Fast Bibliography Pre-Selection via Two-Vector Semantic Representations

Wenchuan Mu
wenchuan_mu@sutd.edu.sg
Singapore University of Technology and Design
Singapore

Junhua Liu
junhua_liu@sutd.edu.sg
Singapore University of Technology and Design
Singapore

Kwan Hui Lim
kwanhui_lim@sutd.edu.sg
Singapore University of Technology and Design
Singapore

Abstract

In academic writing, bibliography compilations is essential but time-consuming, often requiring repeated searches for references. Hence, an efficient tool for faster bibliography compilation is needed. Our work offers a solution to the challenges of managing large-scale bibliographic databases, introducing a new algorithm that improves both efficiency and sensitivity. Using two-vector semantic modelling, bibliographic entries and queries are embedded into the same vector space to select relevant references based on semantic similarity. Experimental results with 3.37 million entries show the method reduces the time needed to generate a manageable subset, streamlining scholarly writing. Our code and dataset are publicly available at https://github.com/cestwc/bibliography-pre-selection.

CCS Concepts

• **Information systems → Retrieval tasks and goals**; **Recommender systems**; *Language models*; *Similarity measures.*

Keywords

Bibliography Selection, Recommendation Systems, Information Retrieval, Neural Networks, Digital Libraries

ACM Reference Format:
Wenchuan Mu, Junhua Liu, and Kwan Hui Lim. 2024. Fast Bibliography Pre-Selection via Two-Vector Semantic Representations. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3677389.3702495

1 Introduction

Compiling a bibliography is crucial but tedious in scientific writing [6, 45], frequently involving an initial broad literature review [14] followed by specific references selection [12, 40]. References can be added either manually by retrieving the official format from databases like ACM Digital Library or by using BibTeX files. While the manual approach is time-consuming, using BibTeX creates large files that slow down manuscript compilation, especially on platforms like Overleaf with file size limits. Dividing bibliographies to meet these constraints is inefficient as academic papers expand.

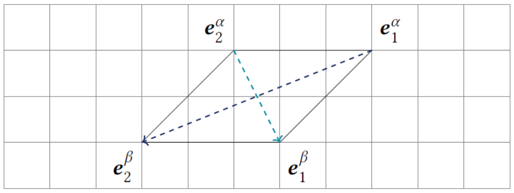

$\boldsymbol{e}_2^{\alpha}$
$\boldsymbol{e}_1^{\alpha}$
$\boldsymbol{e}_2^{\beta}$
$\boldsymbol{e}_1^{\beta}$

Figure 1: An illustration of $x_1 \bar{\rtimes} x_2$. A two-vector model can be visualised as a parallelogram representing two instances of citation information, where query and entry vectors form the vertices. Adjusting these vectors changes the shape, influencing citation probabilities. A rectangle would suggest equal citation likelihood. This analogy simplifies the complex, multidimensional nature of the two-vector representation.

The ideal scenario is where researchers have the exact bibliography they need when they start writing. Recent advancements in digital library systems and tools aim to improve the process of selecting relevant citations. For instance, Kreutz et al. [19] used formal process modelling techniques, while Dutta et al. [10] leveraged machine learning for these purposes. However, these systems still face challenges, such as overwhelming users with irrelevant articles or missing critical documents due to rigid search algorithms.

Contributions. Instead of using an exact subset, we propose a new approach to generate a superset of references based on minimal input, like a topic or title, without requiring detailed sections. The challenge is to minimise omissions. Instead of traditional binary classification, we suggest a two-vector embedding technique. This method embeds both queries and entries in vector space, allowing for efficient selection by measuring semantic similarity while accounting for citation asymmetry, as shown in Figure 1. Our experiments demonstrate high recall and effective subset generation.

