**УДК 004.89**

# АНАЛИЗ ПОИСКОВЫХ АЛГОРИТМОВ СО СЖАТИЕМ БИБЛИОТЕКИ FAISS

**Маслюхин С.М.** (Университет ИТМО, ООО «ЦРТ-инновации»)
**Научный руководитель – д.т.н., Матвеев Ю.Н.** (Университет ИТМО, ООО «ЦРТ-инновации»)

Применение нейросетевых моделей с поиском ответа в базе кандидатов в диалоговых системах связано с поиском по огромной базе кандидатов. Поисковые алгоритмы библиотеки FAISS обеспечивают эффективный поиск в условиях ограниченного количества оперативной памяти. Однако выбор подходящего алгоритма и его параметров является нетривиальной задачей.

Нейросетевые модели с поиском ответа в базе кандидатов часто используются в диалоговых системах, как самостоятельно, так и вместе с генеративными моделями. Для эффективной работы таких моделей требуется поиск по огромной базе кандидатов, состоящей из сотен миллионов текстовых фрагментов, что требует больших вычислительных мощностей и объёмов оперативной памяти. Библиотека FAISS содержит большое количество алгоритмов поиска, как с сжатием, так и без. Применение алгоритмов без сжатия обеспечивает наиболее точный выбор ответа, однако часто не может быть применено в условиях технических ограничений. Алгоритмы со сжатием позволяют адаптироваться под технические ограничения, ключевым из которых является оперативная память, но ведут к снижению точности поиска. Выбор оптимального алгоритма и его параметров, удовлетворяющих техническим требованиям и обеспечивающих наибольшую точность, является нетривиальной задачей. Существующие исследования, посвящённые этой проблеме, рассматривают частные решения применимые в определённых условиях и не могут быть использованы для выбора алгоритма на новых моделях и данных с соблюдением технических ограничений.

Библиотека FAISS содержит различные алгоритмы поиска ближайших векторов. Предполагается, что текстовые фрагменты идентифицируются целыми числами, представлены в виде векторов, и их можно сравнивать через скалярное произведение. Векторы, похожие на вектор запроса, имеют наибольшее скалярное произведение с вектором запроса. Библиотека также поддерживает косинусную близость, поскольку это скалярное произведение нормализованных векторов. Некоторые из методов, используют исключительно сжатое представление векторов и не требуют сохранения оригинальных векторов. Обычно это происходит за счет менее точного поиска, однако эти методы могут масштабироваться до миллиардов векторов в оперативной памяти на одном сервере. Другие методы, такие как HNSW и NSG, добавляют структуру индексации поверх оригинальных векторов, чтобы сделать поиск более эффективным. Библиотека FAISS построена на основе индексов, которые хранят набор векторов и предоставляют функцию для поиска среди них ближайшего вектора с помощью скалярного произведения. Некоторые типы индексов являются простыми, например, индекс полных векторов. Больший интерес представляют составные индексы, объединяющие различные методы, комбинация которых позволяет получить наилучший результат. Большинство доступных индексов обеспечивают компромисс в отношении: времени поиска; точности поиска; памяти, используемой для хранения одного вектора; времени обучения индекса и времени добавления нового вектора в индекс. Теоретический и практический анализ различных составных индексов важен при выборе оптимального решения в соответствии с предъявляемыми техническими требованиями. Также при сравнении различных вариантов индекса следует оценивать корреляцию между теоретической оценкой и оценкой по результатам эксперимента, так как эту зависимость можно в дальнейшем использовать для быстрого выбора оптимального алгоритма и его параметров в новых условиях, что очень важно, так как условия использования модели могут меняться со временем.

Исследование позволяет выбрать оптимальный алгоритм поиска, обеспечивающий наибольшую точность поиска в соответствии с предъявляемыми техническими требованиями. При этом рекомендации на основе теоретического анализа обеспечивают быструю адаптацию поискового алгоритма под новые условия. В результате исследования достигается возможность эффективного применения нейросетевых моделей с поиском ответа в базе кандидатов в диалоговых системах.