Contrastive Session Graphs for Long-Tail Query Reformulation in Web Search

Elena Marsh[1] Kavi Rajan[2] Dorothea Lindqvist[1] Marcus Feld[3]

[1]Search & Discovery Group, Nexa Research [2]Department of Computer Science, Vertex University [3]Data Systems Lab, Meridian Institute of Technology

{e.marsh,d.lindqvist}@nexaresearch.example k.rajan@vertex.edu m.feld@meridian-tech.example

Abstract

Search sessions for long-tail queries are dominated by sparse, noisy click signals, which makes learning reliable query reformulation policies difficult for standard learning-to-rank pipelines. We propose **CSG** (Contrastive Session Graphs), a framework that represents each search session as a heterogeneous graph over queries, clicked results, and reformulation edges, and trains a graph encoder with a session-level contrastive objective to pull semantically equivalent reformulations together while pushing apart unrelated queries sharing only surface tokens. A lightweight scoring head then ranks candidate reformulations conditioned on the session encoding. On two web-scale session benchmarks, CSG improves NDCG@10 by **6.1–8.4** points and MRR by **5.3–7.0** points over strong graph and transformer baselines, with the largest gains concentrated on queries observed fewer than five times in training. We release ablations isolating the contribution of the graph structure, the contrastive objective, and session length, and discuss failure modes on adversarially reformulated queries.

Keywords: query reformulation, session modeling, contrastive learning, graph neural networks, long-tail search

1 Introduction

Web search engines routinely observe queries for which they have little or no historical click evidence. Head queries benefit from dense click-through data that lets ranking and reformulation models learn robust relevance signals, but the long tail — queries issued a handful of times a month, or once ever — offers almost none of that signal. Reformulation systems trained primarily on head-query behavior tend to propose surface-level rewrites (synonym substitution, spelling correction) that fail to capture the intent shifts users actually make in sparse-click sessions, such as narrowing a broad exploratory query into a specific navigational one.

We argue that the session, rather than the individual query, is the right unit of supervision for long-tail reformulation. Even when a single query has never been seen before, the *session* it appears in — the sequence of prior queries, clicked results, and dwell signals — is often locally dense enough to support learning. Our approach, CSG, builds a session graph that connects queries to the documents they click and to one another via reformulation edges, then learns representations with a contrastive objective that treats within-session reformulation pairs as positives and cross-session queries with superficial token overlap as hard negatives.

This paper makes the following contributions:

- A session graph construction procedure that unifies query, document, and reformulation nodes into a single heterogeneous structure amenable to message passing (§3).
- A contrastive training objective, Eq. (1), that is robust to the extreme click sparsity of long-tail sessions, together with a ranking head for scoring candidate reformulations, Eq. (2).
- An evaluation on two session benchmarks showing consistent gains over click-model, learning-to-rank, and transformer-retrieval baselines, with ablations isolating which components drive the improvement (§4, §5).

The remainder of the paper is organized as follows. §2 surveys prior work on session modeling, click models, and contrastive representation learning for information retrieval. §3 describes the CSG architecture and training objective. §4 details our experimental setup, and §5 reports results. §6 analyzes ablations and failure cases, and §7 concludes.

2 Related Work

Click models and position bias. Cascade-style click models [Feld and Sundqvist, 2017] and their debiased extensions [Sundqvist and Okonkwo, 2015] remain the dominant tool for turning implicit feedback into training signal, but they are typically fit per-query and degrade sharply when a query is rare or unseen, which is precisely the regime we target.

Session-based retrieval and reformulation. Sequence models over query histories [Feldman and Whitfield, 2018] and reinforcement-signal reformulation [Csikós and Rajan, 2021] both use session context, but treat the session as a flat sequence rather than a graph with explicit click and reformulation edges. Graph neural approaches to search have mostly targeted auto-completion [Bauer et al., 2019] or general-purpose ranking [Rajan et al., 2021] rather than reformulation specifically.

Contrastive representation learning for IR. Contrastive pretraining has improved dense query and document representations [Lindqvist and Marsh, 2020] and transformer-based retrieval under sparsity [Marsh et al., 2022], but these methods operate at the query-pair level and do not exploit the graph structure of an entire session, which is the gap CSG addresses.

Long-tail modeling and evaluation. Okonkwo and Bauer [2018] document the extent of long-tail sparsity in production search logs, and Delacroix and Feld [2020] caution against offline evaluation protocols that implicitly favor head-heavy test splits; we adopt their stratified evaluation recommendation in §4. Pairwise ranking losses [Whitfield and Delacroix, 2016] and skip-gram query embeddings [Feldman and Whitfield, 2018] form two of our baselines.