2 Bibliography Pre-selection Problem

While pre-loading the full bibliography $\mathbb{B}$ brings convenience to users, it is not scalable due to the large file and being time-consuming to manage, e.g., more than 100 gigabytes if the scope of all articles is as large as Semantic Scholar. The querying procedure shifts from manual searches in digital libraries to automated searches during scholarly writing. Thus, we explore whether we can retain the benefits of a pre-loaded bibliography while avoiding an overly large data volume. More specifically, can we heuristically find a subset of all articles whose citation information is pre-loaded? This subset should cover all literature an author might possibly need.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

JCDL '24, December 16–20, 2024, Hong Kong, China
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702495

Definition 2.1 (Bibliography Pre-selection). Given a digital library $\mathbb{D}$, its associated bibliography set $\mathbb{B}$, and a user prompt query q, bibliography pre-selection produces a bibliography subset $\mathbb{S} \subset \mathbb{B}$.

The user prompt q in Definition 2.1 generally pertains to the topic the author intends to write about. Let α stand for the work an author intends to write. For simplicity, q can be considered a segment of the current $x_\alpha \in \mathbb{B}$, such as the title, abstract, and keywords. We denote the bibliography pre-selection process using S, *i.e.*, $\mathbb{S} \coloneqq S(\mathbb{D}, \mathbb{B}, q)$.

Objective of Bibliography Pre-selection Algorithms. Bibliography pre-selection aims to improve bibliographic search. A desirable pre-selection mechanism should allow users to seamlessly leverage the Bibkey copy-paste strategy for smooth scholarly writing. Two key properties are essential: minimizing undefined Bibkeys (*i.e.*, missing citation information in $\mathbb{S}$) and keeping the data volume of $\mathbb{S}$ manageable for efficiency. Formally, the optimisation of bibliography pre-selection is captured as follows, where uniform distribution $\mathcal{U}$ considers each article in the digital library equally important.

$$\begin{aligned} &\min_{S} \left\{ \mathrm{E}_{\mathbf{z} \sim \mathcal{U}(\mathbb{D})} \left[P\left(t \notin S\left(\mathbb{D}, \mathbb{B}, \mathrm{x}\right) \mid t \in \mathrm{ref}(\mathbf{z})\right)\right]\right\}, \\ &\qquad\qquad \text{subject to} \quad \forall z \in \mathbb{D}. \left|S\left(\mathbb{D}, \mathbb{B}, x\right)\right| < \delta. \end{aligned} \tag{1}$$

where $\mathbf{z}$ is a random variable of an article, and x denotes its associated citation information. $\mathrm{ref}(\mathbf{z})$ denotes references in $\mathbf{z}$. δ is a reasonable constant, *e.g.*, 50,000.

Therefore, when evaluating a bibliography pre-selection algorithm, we examine whether the constraint is satisfied and how effectively the objective is minimised. It is not necessary to let $\mathbb{S} = \mathrm{ref}(z_\alpha)$, as authors do not need to read or use all articles in $\mathbb{S}$.

Overall, the bibliography pre-selection problem is a specific type of personalised or customised bibliography selection, characterised by its properties as described in Equation 1. The aim of this study is to propose a method that effectively addresses this problem.

3 Two-vector Semantic Embedding for Bibliography Pre-selection

We present our solution to the bibliography pre-selection problem. As described in Equation 1, the goal is to select at most δ articles from the entire digital library $\mathbb{D}$, forming a bibliography set $\mathbb{S}$ based on the selection prompt q. The prompt q can be derived from any information related to z_α, such as an abstract or main text, but for simplicity, we assume citation information captures the essence of a work. Thus, we let q be a segment of x_α. In this study, we use information solely from $\mathbb{B}$ rather than $\mathbb{D}$. Features of cited and citing articles are represented by their respective citation information x, and we classify whether two pieces of citation information, *e.g.*, x_1 and x_2, indicate a citation relationship.

