

SUSHI BAR: Visualizing Results for Experiments with Searching Physical Archives

Victor Hugo O. Leão¹[0009–0003–6378–7605], Thiago De Paulo Faleiros¹[0000–0001–5275–8356], and Douglas W. Oard²[0000–0002–1696–0407]

¹ University of Brasilia, Brazil

`victor.pessoal1203@gmail.com, thiagodepaulo@unb.br`

² University of Maryland, College Park, MD, USA

`oard@umd.edu`

Abstract. It is well known that search results differ markedly from one search topic to another. What is less well studied is the effect on search results of different training sets. This demo paper presents a system for jointly visualizing variations in nDCG@5 from different topics and different training sets. The demo uses the NTCIR SUSHI test collection, which models an archival access task in which a few paper documents in a physical archive have been digitized (these documents are the training set) and the goal is to predict the location of undigitized relevant documents in the original (paper) archive. Point estimates and confidence intervals are computed by randomly selecting different training sets to visualize how those vary with choice of topic and choice of retrieval system. The results show that evaluation using a single training set can yield misleading results, and that the combination of point estimates and confidence intervals better characterizes system and topic effects.

Keywords: Visualization · Evaluation · Archival access.

1 Introduction

For summative evaluation, we prefer search systems that do best by some measure of effectiveness (Spärck Jones and Galliers, 1995). Because which system is best can vary from one search topic to another, a common approach is to average the measure over many topics to make claims about which system is likely to be most useful for future topics. In recent years, other sources of variation such as the query used to express a search topic (Culpepper et al., 2021; Penha et al., 2022) and the document collection to be searched (Thakur et al., 2021) have also been characterized. Less well studied has been variation in retrieval effectiveness that results from use of different training sets. It is that variation, across training sets, that motivated us to create the system we will demonstrate.

When training data is abundant, how best to use that data is simply an optimization detail. It is when training data is expensive, and thus limited, that variations from different training data choices are of particular interest to us. For our demo we use the NTCIR-18 SUSHI test collection (Suzuki et al., 2025), which

models such a case.¹ This test collection is designed to model searching a large physical archive (e.g., a building) for paper documents. In the SUSHI search task, documents are stored in folders, folders in boxes, and boxes on shelves. Because some physical collections are very large, it may only be practical to digitize (or otherwise describe) a small fraction of the documents. The SUSHI search task is therefore not to directly find documents, but rather to find where in the archive to look for those documents. Searchers will then typically need to request that one or more boxes be fetched, and then look through the folders in those boxes (Wiedeman, 2019). In SUSHI, the task is to rank folders in decreasing order of the likelihood of finding relevant documents there.

In the SUSHI test collection, the limited training data is modeled by selecting a small number of documents (e.g., a few per box) that we consider as having been digitized, and then the system’s task is to use that information (together with folder labels and the storage order of the boxes) to infer the likelihood of finding relevant documents in a folder, given a query that expresses the search topic. In the NTCIR-18 SUSHI evaluation, each topic was evaluated using one training set, so the results from that evaluation can only tell us which retrieval systems are preferred when the system was trained on that one set. In our demo, we characterize retrieval effectiveness using ensembles of randomly selected training sets. For details on the SUSHI test collection itself we refer the reader to the NTCIR-18 task overview (Suzuki et al., 2025).

Several tools support inspection of retrieval experiments. TrecTools (Palotti et al., 2019) unifies pooling, evaluation, fusion, and significance testing in one Python library and can generate plots and L^AT_EX tables, though that output is static rather than interactive. ASPIRE (Peikos et al., 2025) closes that gap with a Streamlit app offering multi-experiment comparison, per-query analysis, and analysis of relevance judgments. However, both model experiments in which the sources of variation are system, topic, and query, over a collection that is fully indexed. The SUSHI task adds a further source of variation: which documents happen to have been digitized. Our SUSHI BAR system treats the training set as an experimental variable, which makes it possible to see whether a topic is intrinsically hard, or merely hard with some specific training set. It also provides viewers for inspecting the SUSHI test collection and the topics directly, giving newcomers a way into the task before they run any experiments at all.

2 Visualizing the Search Task

SUSHI BAR,² a Streamlit web app, is organized into five main sections accessible through sidebar navigation: Collection Viewer, Task Viewer, Training Set Viewer, Single Experiment Viewer and Two-Experiment Viewer. There is also a How-To Guide that briefly explains the SUSHI task, the archival hierarchy, and

¹ SUSHI stands for Searching Unseen Sources for Historical Information. What we call the SUSHI test collection was used for Subtask A of the NTCIR-18 SUSHI task.