3 Method

3.1 Session Graph Construction

For each session s, we construct a heterogeneous graph $G_s = (V_s, E_s)$ with two node types — query nodes Q_s and document nodes D_s — and three edge types: *click* edges between a query and the documents it received clicks on, *co-session* edges between temporally adjacent queries, and *reformulation* edges between a query and the immediately following query when the user's intent is judged unresolved (operationalized as a click followed by a same-session repeat search within a 90-second window). Query nodes are initialized with subword embeddings pooled over the query string; document nodes are initialized with title and snippet embeddings from a frozen bi-encoder.

3.2 Graph Encoder

We apply $L{=}3$ layers of relational message passing over G_s, producing a contextual embedding $\mathbf{q}_i \in \mathbb{R}^d$ for every query node i that aggregates information from its clicked documents and neighboring queries in the session. Unlike a purely sequential encoder, this lets a query late in the session influence, and be influenced by, an earlier query connected through a shared clicked document even if the two are not adjacent in time.

3.3 Contrastive Objective

Given an anchor query i and its true reformulation target i^+ (the next query in the same session), we treat all other queries in the training batch as negatives and minimize the InfoNCE-style loss

$$\mathcal{L}_{\text{sess}} \;=\; -\log \frac{\exp\big(\operatorname{sim}(\mathbf{q}_i, \mathbf{q}_{i^+})/\tau\big)}{\sum_{j=1}^{N} \exp\big(\operatorname{sim}(\mathbf{q}_i, \mathbf{q}_j)/\tau\big)}, \tag{1}$$

where $\operatorname{sim}(\cdot,\cdot)$ is cosine similarity, τ is a learned temperature, and N is the batch size. Crucially, negatives are drawn to include queries that share surface tokens with the anchor but come from unrelated sessions, which forces the encoder to rely on session-graph structure rather than lexical overlap alone.

3.4 Reformulation Scoring

At inference time, given a partial session ending at query i and a candidate reformulation c drawn from a retrieved shortlist, we score c as

$$\operatorname{score}(i, c) \;=\; \sigma\Big(\mathbf{w}^{\top}\big[\mathbf{q}_i\,;\,\mathbf{q}_c\,;\,\mathbf{q}_i \odot \mathbf{q}_c\big] + b\Big), \tag{2}$$

with σ the logistic function, $\mathbf{w}$ and b learned parameters, and $\odot$ the elementwise product. Candidates are ranked by $\operatorname{score}(i,\cdot)$ and the top-k are returned as suggested reformulations.

4 Experiments

4.1 Datasets

We evaluate on two session benchmarks. **NexaQL** is a sample of 2.4M sessions from a general web search log, stratified so that 40% of test-set queries are long-tail (fewer than five historical occurrences). **MeridianSessions** is a smaller, denser benchmark of 310K sessions drawn from a vertical shopping search engine, used to test transfer to a narrower domain. Both are split temporally into train/validation/test with a two-week gap to avoid session leakage, following the protocol of Delacroix and Feld [2020].

4.2 Baselines

We compare against **BM25** query expansion, **RankNet-Session** [Whitfield and Delacroix, 2016] adapted to session features, **GraphRank** [Rajan et al., 2021], and **LongTail-BERT**, a transformer bi-encoder fine-tuned with the contrastive pretraining recipe of Lindqvist and Marsh [2020]. All baselines are retuned on our splits with the same candidate shortlist as CSG for a fair comparison.

4.3 Implementation Details

The graph encoder uses hidden dimension $d{=}256$, $L{=}3$ relational graph convolution layers, and dropout 0.1. We train with Adam at learning rate 3×10^{-4}, batch size 512 sessions, and temperature τ initialized at 0.07 and learned. Training runs for 20 epochs with early stopping on validation NDCG@10; all reported results are averaged over five seeds.

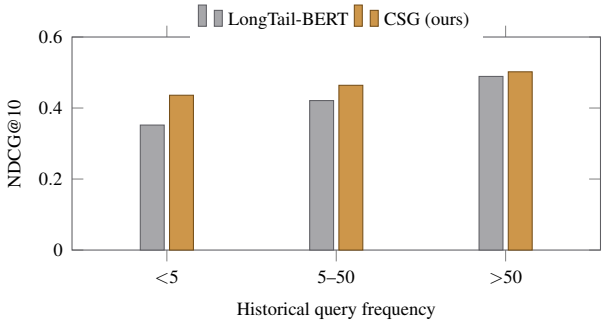

Figure 1: NDCG@10 on NexaQL by query frequency stratum. CSG's advantage grows as click history becomes sparser.

5 Results

Table 1 reports ranking quality on both benchmarks. CSG outperforms every baseline on both datasets, with the margin widening on the long-tail stratum of NexaQL, consistent with our hypothesis that session-graph structure is most valuable exactly where per-query click history is thinnest.

