

BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(HƯỚNG NGHIÊN CỨU)

Tên đề tài : Xây dựng kho dữ liệu đồ thị dựa trên nền tảng dữ liệu lớn cho đề xuất khóa học dựa trên mục tiêu nghề nghiệp CNTT

Sinh viên thực hiện : 19127134 – Nguyễn Gia Hân

19127584 – Mạch Cảnh Toàn

Giảng viên hướng dẫn: TS. Nguyễn Trần Minh Thư

1. Chủ đề và ý tưởng nghiên cứu:

Trong bối cảnh dữ liệu lớn (big data) hiện nay, kho dữ liệu (data warehouse) truyền thống được xây dựng dựa trên dữ liệu quan hệ không còn phù hợp để khai thác cho nhiều dạng dữ liệu khác nhau (có cấu trúc, phi cấu trúc, bán cấu trúc). Đề tài tập trung nghiên cứu xây dựng kho dữ liệu dựa trên cấu trúc đồ thị, cùng với việc xây dựng các kho dữ liệu chức năng (data mart) nhằm phục vụ cho việc lưu trữ, phân tích dữ liệu cho bài toán tư vấn khóa học dựa trên mục tiêu nghề nghiệp trong lĩnh vực CNTT. Đề tài cũng tiến hành xây dựng thực nghiệm, so sánh hiệu suất phân tích dữ liệu trên kho dữ liệu đồ thị và kho dữ liệu quan hệ truyền thống, trên bộ dữ liệu thử nghiệm về Khóa học, nghề nghiệp đã được xây dựng trước đó.

2. Phương pháp nghiên cứu:

- Tiến hành nghiên cứu các bài báo khoa học trong lĩnh vực liên quan, đặc biệt là các bài báo về kỹ thuật xây dựng kho dữ liệu đồ thị, cách ánh xạ kho dữ liệu quan hệ truyền thống sang kho dữ liệu đồ thị, nghiên cứu các độ đo, kỹ thuật thử nghiệm.
- Khảo sát một số trang tuyển dụng trực tuyến, khóa học MOOC, thống kê nhu cầu nghề nghiệp, ... từ đó phân tích yêu cầu của bài toán phân tích dữ liệu tư vấn kỹ năng, khóa học, các loại thống kê quan trọng trong bài toán tư vấn lộ trình học tập cũng như các yêu cầu phân tích dữ liệu liên quan.
- Xây dựng kiến trúc kho dữ liệu đồ thị, các khối dữ liệu chức năng đồ thị đáp ứng yêu cầu của bài toán.

- Xây dựng đường ống dữ liệu, để thực hiện đồ dữ liệu thô, xử lý, và tải dữ liệu và kho dữ liệu đồ thị, cũng như tải dữ liệu đến các kho dữ liệu chức năng.
- Xây dựng bổ sung kho dữ liệu truyền thống trên dữ liệu quan hệ đi kèm, nhằm thực hiện thử nghiệm tốc độ thời gian truy xuất dữ liệu trên kho dữ liệu đồ thị và kho dữ liệu truyền thống. Đặc biệt đối với những truy vấn phân tích dữ liệu dạng phân cấp (có sự kết nối giữa các đối tượng dữ liệu theo độ phức tạp khác nhau)
- Nghiên cứu các nền tảng, công nghệ lập trình để thực hiện việc thử nghiệm bao gồm : nền tảng công nghệ xử lý dữ liệu lớn Spark, ngôn ngữ lập trình Python, CSDL đồ thị Neo4j và ngôn ngữ truy vấn đồ thị Cypher.
- Chạy thực nghiệm, đối sánh kết quả tốc độ thời gian truy xuất dữ liệu trên kho dữ liệu đồ thị và quan hệ, từ đó đưa ra nhận định cho ý nghĩa thử nghiệm.

3. Đóng góp Khoa học và thực tiễn:

- Cung cấp cách nhìn chung về phương pháp xây dựng, triển khai cho kho dữ liệu đồ thị.
- Xây dựng kiến trúc kho dữ liệu đồ thị, phục vụ thống kê, phân tích, tư vấn cho kỹ năng, nghề nghiệp, khóa học theo mục tiêu nghề nghiệp của lĩnh vực CNTT.
- Triển khai và thử nghiệm toàn bộ kết quả nghiên cứu trên nền tảng dữ liệu lớn thông qua các công nghệ Spark, Python, Neo4j, Cypher và một đường ống dữ liệu (data pipeline) từ các tập tin dữ liệu thô, xử lý dữ liệu (thực hiện NER qua các mô hình máy học) để trích xuất thực thể và đưa vào kho dữ liệu đồ thị.
- Cung cấp kết quả thực nghiệm truy xuất dữ liệu trên kho dữ liệu đồ thị và kho dữ liệu quan hệ truyền thống. Kết quả thực nghiệm cho thấy hiệu suất khai thác trên dữ liệu đồ thị tốt hơn dữ liệu quan hệ trong tốc độ đọc, đặc biệt cho các truy vấn dạng phân cấp phức tạp.

4. Quá trình thực hiện:

- Sinh viên nghiêm túc, có trách nhiệm trong quá trình thực hiện đề tài.
- Biết lắng nghe, tiếp thu ý kiến, và điều chỉnh, giải quyết tốt những vấn đề được GVHD nêu ra.
- Chủ động trong việc nghiên cứu, thực hiện, giải quyết công việc.

5. Báo cáo viết:

- Báo cáo viết tốt, trình bày nội dung rõ ràng, đầy đủ cho toàn bộ mục tiêu đề tài đặt ra.

6. Trình bày trước hội đồng:

- Trình bày đầy đủ nội dung, kết quả nghiên cứu rõ ràng, mạch lạc.

7. Công bố khoa học/ ứng dụng thực tế: chưa có công bố

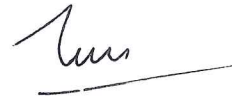
- Chưa có công bố khoa học, tuy nhiên với kết quả đạt được của đề tài có thể viết bài báo khoa học để công bố và sinh viên đang tiến hành công việc sau bảo vệ đề tài.

Đánh giá xếp loại: Xuất sắc

TP.HCM, ngày 05 tháng 04 năm 2023

Giảng viên hướng dẫn

(Ký và ghi rõ họ tên)



Nguyễn Trần Minh Thư