Learning Domain Semantics and Cross-Domain Correlations for Paper Recommendation, page 706–715. Association for Computing Machinery, New York, NY, USA.

Xiaofeng Xu, Santonu Goswami, Jay Gulledge, Stan D Wullschleger, and Peter E Thornton. 2016. Interdisciplinary research in climate and energy sciences. *Wiley Interdisciplinary Reviews: Energy and Environment*, 5(1):49–56.

Dengmei Zhou, Chong Zeng, Jing Xiang, Tao Wang, Zitian Gao, Chunlian An, and Wanxia Huang. 2022. Review on Mn-based and Fe-based layered cathode materials for sodium-ion batteries. *Ionics*, pages 1–12.

A Evaluation of Global Clusters

To investigate the ability of clustering methods for separation of materials science corpora into coherent subdomains, we compare model generated clusters with author-provided keywords treated as gold standard labels.

Dataset. To generate our evaluation dataset domain experts first screened most frequent keywords in a multi-domain dataset of 700k English materials science publications and selected 18 keywords indicative of sub-domains (see Table 2). Next, only a subset of the corpus with documents assigned a single keyword of the 18 sub-domain keywords was retained for evaluation. This yields a corpus of 39,699 documents. A sample of these documents were checked manually to ensure that they represented reasonable paper-keyword assignments. This corpus is available.[3]

Experimental Protocol. We compare the method used in our system, SPECTER (Cohan et al., 2020) to TFIDF and SCIBERT baselines (Beltagy et al.,

---

[3] `https://github.com/olivettigroup/cross-domain-exploration`

	Keywords Present	Keywords Removed
RANDOM	27.37	27.37
TFIDF	**48.08**	39.75
SCIBERT	31.80	31.32
SPECTER	47.93	**43.90**

Table 1: Domain cluster purity for SPECTER, compared to baseline approaches on a corpus of materials science papers and author-provided keywords. Number of clusters = 18. Reported metrics represent cluster purity in percent, averaged over three runs. "Keywords Removed" represents the performance after removal of all category defining keywords (of Table 2) from the text of the title and abstract based on which paper representations are computed.

magnetic materials	carbon materials
ceramics	optical properties
electrochemistry	nanomaterials
alloys	photocatalysis
semiconductors	solar cells
fuel cells	polymers
composite materials	biomaterials
thermodynamics	lithium-ion batteries
thin films	microstructure

Table 2: Keywords representative of clusters in our cluster evaluation dataset.

2019). For evaluation of document clusters, embeddings are computed using publications titles and abstracts. SCIBERT and SPECTER input the papers as "Title [SEP] Abstract". To generate clusters we employ K-Means with $K = 18$. We report average cluster purity over three separate runs of clustering (Manning et al., 2008, C16.3).

Results. Results (Table 1) show SPECTERs' ability to induce subdomains clusters. Surprisingly high dimensional TFIDF embeddings perform similarly to transformer generated SPECTER embeddings (column: "Keywords Present"). An explanation for this might be found in the evaluation dataset we created – clusters based on author provided keywords, which are also likely to appear in the title and abstract of a paper are likely to see term-based methods (i.e. TFIDF) seeing strong performance. To investigate this further, we repeat the evaluation after removing the 18 category defining keywords (Table 2) from abstracts and title text before encoding papers. This experiment (column: "Keywords Removed") strengthens our hypothesis, showing a larger drop in the measured cluster purity values for the TFIDF approach compared to SPECTER. SPECTER's ability to induce clusters in the absence of explicit term overlap provides the value needed in our search setup.

B   Procedure and Interface Design Used in the Case Studies

Participants. Three participants (1 female) whose research areas are in the broader space of material science and engineering were recruited for interviews. The mean age of participants was 24 (SD=4.04). They all reported conducting a review of related work using scholarly search engines as a research facilitator and vehicle for learning.