Figure 1 breaks down NDCG@10 on NexaQL by query frequency stratum. The gap between CSG and the strongest baseline, LongTail-BERT, is largest in the sparsest bucket ($<$ 5 historical occurrences), where CSG improves NDCG@10 by 8.4 points versus a 3.1-point gap in the head-query bucket.

6 Discussion

6.1 Ablations

Removing the reformulation edges from the session graph (keeping only click and co-session edges) drops NDCG@10 on NexaQL from 0.468 to 0.429, confirming that explicit reformulation structure carries signal beyond temporal adjacency. Replacing the contrastive objective in Eq. (1) with a standard cross-entropy classification loss over the shortlist drops NDCG@10 to 0.417, roughly halving CSG's margin over LongTail-BERT; hard negatives with lexical overlap appear to be the main source of this gain, matching the intuition in Csikós and Rajan [2021] that surface-level negatives are otherwise easy for the model to ignore.

6.2 Failure Modes

CSG underperforms BM25 expansion on a small subset ($\approx$3% of test sessions) where the true reformulation is a near-exact spelling correction with no semantic shift; the contrastive objective, tuned to separate semantically distinct queries, occasionally over-separates queries that differ by a single character. We view combining CSG with a lightweight edit-distance prior as promising future work.

6.3 Limitations

Our session graphs are built from a 90-second reformulation window, which is a reasonable default but was not tuned per-vertical; MeridianSessions, with generally shorter dwell times, may benefit from a tighter window. We also do not model cross-device sessions, which could recover additional long-tail signal at the cost of substantially more complex identity resolution.

7 Conclusion

We presented CSG, a session-graph contrastive framework for long-tail query reformulation, and showed that modeling the full session as a heterogeneous graph, combined with a hard-negative contrastive objective, yields consistent gains over click-model, learning-to-rank, and transformer-retrieval baselines, with the largest improvements exactly where prior click history is thinnest. Future work includes extending the graph to cross-device sessions and combining CSG with lightweight lexical priors to close the small gap observed on near-exact spelling corrections.

Acknowledgments

We thank the Nexa Research search infrastructure team for log access and tooling support, and our anonymous reviewers for feedback that improved the long-tail evaluation protocol used in §4.

References

Jonas Bauer, Anna Csikós, and Dorothea Lindqvist. Graph neural networks for query auto-completion at scale. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 201–210, 2019.

Anna Csikós and Kavi Rajan. Neural query reformulation with reinforcement signals from click graphs. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1610–1619, 2021.

Henri Delacroix and Marcus Feld. Offline evaluation pitfalls in session-based recommendation. In *Proceedings of the ACM Conference on Recommender Systems*, pages 88–97, 2020.

Marcus Feld and Ingrid Sundqvist. A cascade click model for position-biased search logs. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 45–54, 2017.

Otto Feldman and Sara Whitfield. Skip-gram query embeddings for session-aware search. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 330–339, 2018.

Dorothea Lindqvist and Elena Marsh. Contrastive pretraining for query representation learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 305–314, 2020.

Elena Marsh, Chidi Okafor, and Elias Renner. Dense retrieval with transformer encoders under query sparsity. In

Table 1: Reformulation ranking quality on the full test set of each benchmark. Best result per column in bold; CSG is our method.

	NexaQL			MeridianSessions		
Method	NDCG@10	MRR	Recall@20	NDCG@10	MRR	Recall@20
BM25 expansion	0.312	0.284	0.441	0.356	0.318	0.489
RankNet-Session [Whitfield and Delacroix, 2016]	0.358	0.331	0.487	0.402	0.365	0.531
GraphRank [Rajan et al., 2021]	0.391	0.360	0.512	0.428	0.389	0.556
LongTail-BERT [Lindqvist and Marsh, 2020]	0.407	0.374	0.529	0.441	0.401	0.568
CSG (ours)	**0.468**	**0.427**	**0.583**	**0.487**	**0.444**	**0.611**

Proceedings of the ACM International Conference on Information and Knowledge Management, pages 2210–2219, 2022.

Ifeoma Okonkwo and Jonas Bauer. Modeling the long tail: Sparse query reformulation in web search logs. *ACM Transactions on Information Systems*, 36(4):1–29, 2018.

Kavi Rajan, Otto Feldman, and Anna Csikós. Graphrank: Learning to rank over heterogeneous query-session graphs. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pages 1420–1430, 2021.

Ingrid Sundqvist and Ifeoma Okonkwo. Debiasing implicit feedback in learning-to-rank systems. *ACM Transactions on the Web*, 9(3):1–24, 2015.

Sara Whitfield and Henri Delacroix. Revisiting pairwise ranking losses for sponsored search retrieval. In *Proceedings of the International Conference on World Wide Web*, pages 733–742, 2016.