CS59200-VecDB: Databases for AI: Vector Databases

Instructor

- Jianguo Wang
- Email: csjgwang@purdue.edu

Course Description

Vector databases have recently emerged as a hot topic in the broader realm of Databases for AI. The surge of interest is largely fueled by large language models (LLMs), where vector databases help overcome inherent limitations such as hallucinations, lack of domain expertise, and the inability to incorporate real-time information. This is enabled by the new paradigm of Retrieval-Augmented Generation (RAG), in which vector databases act as external knowledge bases, delivering relevant context to LLMs via vector search. While vector search itself is not new, modern vector databases face a host of system-level challenges, which we will explore in depth in this course.

In this seminar, we will cover the foundations of vector databases, covering vector indexing techniques such as quantization- and graph-based indexes, as well as memory- and disk-optimized indexes. We will discuss how to automatically tune these indexes and learn advanced vector search that integrates both vector and non-vector data. We will also cover the design trade-offs between specialized and integrated vector databases and highlight recent techniques from top-tier research papers and modern vector databases.

Format

This course is structured as a seminar with a semester-long project. Students will read and discuss recent papers from top-tier venues and conduct a research project on vector databases. Projects may be done individually or in teams of up to three, subject to the scope and instructor approval. A list of papers will be provided, from which students will select topics to study and present in class. The seminar will also feature guest speakers with extensive expertise in vector databases.

Prerequisites

No prior experience with vector databases is required. However, familiarity with data structures (e.g., CS251), databases (e.g., CS348 or CS448), and introductory AI/ML (e.g., CS242 or CS243) will be helpful.

Credits

Number of credits is 3.

Grading Criteria

- Paper Presentation: 10%
- Midterm exam: 25%
- Final project: 40%
- Final exam: 25%

Plan of Study

- Intend to be included as both PhD and MS plan of study.