Procedure. Participants prepared a query paper and a description of their research problem (Table 3) and chose search engines they personally use the most and are familiar with for the first stage (20 min) of the interview. Using the search engines of their choice, participants in the first stage looked for as many related work as possible for the query domain. The goal of this stage was to gain a sense of 'what is out there' which served as context for their judgement on retrieval results in the subsequent stage using our system. In the next stage, participants interacted with our system with the focus of finding different ideas than what they have found in the first stage (20 min). For the papers they found interesting, they provided feedback on two dimensions, 1) *Novelty* with the following ternary response options (1: *"I have seen this exact paper before"*; 2: *"I have not seen this exact paper but similar ideas before"*; 3: *"I have not seen anything like this before,"* similar to the scale used in (Kang et al., 2022a)) and 2) *Relevance* with binary response options. The collected feedback was used in the end-of-session debriefs to aid both the interviewer's and participants' memory and discussion grounded on specific examples.

C   Related Work

Our work bears resemblance to a number of prior lines of work. We group these broadly into methods for exploratory search of text corpora, and methodological NLP and information retrieval work which bears resemblance to parts of our system.

C.1   Exploratory Search for Text Corpora

A large body of work has explored the problem of exploratory search for text corpora, with searchers

PID	Research Domains	Seed Paper & Query sentence	Baseline Search System
1	New ways of producing cement using industrial wastes to lower the carbon footprint and maintain sustainability in commercial bldgs	Query sentence was the 2nd sentence in the abstract of Traynor et al. (2020)	CONNECTEDPAPERS[4]
2	New electrolyte materials for simultaneously increasing the capacity and stability of batteries	Query sentence: 8th abstract sentence in Abhilash et al. (2013)	SCIFINDER[5]
3	Understanding the barriers to scaling up the manufacturing of effective cathode materials	Query sentence: 2nd abstrct sentence in Zhou et al. (2022)	GOOGLE SCHOLAR

Table 3: Participants' research domains, seed papers used in the study, and baseline search engines they chose.

often interested in learning and discovery of new content. These search tasks are often characterized by an open-ended and multifaceted information need and involve a sizable browsing component (White and Roth, 2009, Sec 2.4). Given the importance of these aspects, we group the ways in which prior work represents documents in a system, this in turn, affect the ways that users query a search system. In describing past work we also highlight the ways in which they facilitate browsing.

Exploration via rhetorical structure. This line of work primarily consists of designing a schema of labels to describe the rhetorical structure elements of documents, automatically or manually extracting these elements from documents, and indexing and retrieving documents along these aspects in response to a query. Document representation with rhetorical structure elements follows from a large body of work demonstrating that problem solving is often characterized by descriptions of the Situation, Problem, Solution, and Evaluation (Heffernan and Teufel, 2018, Sec 1). In practice, a primary bottleneck of this line of work is its reliance on a schema of labels - making transitions to new corpora challenging. The closest prior work of this type is that of analogical search for creative ideation applied to science and design (Hope et al., 2017, 2021; Kang et al., 2022a). The method of query specification in this body of work is primarily via short text queries. In contrast, our system adopts a faceted query-by-example approach where searchers query with an example document and indicate an aspect of the document they would to focus on via selecting text spans, with documents indexed at the sentence level (Mysore et al., 2021). In this respect, work of Chan et al. (2018) and Neves et al. (2019) bear the closest resemblance to ours in supporting querying with documents in combination with a fixed schema of aspect labels.

While most work highlighted above retrieve a list of documents, Chan et al. (2018) and Hope et al. (2021) present exceptions in allowing browsing via domain-like structures - bearing resemblance to our approach. While Chan et al. (2018) search across domain structures informed by document metadata, our system relies on a set of automatically built global domain clusters and a set of query specific local clusters. Similarly, Hope et al. (2021), allow exploration of purpose and mechanism aspects via a hierarchical organization of these aspects constructed with a rule-mining approach. This approach, while well suited to use cases where aspects can be pre-extracted proves challenging to operationalize in the absence of these extractions. Our work complementarily prioritizes flexibility in query specification without a schema based document representation.

