MMLF: Multi-query Multi-passage Late Fusion Retrieval

**Yuan-Ching Kuo , Yi Yu, Chih-Ming Chen , Chuan-Ju Wang**

Academia Sinica, 2The Ohio State University

---

**Abstract**

MMLF là một phương pháp truy hồi thông tin mới, khai thác mô hình ngôn ngữ lớn (LLM) để tạo các truy vấn phụ (sub-queries), mở rộng thành các đoạn văn (passages), truy xuất độc lập, sau đó hợp nhất kết quả bằng kỹ thuật Reciprocal Rank Fusion (RRF). Phương pháp này cải thiện rõ rệt độ chính xác trên 5 tập dữ liệu chuẩn BEIR.

**Introduction**

1. Bài toán truy hồi thông tin vẫn gặp khó khăn khi truy vấn ngắn hoặc mơ hồ.
2. Truy hồi truyền thống (như BM25) dựa vào từ vựng, không hiệu quả khi không có sự trùng lặp từ.
3. Mô hình dense retrieval và LLM giúp cải thiện bằng cách biểu diễn ngữ nghĩa truy vấn.
4. Các phương pháp hiện tại như Query2Doc, Chain-of-Thought, MILL tuy hiệu quả nhưng vẫn có hạn chế, đặc biệt là trong cách xử lý kết quả truy hồi.
5. MMLF được đề xuất như một pipeline hai giai đoạn nhằm tối ưu hóa quá trình phân tích truy vấn và hợp nhất kết quả.

**Pipeline Đề Xuất: MMLF**

Gồm 3 bước chính:

1. Tạo truy vấn phụ (Multi-query Generation)

Sử dụng LLM và prompt MQR để tạo ra 3 truy vấn phụ phản ánh nhiều khía cạnh của truy vấn gốc.

1. Mở rộng truy vấn thành đoạn văn (Query-to-passage Expansion)

Mỗi truy vấn phụ được LLM mở rộng thành đoạn văn giàu ngữ cảnh (pseudo-document) bằng prompt CQE.

1. Hợp nhất danh sách xếp hạng (Ranked List Fusion)

Truy xuất tài liệu từ mỗi đoạn văn (và truy vấn gốc), sau đó hợp nhất kết quả bằng RRF thay vì nối chuỗi như phương pháp MILL.

**Experiments**

1. Dữ liệu và đánh giá

Sử dụng 5 tập dữ liệu BEIR quy mô nhỏ: DBPEDIA, FIQA-2018, NFCORPUS, TREC-COVID, TOUCHE-2020.

Đánh giá bằng Recall@1k và nDCG@10.

1. Thiết lập thực nghiệm

Dùng LLaMA-3-70B-Instruct và bộ mã hóa e5-small-v2.

Tạo 3 truy vấn phụ cố định.

1. So sánh với các phương pháp khác

So sánh với 5 phương pháp: Raw Query, Query2Doc, CoT, LC-MQR (với RRF), MILL.

MMLF vượt trội nhất, tăng trung bình 4% Recall@1k so với MILL.

1. Kết quả chính

MMLF có kết quả tốt nhất ở cả Recall@1k và nDCG@10 trên tất cả 5 tập dữ liệu.

1. Nghiên cứu tách biệt (Ablation Study)

5.1 So sánh kỹ thuật hợp nhất:

RRF vượt trội hơn CombSUM và concat trong Recall@1k, nhưng CombSUM tốt hơn về nDCG@10.

5.2 Vai trò của truy vấn gốc:

Kết hợp truy vấn gốc với các đoạn văn cải thiện kết quả rõ rệt. “RRF w/ q included” đạt Recall@1k cao nhất.

5.3 So sánh pipeline tạo truy vấn:

MQ2MP (hai bước: tạo sub-query rồi mở rộng) vượt trội hơn các phương án MQ (chỉ tạo sub-query) và MP (chỉ tạo passage).