3.1 Two-vector Represented Feature Distance

We define the citation link $x_1 \rtimes x_2$ to indicate that article z_2 should cite or actually cites article z_1. We then use supervised machine learning to predict whether $x_1 \bar{\rtimes} x_2$ or $\neg x_1 \bar{\rtimes} x_2$.

Feature Encoding. Neural networks extract features from citation information differently depending on whether an article is citing or cited. Models like recurrent neural networks or transformers map raw text input into dense vector spaces [29, 42]. Tokens are first converted into high-dimensional one-hot vectors, which are then processed through layers that capture semantic relationships [9]. For example, each token $(x_i)_j$ from a vocabulary of size v is represented as a v-dimensional one-hot vector $\boldsymbol{v}_{i,j}$. A dense token vector in embedding space dimension m is then produced by multiplying the embedding matrix $\boldsymbol{W} \in \mathbb{R}^{v \times m}$ with $\boldsymbol{v}_{i,j}$, yielding $\boldsymbol{W}^T \boldsymbol{v}_{i,j}$.

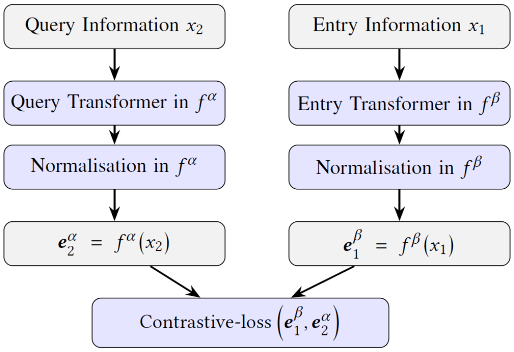

Query Information x_2
Entry Information x_1
Query Transformer in f^α
Entry Transformer in f^β
Normalisation in f^α
Normalisation in f^β
$\boldsymbol{e}_2^\alpha = f^\alpha(x_2)$
$\boldsymbol{e}_1^\beta = f^\beta(x_1)$
Contrastive-loss $\left(\boldsymbol{e}_1^\beta, \boldsymbol{e}_2^\alpha\right)$

Figure 2: Our training involves two distinct models with potentially identical architectures but different parameters. Features are extracted separately from the query and entry texts, generating fixed-size vector representations, $\boldsymbol{e}^\alpha$ and $\boldsymbol{e}^\beta$. We then compute the cosine similarity between vectors $\boldsymbol{e}_1^\beta$ and $\boldsymbol{e}_2^\alpha$ to quantify the similarity between the query and entry, with values ranging from [-1, 1] or distances from [0, 2].

Neural networks then capture contextual information in sequences of tokens to extract higher-level features. A function $\left(\boldsymbol{W}^T \boldsymbol{v}_{i,0}, \boldsymbol{W}^T \boldsymbol{v}_{i,1}, \ldots, \boldsymbol{W}^T \boldsymbol{v}_{i,j}, \boldsymbol{W}^T \boldsymbol{v}_{i,j+1}, \ldots\right) \mapsto \boldsymbol{e}_i$ is learned to map a sequence of token embeddings to encoded features $\boldsymbol{e}_i$. The entire feature encoding from x_i to $\boldsymbol{e}_i$ is learnable, denoted as $f : x_i \mapsto \boldsymbol{e}_i$. As the network trains on large text corpora, it adjusts weights to create embeddings where similar x have similar vector representations, capturing the semantics and nuances of each [34]. Similar vector representations typically have shorter distances between them, which is a fundamental principle in vector space models [28]. A threshold can often be used to decide if two vectors are "close enough" to be considered similar, facilitating various applications such as clustering and classification [21].

Two-vector Embedding. An article is more likely to cite a similar one rather than a very different one. Feature encoding of x allows us to estimate similarities between articles. However, most similarity measures, such as L^2 and L^1-distances, are symmetric, implying mutual citation, which is rare. To avoid this unnecessary constraint, we propose breaking the symmetry for greater flexibility (Figure 1).

