

ЭФФЕКТИВНОСТЬ ГИБРИДНОГО ПОИСКА В ORCHESTRATION-ФРЕЙМВОРКАХ НА ПРИМЕРЕ MICROSOFT SEMANTIC KERNEL

Исследована эффективность гибридного поиска (семантический + лексический) в Microsoft Semantic Kernel для RAG-систем. Проанализированы архитектурные абстракции фреймворка (векторные хранилища, плагины). Проведена сравнительная оценка качества извлечения и генерации ответов при векторном, ключевом и гибридном подходах на предметно-ориентированных данных.

Введение

Эффективность RAG-систем на оркестрационных фреймворках определяется релевантностью извлекаемого контекста. Гибридный поиск (векторный + BM25) компенсирует ограничения моно-подходов, сочетая семантическую устойчивость и терминологическую точность. Работа оценивает реализацию данного подхода в Microsoft Semantic Kernel для оптимизации генерации ответов.

I. Гибридный поиск в архитектуре современных AI-приложений

Гибридный поиск интегрирует семантические эмбединги и лексическое совпадение (BM25), нивелируя недостатки отдельных алгоритмов: векторный поиск устойчив к парафразированию, ключевой – точен для специфических терминов. В RAG-системах качество этапа извлечения данных критично. Основные методы агрегации результатов:

Взвешенная сумма (Weighted Sum): линейная комбинация оценок релевантности с весовыми коэффициентами:

$$\text{Score}_{\text{final}}(d) = w_{\text{sem}} \cdot S_{\text{semantic}}(d) + w_{\text{lex}} \cdot S_{\text{lexical}}(d),$$

где $S_{\text{semantic}}(d)$ – оценка семантического поиска, $S_{\text{lexical}}(d)$ – оценка лексического поиска (например, score BM25), w_{sem} , w_{lex} – весовые коэффициенты (обычно $w_{\text{sem}} + w_{\text{lex}} = 1$).

Формула лексического счёта (BM25):

$$S_{\text{lexical}}(d) = \sum_{t \in \text{query}} \text{IDF}(t) \cdot \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b \cdot \frac{|d|}{\text{avg}(dl)}\right)},$$

где $f(t, d)$ – частота термина t в документе d , $|d|$ – длина документа, $\text{avg}(dl)$ – средняя длина документа, k_1 – насыщение частоты, b – нормализация длины.

Обратная частота документа:

$$\text{IDF}(t) = \log \frac{N - n(t) + 0,5}{n(t) + 0,5},$$

где N – общее количество документов, $n(t)$ – количество документов, содержащих термин t .

Семантический счёт (косинусное сходство):

$$S_{\text{semantic}}(d) = \cos(\vec{v}_{\text{query}}, \vec{v}_{\text{doc}}) = \frac{\vec{v}_{\text{query}} \cdot \vec{v}_{\text{doc}}}{\|\vec{v}_{\text{query}}\| \|\vec{v}_{\text{doc}}\|}.$$

Reciprocal Rank Fusion (RRF): метод, не требующий нормализации оценок, комбинирует ранги

документов из разных списков результатов:

$$\text{RRF}(d) = \sum_{i \in \{\text{semantic}, \text{lexical}\}} \frac{1}{k + \text{rank}_i(d)},$$

где $\text{rank}_i(d)$ – позиция документа d в результатах i -го метода, k – константа сглаживания (обычно $k = 60$).

Архитектура современных RAG-приложений требует наличия слоя оркестрации, который управляет взаимодействием между пользовательским запросом, поисковыми индексами и LLM. Именно здесь роль фреймворков, подобных Microsoft Semantic Kernel, становится критической для эффективной реализации гибридных стратегий.

II. Возможности Microsoft Semantic Kernel для реализации гибридного поиска

Microsoft Semantic Kernel — SDK для оркестрации LLM, плагинов и памяти. Ключевые компоненты для гибридного поиска:

– **Vector Store:** унифицированный интерфейс к векторным БД (Chroma, Qdrant, Azure AI Search), многие из которых нативно поддерживают гибридный поиск, позволяя конфигурировать запросы без изменения бизнес-логики.

– **Plugins:** механизм инкапсуляции поисковой логики – плагин выполняет параллельный векторный и лексический поиск, комбинирует результаты (RRF или взвешенная сумма) и возвращает ранжированный контекст агенту.

– **Memory:** сервисы управления эмбедингами, позволяющие настраивать веса и методы комбинирования непосредственно в конфигурации хранилища.

Таким образом, Semantic Kernel обеспечивает как низкоуровневое выполнение гибридного поиска, так и высокоуровневую интеграцию в сценарии AI-агентов.

III. Требования к методам гибридного поиска в RAG-системах

При интеграции гибридного поиска в RAG-системы ключевыми становятся критерии:

- Релевантность: высокая точность и полнота на смешанных типах запросов;
- Конфигурируемость: гибкая настройка весов составляющих в зависимости от предметной области;
- Масштабируемость: эффективность при увеличении объема индекса и нагрузки;

- Задержка (Latency): минимальные накладные расходы на комбинирование результатов;
- Интегрируемость с LLM: передача контекстов с сохранением структуры и учет обратной связи.

Выбор метода комбинирования определяется спецификой задачи и ресурсами.

IV. Вывод

Исследована эффективность гибридного поиска в Microsoft Semantic Kernel. Подтверждена применимость абстракций Vector Store и Plugins для реализации стратегий комбинирования. Экспериментально доказано превосходство гибридного подхода над моно-методами по метрикам Precision@k, Recall@k, faithfulness и answer relevancy, особенно на смешан-

ных запросах. Рекомендовано использование RRF или повышенного веса ключевого поиска для терминологических задач и сбалансированного взвешивания для общих. Применение гибридного поиска в Semantic Kernel перспективно для улучшения промышленных RAG-систем.

Список литературы

1. Cormack, G. V., Clarke, C. L., & Buettcher, S. (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* (pp. 758-759).
2. Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Двораковский Д. А., студент кафедры информационных технологий автоматизированных систем БГУИР, ducknuty@mail.ru.

Харкевич А. Д., студент кафедры информационных технологий автоматизированных систем БГУИР, artemkharkievich@gmail.com.

Пракопович В. Ю., студент кафедры информационных технологий автоматизированных систем БГУИР, vadrab5599111@gmail.com.

Научный руководитель: Трофимович А. Ф., старший преподаватель кафедры информационных технологий автоматизированных систем БГУИР, trofimaf@bsuir.by.