² Online demo is available at <https://tinyurl.com/sushibarapp>; code for local use is at https://github.com/victorleao/SUSHI_Information_Retrieval_Archives/

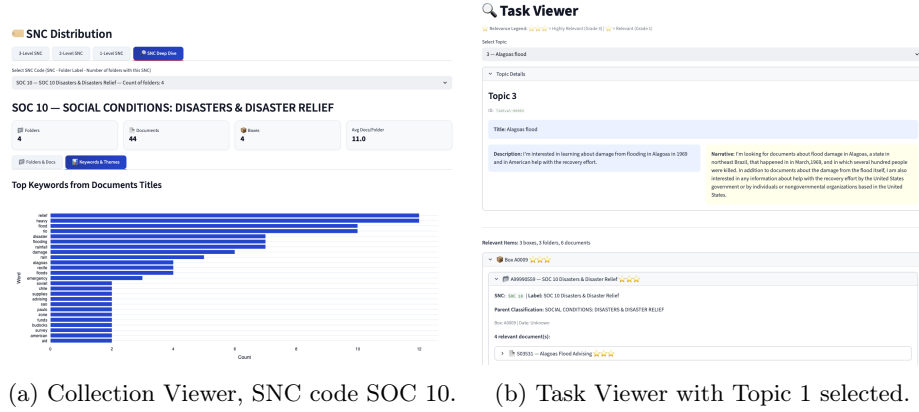


Fig. 1: Collection and Task Viewers.

task-specific nomenclature. All data is read directly from the raw test collection files or from results generated by experiments. This design choice allows a user with the SUSHI collection and a set of experiment result files to launch SUSHI BAR locally. Researchers with new ways of generating training sets, or even new similarly-structured test collections, could use the system locally as well.

The Collection Viewer shows the structure of the test collection and provides summary statistics. It shows the distribution of folders per box, and of documents per folder. One unusual characteristic of the SUSHI collection is that every folder is labeled with a Subject-Numeric Classification (SNC) code,³ so it contains a number of ways of seeing how SNC codes were actually used. For example, the Collection Viewer shows the title and a GPT-4o summary for any document in a folder labeled with a user-selected SNC code. Aggregate statistics for that document set can also be explored using an “SNC Deep Dive” to see a details for folders labeled with a specific SNC code. For example, Figure 1a shows that the most common words in the titles of documents found in folders labeled with SNC code “SOC 10” (Disasters & Disaster Relief) are *relief*, *heavy* and *flood*.

The Task Viewer (Figure 1b) shows the test collection’s 45 topics and the documents that are relevant to each. Topics contain three fields: Title (short query-like statement of what is sought), Description (sentence or two on what is sought), and Narrative (details on how to judge relevance to this topic). A query for each topic can be constructed by concatenating some subset of the topic fields. For each document that is relevant to a topic, the Task Viewer shows the location (box and folder) of that document and its degree of relevance (one star for relevant, three for highly relevant). Given the query, the task is to rank folders that contain relevant documents as highly as possible, with folders containing highly relevant documents ahead of all other folders.

³ SNC codes were manually assigned when documents were placed in folders using the State Department filing system of the time. They provide an indication of broad themes addressed by the documents in a folder, and they exist for every folder.

In reality, all of the documents in the SUSHI collection have been digitized, with each assigned a title by human indexers (even for documents that did not originally have a title). To model incomplete digitization, we mask out the digitized content and titles for most of the collection to produce a training set that averages 5 documents per box. We define three types of training masks. For our *Uniform* condition, we draw five documents per box without replacement. For our *Skewed* condition, we draw a variable number of documents per box, using a Zipfian distribution that models actual digitization rates.⁴ Sampling was done without regard to relevance, so a training set can include documents that are relevant to some topics. All topics in a run share the same mask.⁵ We also ran an *All docs* condition, indexing the full collection, to show how much better a folder ranking system could do if the entire collection had been digitized.

We are interested in training set effects on folder ranking, so SUSHI BAR includes a Training Set Viewer (not shown). It presents summary statistics for the training set, such as number of documents, how those documents are distributed over the 126 boxes, and how they are distributed over the folders in each box. These same statistics can also be aggregated by SNC code. A breakdown of folders containing documents relevant to each topic that shows how many of those folders have at least one document in the training set is also available.