We use two distinct vector models to represent the citing and cited articles differently. Specifically, when an x is (from) the citing work, its information is encoded by f^α, and when x is potentially the cited work, its information is encoded by f^β. Thus, we have:

$$\boldsymbol{e}_i^\alpha = f^\alpha(x_i), \quad \text{and} \quad \boldsymbol{e}_i^\beta = f^\beta(x_i). \tag{2}$$

Note that these two representation schemes map from the same input citation information space to the same output vector space.

Table 1: Demonstration of a small subset of the parallel dataset, where the inputs are x_α and x_β, and the label is the citation link between them. For simplicity, we include only the title information (a segment of x) in this demonstration. Note that the pre-selection links are to be trained based on the ground truth labels of the citation link $\rtimes$.

x_α, ***e.g.*, query title**	x_β, ***e.g.*, entry title**	**Label**
Emotion Cause Extraction - A Review of Various Methods and Corpora	Incongruent Headlines: Yet Another Way to Mislead Your Readers	`False` $(\neg x_\beta \rtimes x_\alpha)$
Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility	Evaluating and Enhancing the Robustness of Neural Network-based Dependency Parsing Models with Adversarial Examples	`True` $(x_\beta \rtimes x_\alpha)$

Thus, given two inputs $x_1, x_2 \in \mathbb{B}$, we can directly measure the distance between $f^\alpha(x_2)$ and $f^\beta(x_1)$, *e.g.*, L^2-distance.

Key Innovation. The proposed two-vector embedding separates vector models for queries and entries, making the learning of citation links reasonable, especially when a symmetric distance measurement between vectors is retained.

Same Vector Space but Slightly Shifted Encoding. To build two feature extractors mapping into the same vector space, we must ensure that for any article $\boldsymbol{z}$, the vectors $\boldsymbol{e}^\alpha$ and $\boldsymbol{e}^\beta$ are close to each other. The idea is that if a comprehensive semantic representation is created for all citation information in $\mathbb{B}$, not just for determining citation links, there would be one best representation scheme or a group of equivalent schemes that can be deterministically transformed into each other. Thus, $\boldsymbol{e}^\alpha$ and $\boldsymbol{e}^\beta$ should not deviate significantly from their "best-representing feature vector", denoted as $\boldsymbol{e}^\Omega$.

Our two-vector model uses two distinct models to transform citation information into vectors within the same dimensional space for comparison. The innovation is in using separate models for queries and entries, allowing for nuanced semantic representation based on whether the context is citing or cited. This approach provides two distinct distance metrics that capture the asymmetry of bibliographic relationships, leading to a more accurate citation context representation. Existing sentence embeddings are insufficient due to difficulties in aligning different representations and their inability to capture scholarly nuances. Therefore, we must train two feature extractors specifically for this task.

Vector Model Learning. In this section, we describe how f^α and f^β learn from the database $\mathbb{B}$ for accurate representation.

Parallel Dataset. We first transform $\mathbb{B}$ into a parallel dataset, where the input features are pairs of citation information in the α and β roles. The label is a binary value indicating whether the work in the α role cites the work in the β role. Since $\mathbb{B}$ is finite at any given time, the size of this parallel dataset, approximated by $|\mathbb{B} \times \mathbb{B}|$, is also finite. This dataset can be used to train various binary classifiers.

We train the proposed two-vector model using a parallel dataset, arranging f^α and f^β in a Siamese architecture, as shown in Figure 2. Unlike traditional Siamese networks, which use symmetric distance functions and shared weights between extractors [5, 39], we apply separate models for asymmetric inputs [44], mapping them into the same embedding space. The network learns to assess similarity between text pairs, with similarity scores ranging from 0 (dissimilar) to 1 (similar). A contrastive loss function encourages the network to minimise distances for similar pairs and maximise them for dissimilar ones [35], enhancing its ability to detect similarities.