Other work also bear resemblance to the ones presented above. Jain et al. (2018) learn aspect disentangled paper representations with *aspect specific* encoders trained with similarity data aligned to aspects in biomedical abstracts. Chakraborty et al. (2016) present a system for retrieval of similar papers grouped along a schema of aspects based on citation and content similarities to a query. Ostendorff et al. (2020), present an approach to classify the aspect of similarity between a pair of papers rather than producing a ranking over documents. Portenoy et al. (2022) present a system for exploring *authors* across knowledge domains.

Exploration via topical structures. A separate line of work has also leveraged topical structures such as knowledge base entities (concepts), keywords, or other model inferred topic structures. Hope et al. (2020) present a system for exploration of biomedical concepts (e.g. proteins, diseases,

drugs) and research groups. Similarly, Shukla and Hoeber (2021) enhance academic search engine result pages with keywords available in document metadata, and Sorkhei et al. (2017) allow exploration of a corpus of papers via topics inferred with a LDA model. Besides the bottleneck of concept extraction in some of these methods, a key challenge of concept based exploration, as noted in Hope et al. (2020) and Sultanum et al. (2020), stems from users desire to continually explore additional concepts unavailable in a system or add concepts of their own. We alleviate this drawback in part by allowing users to directly query with text spans in a document context. That said, the two methods provide complementary value – while topical structures provide value in gaining an overview of a corpus our approach is likely to allow expression of more complex queries.

C.2 Search and Recommendation

A body of work in NLP, information retrieval, and recommendation bear resemblance to parts of our system. We group these into query-by-example methods and methods for cluster ranking and result diversification. While the former contextualizes our document representation and querying mechanism the latter contextualizes our result presentation.

Query-by-Example. This body of work primarily intends to retrieve documents given a query document and is well suited to exploratory search given its ability to express more context for under-specified queries (Ksikes, 2014; Lissandrini et al., 2019). This is best represented in El-Arini and Guestrin (2011), who present a system for personalized recommendation of papers given a set of user trusted papers - with documents presented as a diversified ranking of candidate documents based on citation network similarities. A large body of work under citation recommendation also recommends documents similar to a query document Färber and Jatowt (2020). Recent work has also seen strong performance for a variety of scientific paper recommendation tasks by training transformer bi-encoder models on citation network data (Cohan et al., 2020; Ostendorff et al., 2022). In contrast with this work, our work intends to model aspect conditional similarities. Closest work to ours is presented by Mysore et al. (2022) who extend these transformer bi-encoder models to a sentence level multi-vector model intended for finer-grained aspect conditional document similarities trained with co-citation data and textual supervision.

Cluster retrieval and diversification. In our generation of results based on a set of global clusters and a set of local clusters, our work departs from the standard IR assumption of treating candidate documents as independent from each other - Pang et al. (2021) overview this line of work. In doing so, our work relates to lines of work on search result diversification and cluster retrieval. Work in search result diversification most often intends to improve retrieval performance for ambiguous queries by optimizing the novelty of retrieved results or optimizing for coverage of all possible interpretations of a query, Santos et al. (2015) surveys this body of work in the context of search, while Kaminskas and Bridge (2016) survey similar work in the context of recommendations. Close work in cluster retrieval comes from Levi et al. (2018), who investigate the appropriate queries for which results should be grouped on a search result page. These lines of work differ from ours in intending to improve performance for ad-hoc search, our work on the other hand intends to use clusters of documents as a means of controlling "distance" to a query in an exploratory search setup. In this respect, Xie et al. (2021) represents closely related recent work in training a model for citation recommendation across-domain boundaries while leveraging an existing ontology of domain concepts to learn the domain specific semantics - training similar models for our setup represent exciting future steps.

Finally, cross-domain retrieval (Cohen et al., 2018; Tran et al., 2019) aims to improve in-domain performance of retrieval models in a transfer learning setup while our work focuses on retrieval across domain boundaries.