3 Experiment Viewers

We demo SUSHI BAR with experiments that follow the multi-stage design introduced in Oard et al. (2025) in which the folders containing training documents are first scored, then those scores are used to estimate scores for other folders, then all of those scores are combined with scores based on folder labels alone (which are known for every folder). We adopt the run naming notation from Leao et al. (2026) in which the runs differ in **Training Sample Type** (**U5**: Uniform 630 docs (5 per box), **K5**: sKewed 630 docs, or **A-**: All 31,681 docs); **Query Type** (**T-**: Title only, **TD-**: Title and Description concatenated, or **TDN**: all three fields); **Training Document Scoring** (**B**: BM25F (Robertson et al., 2004), **L**: BM25F with Learned weights,⁶ **C**: ColBERT (Santhanam et al., 2022), **E**: cosine similarity between query and document *all-mpnet-base-v2* Embeddings (Mitra and Craswell, 2018), **X**: Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) of B and C, **Y**: RRF of B and E, **Z**: RRF of B, C and E, **V**: RRF of L and C, or **W**: RRF of L, C and E), **Indexed Fields for Training Documents** (some combination of **T**: Title metadata, **O**: OCR, **F**: labels on Folders containing the documents, and **S**: GPT-4o Summaries of the OCR); **Score Propagation to Folders Outside the Training Set** (related to a scored folder by **m**: exact SNC Match, **x**: same boX, or **s**: Similar SNC, rescaled so no

⁴ Modeled using 126 US National Archives record group digitization rates (National Archives and Records Administration, 2023), as first done by Suzuki et al. (2024).

⁵ The NTCIR-18 SUSHI task specified 3 Uniform masks, each assigned to 15 topics. For each of our 30 runs we generated a single training mask for all 45 topics.

⁶ Weights were learned on held out topics provided for the NTCIR-18 SUSHI dry run.

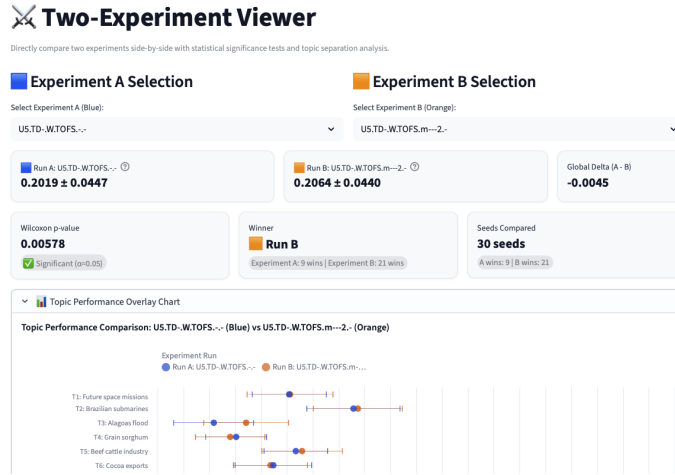


Fig. 2: Comparing per-topic results with and without score propagation.

propagated score is placed above rank **1**, **2** or **3**); and **Full-Collection Folder Label Scoring** (**b**: BM25 or **c**: ColBERT). A dash in any position indicates an option that was not used. Each experiment is named using a dotted notation such as U5.TD-.B.T-FS.m--2.c (Uniform sample w/5 documents per box, Title and Description queries, BM25F over the Title, Folder label and Summary fields (but not OCR), score propagation to folders with an exact SNC match, no higher then rank 2, and full-collection folder label search with ColBERT).

The Single Experiment Viewer (not shown) depicts how retrieval effectiveness varies across topics and training sets. It displays a dumbbell plot with one point estimate per topic showing mean nDCG@5 across the 30 training sets together with a 95% confidence interval. Unlike confidence intervals over topics, as reported in the NTCIR-18 SUSHI overview (Suzuki et al., 2025), these intervals characterize variability in per-topic results from the choice of training set.

The Two-Experiment Viewer (Figure 2) allows selection of any two experiments, depicting their per-topic comparison using paired dumbbell plots. Per-topic details are also provided for which system performed notably better, notably worse, or about the same (within 10%). Statistical significance for differences is based on a Wilcoxon signed-rank test, averaging over topics and pairing by training set, with no correction for multiple tests.

4 Examples of Using the System for Analysis

One result from the NTCIR-18 SUSHI task was that for 11 of the 45 topics, no system found any folder with any relevant documents in their top 5 results (i.e., per-topic nDCG@5=0) (Suzuki et al., 2025). With multiple training sets, we can now see that this is a training set effect, since even BM25F (e.g., U5.TD-.B.T.OFS.----.-, a system tested at NTCIR-18) yields a non-zero

averages for each topic over 30 training sets. Indeed, for our strongest overall system with uniform sampling and TD- topics (U5.TD-.W.TOFS.s--2.c), 95% confidence intervals include zero only for 6 of the 45 topics.

