0.1. *Шашок Н.А., Кожемякина Э.Д.* Разработка архитектуры системы векторного поиска с привязкой эмбеддингов к исходным документам для вопрос-ответной системы

Современные информационные системы должны обрабатывать и анализировать данные высокой сложности и многомерности, а также различные их форматы, включающие графические, звуковые и текстовые данные. Традиционные хранилища данных и алгоритмы поиска не предназначены для единого корпуса разнородных данных: без дополнительных разработок по внесению метаданных, описаний и ключевых слов, осуществление поиска данных разного формата, имеющих схожую смысловую нагрузку, представляется трудоемкой задачей. Одним из вариантов решения этой задачи является использование векторных баз данных, оптимизированных для эффективного и точного поиска в векторном пространстве и для хранения эмбеддингов – представлений объектов в виде числовых массивов, являющихся, в том числе, «смысловыми» проекциями объектов. Однако использование эмбеддингов без привязки к исходным данным и документам большого размера, что, как правило, не реализуется в большей части векторных хранилищ, не является полноценным решением задачи.

При разработке системы поиска данных для вопрос-ответной системы представляется необходимым использование гибридной модели хранения, так как векторные базы данных не предназначены для хранения данных большого объема в форматах, отличных от векторного. Такое хранение данных отвечает специфике создаваемых эмбеддингов как представлений отдельных слов или небольших частей данных в виде векторов, но не подходит для решения поставленной задачи. Также в задаче построения системы вопрос-ответ необходимо предоставлять пользователю возможность удостовериться в истинности ответа, либо получить дополнительные документы по интересующей тематике.

Таким образом, необходимо использовать векторное хранилище в связке с хранилищем другого типа, которое позволит обрабатывать большие объемы данных произвольного размера. Под такую постановку задачи подходят реляционные базы данных, хранилища, способные работать с форматом parquet, файловые хранилища, а также иные хранилища, либо их набор.

Для реализации описанного решения представляется целесообразным определить некоторый программный модуль-адаптер в качестве прослойки между векторной базой данных и иным хранилищем. Такой программный модуль должен реализовывать взаимосвязь между векторными данными и данными в соответствующем программному модулю хранилищу по запросу на вывод исходных данных, связанных с некоторым вектором или их набором. Также представляется необходимым разработать некоторый более общий программный модуль, позволяющий агрегировать связи между векторным хранилищем и иными хранилищами, для каждого из которых предоставлен аналогичный описанному выше программный модуль-адаптер с идентичным программным интерфейсом, и возвращать исходные данные, к которым привязаны эмбеддинги, вне зависимости от хранилища, в котором эти данные находятся. Такая архитектура предполагает привязку к набору векторов в векторном хранилище некоторых метаданных, которые могли бы однозначно определить, из какого хранилища происходят изначальные данные.

Описанный подход позволяет хранить исходные мультимодальные данные, использовать их в задачах обработки естественного языка, а также решать множество архитектурных проблем, связанных с расширяемостью, поскольку модуль поиска, векторная БД, адаптеры к хранилищам данных других форматов и сами эти хранилища данных могут поставляться и расширяться независимо друг от друга, что потенциально повышает гибкость системы.

Научный руководитель — д.т.н., к.филол.н. Гавенко О. Ю.