Negative Sampling. A challenge arises due to the imbalance in the dataset, where True citation links are rare, making up no more than 1 in a million. Most articles cite only a few tens or hundreds of others, leaving the majority uncited. To address this, we use negative sampling during training [29]. The neural networks learn from every True citation link but only from a subset of False links. When learning False links, an uncited work is randomly selected from the same or earlier years. The sampling ratio λ controls the number of negative samples relative to positive ones. Typically, $\lambda = 1$ means one negative sample per positive sample, maintaining balance. Adjusting λ allows for fine-tuning, with higher values increasing negative samples to enhance precision.

3.2 Fast Vector Search

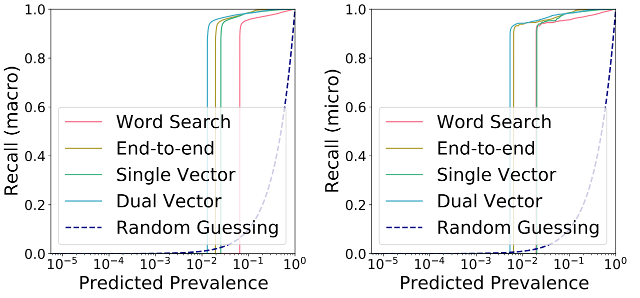
As each citation information x_i receives its two features $\boldsymbol{e}_i^\alpha$ and $\boldsymbol{e}_i^\beta$, we can determine the citation link statuses using vector search. This method is efficient because each feature is calculated only once, rather than recalculating it for every link determination. We use the K-D Tree [1] to fast select citation information instances most pertinent to the α instance. By querying the K-D Tree, we rapidly identify the k closest neighbours to $\boldsymbol{e}_i^\alpha$, among $\{\boldsymbol{e}_{i'}^\beta \mid i' = 1, 2, \ldots, |\mathbb{B}|; i' \neq i\}$. These k instances represent our pre-selected (most relevant) bibliography.

This approach reduces the computational load and time required by avoiding repetitive computations for each instance. The time complexity of selecting the k closest $\boldsymbol{e}_{i'}^\beta$ to a $\boldsymbol{e}_i^\alpha$ involves the following steps. For each of the $|\mathbb{B}|$, or simply n, entries in the $\mathbb{B}$, calculating the distance to the query vector results in $O(n)$ time. Sorting all distances has a time complexity of $O(n \log n)$. Once the k smallest distances are identified, retrieving the corresponding k entries is $O(k)$, typically negligible compared to the earlier steps. Thus, the overall time complexity is dominated by the sorting step $O(n \log n)$.

4 Experiments and Results

We conduct experiments to evaluate the performance of our bibliography pre-selection method. The goal is to integrate seamlessly into researchers' workflows, minimizing disruption. The pre-selected bibliography should achieve high recall, covering as many relevant works as possible while avoiding excessive entries that could hinder writing efficiency. We next describe the common train/test dataset setup for the experiments, followed by the experiments and results.

Research Question. We investigate the minimum subset size required for full recall and the recall-size trade-off. This ensures that the selected bibliography is comprehensive, supporting high-quality research with minimal manual effort. Intuitively, perfect recall can



(a) Micro recall w.r.t PP

(b) Macro recall w.r.t PP

Figure 3: Recall with respect to positive prevalence (PP) selection. (a) micro recall calculates for all query-entry pairs in the test dataset. (b) focuses on macro recall for each query, determining the proportion of required bibliographic entries successfully included for each query title.

be achieved by pre-selecting the entire set if the subset size is not constrained. However, as the set size constraint becomes tighter, achieving comprehensiveness becomes more challenging. This relationship can be illustrated with a PP-recall curve, where the horizontal axis represents the proportion of elements in the subset, and the vertical axis represents the recall. Figure 3a demonstrates this recall-size trade-off for various pre-selection mechanisms. Specifically, our method is compared with three baselines and generally performs best. When subset sizes are equal, our method has the highest recall; when recalls are the same, it achieves the smallest subset size.

