Planes, Trains and Automobiles: Leverage Multimodal In-Mission Signals for Shopping Journeys

Viet Ha-Thuc
Amazon Inc.
Palo Alto, CA, USA
vietth@amazon.com

Shasha Li
Amazon Inc.
Palo Alto, CA, USA
shashli@amazon.com

Arnau Ramisa
Amazon Inc.
Palo Alto, CA, USA
aramisay@amazon.com

Xinliang Zhu
Amazon Inc.
Palo Alto, CA, USA
xzhu@amazon.com

ABSTRACT

Modern search systems offer multiple ways for expressing information needs, including image, voice, and text. Consequently, an increasing number of users seamlessly transition between these modalities to convey their intents. This emerging trend presents new opportunities for utilizing queries in different modalities to help users complete their search journeys efficiently. In this proposal, we introduce an approach to segmenting a multimodal query stream into missions, demonstrate how these in-mission queries can enhance search ranking, and outline key areas for future research.

ACM Reference Format:
Viet Ha-Thuc, Shasha Li, Arnau Ramisa, and Xinliang Zhu. 2024. Planes, Trains and Automobiles: Leverage Multimodal In-Mission Signals for Shopping Journeys. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3627673.3679067

1 INTRODUCTION

Many search systems, from web search to commercial search, now provide multiple ways for users to articulate their information needs, including image, voice, and text queries. An increasing number of users interchangeably rely on these services, seamlessly switching between them to complete their tasks. For example, in commercial search, users might start with a broad keyword query like "Calvin Klein dress size 10" and browse the results to find a dress that suits their style. They can then use the image of that dress as input for a visual search to find similar items. As more users engage in interactive and multimodal searches, it becomes crucial to leverage signals across text, image, and voice queries within the same mission to help users complete their shopping journeys.

In this proposal, we first address the challenges of segmenting a multimodal query stream at an industrial scale. Next, we introduce a ranking approach that combines multimodal in-mission signals. Our experiments with Amazon Visual Search demonstrate that incorporating these signals significantly enhances result quality. Finally, we explore various future research directions on leveraging in-mission signals to further enhance the search experience.

2 SEGMENT QUERIES ACROSS MODALITIES

There has been prior work on segmenting a user's query stream into sequences of related queries. For example, He et al. [4] utilized various timeouts and overlapping terms between consecutive queries to segment query streams. Boldi et al. [2] proposed a method that builds a query-flow graph and uses an ML approach to identify query chains. However, these previous works primarily dealt with keyword queries. Additionally, some of these methods may not meet the scalability and latency requirements for real-time applications due to their complexity. In this proposal, we focus on a specific use case: given a visual query (an image) and a stream of text queries, determine which text queries belong to the same mission as the visual query. However, the approach presented here can be extended to accommodate other combinations of query modalities.

Inspired by He et al. [4], we also consider the similarity between queries. However, since our queries span different modalities, we leverage recent advances in large vision-language models (LVLM) [6, 7, 9]. Specifically, we train an LVLM using an architecture similar to Align before Fuse [7], including a text encoder, an image encoder, and a fused multimodal encoder. The positive training data consists of tuples of visual query, text query, clicked product image, clicked product text, where visual and text queries might be from different sessions but lead to the same product. Using clicked product data allows pairing numerous visual and text queries for LVLM training. With the model, both text and visual queries are mapped into a common embedding space to compute a cosine similarity.

We train a logistic regression model that combines time difference and semantic similarity. For high-precision ground truth, we sample about 1,000 visual-text query pairs issued by the same users within 30 minutes and manually label if they belong to the same shopping mission. Using time difference alone yields acceptable performance (Table 1). However, combining both features significantly enhances performance, with Average Precision improving from 0.61 to 0.81 (+33%) and ROC AUC increasing from 0.66 to 0.78 (+18%). Note that our approach enables parallel classification of query pairs, avoiding scanning over the query stream as in the clustering or graph-based methods, thus keeping latency low.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10.
https://doi.org/10.1145/3627673.3679067

Table 1: Multimodal query segmentation performances

	Average Precision	ROC AUC
Baseline	0.61	0.66
Logistic Regression	0.81	0.78

Table 2: Relative improvements of multimodal rankers over the baseline. All of the improvements are statistically significant by paired t-tests.