Another NTCIR-18 result was finding that score propagation had no measurable benefit in those experiments. In particular, Oard et al. (2025) found that no system ranked any relevant folder in the top five for 5 of the 6 topics that had no relevant document in the training set (for the sixth, success resulted not from score propagation but from full-collection search of folder labels). We can now use our Two-Experiment Viewer to make per-topic comparisons between U5.TD-.W.TOFS.m--2.- and U5.TD-.W.TOFS.----.-, the same system without score propagation. A direct comparison (Figure 2) shows improvements in expectation (i.e., in nDCG@5 when averaged over 45 topics) for score propagation for 21 of the 30 training sets. This improvement is statistically significant at $p < 0.01$. However, comparing the other score propagation approach that we tried (U5.TD-.W.TOFS.m--1.-) shows no statistically significant difference from no score propagation. From this we conclude that (1) score propagation can help, (2) how that score propagation is implemented makes a difference, and (3) experiments in which multiple training sets are used are needed to see those effects.

5 Conclusion

We present SUSHI BAR, a visualization system for inspecting results of folder-ranking experiments in sparsely digitized paper archives. One important contribution it brings is the ability to visualize how retrieval effectiveness varies across training sets, rather than reporting a single score obtained from one sample of digitized documents. A second is that the same system makes the SUSHI task itself more easily explored, letting a researcher trace a topic from its relevant documents to the folders and boxes that hold them, and see how much of that evidence a given training set actually exposes. Because this task is less well studied than the problem of searching fully accessible digital content, SUSHI BAR is designed to illustrate and explore novel aspects of the task, such as the presence of SNC codes and the dependence on the how the sparsely digitized training set was chosen. Our results show that evaluating against a single training set can substantially misrepresent topic difficulty, and that using a single training set can preclude accurate measurement of some effects from differences in system design. Thus we have found beneficial effects for both summative and formative evaluation from the use of multiple training sets. Our system can easily be run locally to ingest new experiment results, and thus it can now be used by others to test and compare novel system designs.

Acknowledgments

First author supported in part by Federal District Research Support Foundation (FAPDF) Grant 02/2025-FAPDF Participa; second author by National Council for Scientific and Technological Development CNPq Grant 200149/2025-3.

Bibliography

- Gordon V. Cormack, Charles L.A. Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 758–759.
- J Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2021. Topic Difficulty: Collection and Query Formulation Effects. *ACM Transactions on Information Systems* 40, 1 (2021), 1–36.
- Victor Hugo O. Leao, Thiago De Paulo Faleiros, and Douglas W. Oard. 2026. Characterizing Training Data Effects with the NTCIR-18 SUSHI Test Collection. In *Proceedings of the 12th International NTCIR Workshop on Evaluating Information Access*.
- Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval* 13, 1 (2018). <https://doi.org/10.1561/15000000061>
- National Archives and Records Administration. 2023. Record Group Explorer. <https://www.archives.gov/findingaid/record-group-explorer>, visited August 3, 2026.
- Douglas W. Oard, Shashank Bhardwaj, and Emi Ishita. 2025. Biting into SUSHI: The University of Maryland at NTCIR-18. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*.
- João Palotti, Harris Scells, and Guido Zuccon. 2019. TrecTools: an Open-source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1325–1328.
- Georgios Peikos, Wojciech Kusa, and Symeon Symeonidis. 2025. ASPIRE: Assistive System for Performance Evaluation in IR. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval*. 65–71.
- Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *European Conference on Information Retrieval*. 397–412.
- Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. 42–49.
- Keshav Santhanam, Omar KhatTab, Christopher Potts, and Matei Zaharia. 2022. PLAID: An Efficient Engine for Late Interaction Retrieval. arXiv:2205.09707 [cs.IR] <https://arxiv.org/abs/2205.09707>
- Karen Spärck Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An analysis and review*. Springer Science & Business Media.

- Tokinori Suzuki, Douglas W. Oard, Shashank Bhardwaj, Emi Ishita, and Yoichi Tomiura. 2025. NTCIR-18 SUSHI Pilot Task Overview. In *Proceedings of the 18th NTCIR Conference Evaluation of Information Access Technologies*.
- Tokinori Suzuki, Douglas W. Oard, Emi Ishita, and Yoichi Tomiura. 2024. Searching for Physical Documents in Archival Repositories. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2614–2618.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Gregory Wiedeman. 2019. The Historical Hazards of Finding Aids. *The American Archivist* 82, 2 (2019), 381–420.