Scholarly Dataset. We created a citation link parallel dataset using a subset of Semantic Scholar, selecting ACL Anthology articles as α articles. References not found in the ACL anthology are sourced from Semantic Scholar, resulting in 3.37 million samples. We used five-fold cross-validation, alternating between training and validation. Bibliographic inputs include titles and abstracts for efficient querying. We also tested with the PubMed Diabetes dataset [11], consisting of 19,717 α articles from a different domain.

Baselines and Proposed Method. We compared the proposed method against various baselines, including:

Word Search: A fast and straightforward rule-based method that uses ROUGE [23] to determine overlapping similarity scores based on the words used between two texts. Such word-based similarity search methods are also commonly used for information retrieval and recommendation systems [31].

End-to-end: A common strategy that concatenates all inputs (x_α and x_β) with separator tokens like `[SEP]` and apply an end-to-end binary classifier (BERT [9] in our case), which is often used in tasks like natural language inference [3], question answering [22], or topic-focused summarisation [30, 41].

Single Vector: A third baseline method is sentence embedding (RoBERTa-base), where similarity is evaluated from embedding vectors [35], and if greater than a threshold, a $\bar{\rtimes}$-link is given.

Two-Vector Model (Ours): The key innovation in the proposed method is using two slightly shifted features. We use transformer encoders with a 12-layer RoBERTa base model [25] to extract a normalised 768-dimensional vector from input text. These vector models are fine-tuned from single-vector sentence embedders for five epochs with a batch size of 8 and a learning rate of 2×10^{-4}, using the AdamW optimiser [26]. The training schedule includes a 10,000-step warm-up, linear learning rate decay, a 0.1 dropout rate, and a 1.0 maximum gradient norm for stability.

Metrics: With some notation abuse, recall (true positive rate) is $\left(\#\left(\bar{\rtimes}, \rtimes\right)/\#\left(\bar{\rtimes}\right)\right)$. Predicted prevalence (PP) is $\left(\#\left(\bar{\rtimes}\right)/\left(\#\left(\bar{\rtimes}\right)+\#\left(\neg\bar{\rtimes}\right)\right)\right)$. We use the macro and micro PP-recall curves to show the recall-size trade-off, which can be converted from a traditional ROC curve by $\mathrm{PP} = (\mathrm{FPR} \times \mathrm{N} + \mathrm{TPR} \times \mathrm{P})/(\mathrm{P} + \mathrm{N})$.

Result. Our two-vector model, evaluated under different positive rate thresholds, matches the performance of a binary classifier trained on the same data and outperforms the single-model system or keyword search. We aim to find the minimum subset size for full recall in the bibliography selection. By varying the predicted prevalence, we observed increases in the recall. At a threshold of 1.09×10^{-3} or higher, 95% of queries successfully retrieved all required entries. This balance demonstrates high correctness in bibliography selection while reducing review volume.

5 Related Work

Bibliographic systems have evolved from manual processes to AI-driven models [17]. Early systems, as from Luhn [27], used automated indexing with simple statistics, followed by the Boolean model for complex queries [37]. The vector space model [36] improved document ranking with cosine similarity, and machine learning models like BERT [9] further enhanced search accuracy. However, challenges like query ambiguity and document diversity remain [7, 32]. Semantic analysis in bibliographic selection has advanced with keyword indexing [15], Boolean logic, controlled vocabularies [18], and topic modelling [38]. With the recent advancement of natural language processing and Latent Semantic Indexing [8], there is a significant improvement in concept understanding, and ontology-based systems with an enhanced contextual understanding of terms [13, 24].

Recent research aim to improve digital library functionality [43]. Chekuri et al. [4] developed a system for long documents with advanced search features, while Kreutz et al. [20] introduced SchenQL, a query language for complex bibliographic searches. Personalisation has also become a key focus, with Nomoto [33] proposing a method for re-ranking search results based on user preferences. Additionally, Bevendorff et al. [2] introduced the Scientific Multi-Authorship Corpus (SMAuC) to study collaboration patterns in academic publications, informing future bibliography systems.