	MAP	NDCG@3	NDCG@7
Explicit in-mission signals	4.4%	6.4%	5.4%
Explicit and implicit signals	5.4%	8.0%	6.5%

3 MULTI-MODAL SIGNALS FOR VISUAL SEARCH RANKING

Given a visual search query, such as a dress image, previous text queries from the same mission, like "Calvin Klein dress size 10," provide useful context on the searcher's intent and preferences. These text queries can be leveraged to generate additional ranking features to improve visual search quality. While visual match features between the query and a product strongly predict how well a result captures the searcher's desired style, lexical matches can indicate if the result meets specific requirements, such as brand name, size, or model number. Thus, lexical and visual matches often complement each other. Additionally, we compute the similarity between the fused multimodal embedding described earlier and the product embedding.

For our ranking experiment, we collect visual queries from the Amazon Lens service preceded by at least one text query within 30 minutes. Labels for the results (products) are inferred from user actions, as in many industrial systems [3]. If the text queries are classified as in-mission by the segmentation model, they would be used to generate the ranking features. Additionally, many visual queries also contain textual content, so we apply an OCR recognizer to extract these implicit text signals. Because raw OCR content can be noisy (e.g., "Directions" or "Copyright"), we filter it using a query tagging model [8] to retain only relevant tags, such as brand names, models, and colors. These filtered texts are then matched with the text fields on the products.

The baseline ranker uses visual match features, product quality, and popularity features. As shown in Table 2, incorporating features from previous in-mission queries significantly improves ranking quality: +4.4% on MAP, +6.4% on NDCG@3, and +5.4% on NDCG@7. Adding implicit text-match features further increases these metrics to 5.4%, 8% and 6.5%, respectively. The strong empirical improvements confirm the benefits of using multimodal signals in visual search ranking.

4 FUTURE DIRECTIONS

Interactive searching across multiple modalities is still in its early stages and opens up numerous intriguing opportunities. Besides ranking, multimodal in-mission signals can be leveraged throughout various stages of an IR system. For example, previous text queries have been vital for query auto-completion, particularly when the typed prefix is short or ambiguous [5]. Extending this approach from text-only to multimodal inputs would significantly enrich the contextual information. Similarly, during the retrieval phase, a fused multimodal embedding combining current and previous queries can be used to perform a kNN search.

Additionally, instead of using the queries, another direction is to utilize user actions from previous in-mission queries. Leveraging past clicks within the same query session, a form of relevance feedback [10], has been shown to be effective [1]. By extending this to actions across different query sessions and modalities, we can increase signal density and enhance overall search effectiveness.

Besides *previous* in-mission queries, queries issued *after* the current search can be leveraged to augment log-based training data. Typically, abandoned queries lack positive results. However, actions from subsequent in-mission queries can infer relevant results for abandoned queries. Since query segmentation in this context runs offline, we can utilize more sophisticated features not feasible in real-time processing. One useful feature is the similarity of the result sets: if two queries, even across different modalities, belong to the same mission, their result sets will be similar.

Another promising research direction is to learn a fused embedding from a multimodal in-mission query stream that performs well both when queries complement each other and when the user's intent evolves over the course of their journey. Existing LVLMs like CLIP [9], ALBEF [7], and BLIP [6] are effective for the former but not the latter, as they treat input pairs symmetrically. It's important to note that intent drift—such as transitioning from an image of Adidas running shoes to a follow-up voice or text query "Nike"—is different from the noise in image-text pairs used to train current LVLMs. For interactive multimodal search, the model needs to account for the temporal order of queries, distinguish between complementary query aspects and evolving aspects, and recognize which source to prioritize when the user's intent changes.

REFERENCES

[1] Keping Bi, Choon Hui Teo, Yesh Dattatreya, Vijai Mohan, and W. Bruce Croft. 2019. Leverage Implicit Feedback for Context-aware Product Search. In *Proceedings of the SIGIR 2019 Workshop on eCommerce*.

[2] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. 2008. The query-flow graph: model and applications. In *Proceedings of the 17th ACM CIKM*. 609–618.

[3] Viet Ha-Thuc and Shakti Sinha. 2016. Learning to Rank Personalized Search Results in Professional Networks. In *Proceedings of the 39th ACM SIGIR*.

[4] Daqing He, Ayse Göker, and David J. Harper. 2002. Combining evidence for automatic Web session identification. *Inf. Process. Manag.* 38, 5 (2002).

[5] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In *The 37th International ACM SIGIR*. ACM, 445–454.

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. 12888–12900.

[7] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[8] Mehdi Manshadi and Xiao Li. 2009. Semantic Tagging of Web Search Queries. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics ACL*. 861–869.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. 8748–8763.

[10] J. J. Rocchio. 1971. Relevance feedback in information retrieval. In *The Smart retrieval system - experiments in automatic document processing*. 313–323.