6 Conclusion

We proposed a two-vector representation technique to enhance bibliographic selection by improving efficiency and accuracy through semantic analysis. Addressing citation relationship asymmetry and utilising advanced algorithms balances recall and operational speed while reducing computational resources. This scalable approach shows promise for larger databases, offering a valuable tool for academic research. Future work could involve more sophisticated models and improved user interaction, and extending beyond bibliography selection to recommendation tasks, such as identifying similar items to recommend to users [16].

Acknowledgments

This research is supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20123-0015). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore. W. Mu and K. H. Lim are also supported by a CoT R&D Programme (Award No. COT-V2-2020-1).

References

[1] Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (1975), 509–517.

[2] Janek Bevendorff, Philipp Sauer, Lukas Gienapp, Wolfgang Kircheis, Erik Körner, Benno Stein, and Martin Potthast. 2023. SMAuC - The Scientific Multi-Authorship Corpus. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 25–29. https://doi.org/10.1109/JCDL57899.2023.00013

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

[4] Satvik Chekuri, Prashant Chandrasekar, Bipasha Banerjee, Sung Hee Park, Nila Masrourisaadat, Aman Ahuja, William A. Ingram, and Edward A. Fox. 2023. Integrated Digital Library System for Long Documents and their Elements. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 13–24. https://doi.org/10.1109/JCDL57899.2023.00012

[5] S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 539–546 vol. 1. https://doi.org/10.1109/CVPR.2005.202

[6] Harris M. Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society* 1, 1 (01 Mar 1988), 104. https://doi.org/10.1007/BF03177550

[7] Steve Cronen-Townsend, W Bruce Croft, et al. 2002. Quantifying query ambiguity. In *Proceedings of HLT*, Vol. 2. 94–98.

[8] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[10] Debanjan Dutta, Dipasree Pal, Dwaipayan Roy, and Mandar Mitra. 2023. Bibliography Counselor: A Citation Recommendation Tool. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 260–262. https://doi.org/10.1109/JCDL57899.2023.00051

[11] Lise Getoor Galileo Mark Namata, Ben London and Bert Huang. 2012. Query-Driven Active Surveying for Collective Classification. In *International Workshop on Mining and Learning with Graphs*. MLG, Edinburgh, Scotland.

[12] Eugene Garfield. 1955. Citation Indexes for Science. *Science* 122, 3159 (1955), 108–111. https://doi.org/10.1126/science.122.3159.108 arXiv:https://www.science.org/doi/pdf/10.1126/science.122.3159.108

[13] Nicola Guarino and Pierdaniele Giaretta. 1995. Ontologies and knowledge bases. *Towards very large knowledge bases* (1995), 1–2.

[14] Chris Hart. 2018. *Doing a Literature Review: Releasing the Research Imagination*. SAGE Publications Ltd, London. 352 pages. http://digital.casalini.it/9781526423146

[15] Markus Heckner, Susanne Mühlbacher, and Christian Wolff. 2008. Tagging tagging. Analysing user keywords in scientific bibliography management systems. (2008).

[16] Ngai Lam Ho, Roy Ka-Wei Lee, and Kwan Hui Lim. 2023. SBTREC - A Transformer Framework for Personalized Tour Recommendation Problem with Sentiment Analysis. In *2023 IEEE International Conference on Big Data (BigData)*. 5790–5798. https://doi.org/10.1109/BigData59044.2023.10386486

[17] James Oluwaseyi Hodonu-Wusu. 2024. The rise of artificial intelligence in libraries: the ethical and equitable methodologies, and prospects for empowering library users. *AI and Ethics* (19 Feb 2024). https://doi.org/10.1007/s43681-024-00432-7

[18] Yufeng Jing and W Bruce Croft. 1994. *An association thesaurus for information retrieval*. University of Massachusetts, Department of Computer Science.

[19] Christin Katharina Kreutz, Martin Blum, Philipp Schaer, Ralf Schenkel, and Benjamin Weyers. 2023. Evaluating Digital Library Search Systems by Using Formal Process Modelling. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 1–12. https://doi.org/10.1109/JCDL57899.2023.00058

[20] Christin Katharina Kreutz, Martin Blum, and Ralf Schenkel. 2022. SchenQL: a query language for bibliographic data with aggregations and domain-specific functions. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 37, 5 pages. https://doi.org/10.1145/3529372.3533282

[21] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 1188–1196. https://proceedings.mlr.press/v32/le14.html

[22] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A Dataset for Multi-Span Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1250–1260. https://doi.org/10.18653/v1/2022.naacl-main.90

[23] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[24] Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. 2024. LARA: Linguistic-Adaptive Retrieval-Augmentation for Multi-Turn Intent Classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Miami, Florida.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[26] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. *arXiv preprint arXiv:1711.05101*. https://openreview.net/forum?id=Bkg6RiCqY7

[27] H. P. Luhn. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development* 1, 4 (1957), 309–317. https://doi.org/10.1147/rd.14.0309

[28] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[30] Wenchuan Mu and Kwan Hui Lim. 2024. Label-Free Topic-Focused Summarization Using Query Augmentation. In *2024 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN60899.2024.10650321

[31] Wenchuan Mu and Kwan Hui Lim. 2024. Modelling Text Similarity: A Survey. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Kusadasi, Turkiye) *(ASONAM '23)*. Association for Computing Machinery, New York, NY, USA, 698–705. https://doi.org/10.1145/3625007.3627305

[32] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 183–188.

[33] Tadashi Nomoto. 2012. Re-ranking bibliographic records for personalized library search. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (Washington, DC, USA) *(JCDL '12)*. Association for Computing Machinery, New York, NY, USA, 125–128. https://doi.org/10.1145/2232817.2232841

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, Kentaro Inui, Jing Jiang,

Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[36] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[37] Gerard Salton and Michael E Lesk. 1965. The SMART automatic document retrieval systems—an illustration. *Commun. ACM* 8, 6 (1965), 391–398.

[38] Trisha Singhal, Junhua Liu, Lucienne T. M. Blessing, and Kwan Hui Lim. 2021. Analyzing Scientific Publications using Domain-Specific Word Embedding and Topic Modelling. In *2021 IEEE International Conference on Big Data (Big Data)*. 4965–4973. https://doi.org/10.1109/BigData52589.2021.9671598

[39] Trisha Singhal, Junhua Liu, Wenchuan Mu, Lucienne T. M. Blessing, and Kwan Hui Lim. 2024. Photozilla: An Image Dataset of Photography Styles and its Application to Visual Embedding and Style Detection. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Kusadasi, Turkiye) *(ASONAM '23)*. Association for Computing Machinery, New York, NY, USA, 445–449. https://doi.org/10.1145/3625007.3627476

[40] Linda C Smith. 1981. Citation analysis. (1981).

[41] Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve Query Focused Abstractive Summarization by Incorporating Answer Relevance. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 3124–3131. https://doi.org/10.18653/v1/2021.findings-acl.275

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[43] Norha M Villegas, Cristian Sánchez, Javier Díaz-Cely, and Gabriel Tamura. 2018. Characterizing context-aware recommender systems: A systematic literature review. *Knowledge-Based Systems* 140 (2018), 173–200.

[44] Xiao Wang, Haoqi Fan, Yuandong Tian, Daisuke Kihara, and Xinlei Chen. 2022. On the importance of asymmetry for siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16570–16579.

[45] Paulus Franciscus Wouters et al. 1999. *The citation culture*. Ph. D. Dissertation. Universiteit van Amsterdam Amsterdam.