fit@hcmus

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

KHAI THÁC LUẬT KẾT HỢP DỰA TRÊN MẪU ĐỒ THỊ CON PHỔ BIẾN

(Mining Association Rules based on Frequent Subgraph Patterns)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– GS. TS. Lê Hoài Bắc (Khoa Công nghệ Thông tin)

[Nhóm] Sinh viên thực hiện:

1. Nguyễn Thị Tình (MSSV: 1612703)
2. Nguyễn Thanh Tuấn (MSSV: 1612774)

Loại đề tài: Nghiên cứu.

Thời gian thực hiện: Từ *tháng 01/ năm 2020* đến *tháng 07/ năm 2020*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Trong thời đại hiện nay, cùng với sự phát triển mạnh mẽ của Internet cũng như Công nghệ thông tin, bài toán Khai thác Dữ liệu (Data Mining) với mục đích tìm kiếm khối thông tin có ích, tiềm ẩn và mang tính dự đoán trong khối cơ sở dữ liệu

lớn đang ngày càng phổ biến. Đặc biệt, bài toán Khai thác Tập phổ biến và Luật kết hợp được chú trọng hơn cả với mục đích tìm các thành phần hay xuất hiện cùng nhau và khai thác mối quan hệ giữa chúng. Bài toán này được áp dụng rộng rãi trong nhiều lĩnh vực: Thương mại; Y học; Phát hiện gian lận, đạo văn; Công cụ tìm kiếm; Hệ thống Đề xuất; ...

Đã có nhiều phương pháp được đề xuất để giải quyết các bài toán trên, tuy nhiên chủ yếu là ở cơ sở dữ liệu dạng giao dịch (transaction). Vì vậy, trong khóa luận này, chúng tôi tiến hành nghiên cứu và xây dựng thuật toán minh họa để khai thác Tập phổ biến và sinh Luật Kết hợp trên dữ liệu dạng đồ thị.

2.2 Mục tiêu đề tài

- Giải quyết bài toán tìm tập phổ biến trên đồ thị.
- Giải quyết bài toán tìm luật kết hợp trên đồ thị.

2.3 Phạm vi của đề tài

Khóa luận bao gồm triển khai hai thuật toán chính là Khai thác Mẫu đồ thị con phổ biến FPMiner và Sinh Luật kết hợp mẫu đồ thị ruleGen với dữ liệu thực nghiệm dạng đồ thị. Dữ liệu thực nghiệm gồm hai loại: dữ liệu thực tế có sẵn và dữ liệu do nhóm tự sinh ra.

2.4 Cách tiếp cận dự kiến

- Đầu tiên, thu thập dữ liệu đồ thị thực tế có sẵn và tiền xử lý dữ liệu.
- Tìm hiểu các cấu trúc dữ liệu đồ thị để tiến hành xây dựng cấu trúc dữ liệu phù hợp.
- Cài đặt thuật toán Khai thác Mẫu đồ thị con phổ biến dựa trên ý tưởng của bài báo [1] với dữ liệu tập trung, chiến lược kiểm tra Mininum DFS Code của thuật toán gSpan ([2]) và chiến lược tỉa nhánh dựa trên tính đơn điệu.
- Cài đặt thuật toán Sinh Luật Kết hợp Mẫu đồ thị.

- Sinh các tập dữ liệu dựa trên mẫu có sẵn để kiểm chứng kết quả của các thuật toán đã cài đặt.

2.5 Kết quả dự kiến của đề tài

- Cài đặt thành công thuật toán Khai thác Mẫu đồ thị con phổ biến.
- Cài đặt thành công thuật toán Sinh luật Kết hợp Mẫu đồ thị.

2.6 Kế hoạch thực hiện

2.6.1 Các mốc thời gian dự kiến

Bắt đầu	Kết thúc	Nội dung thực hiện
01/01/2020	01/02/2020	Tìm hiểu tổng quan về các đề tài trong lĩnh vực Khai thác Dữ liệu và lựa chọn đề tài thích hợp.
01/02/2020	31/03/2020	Tìm hiểu tổng quan về các bài toán trong lĩnh vực Khai thác Dữ liệu trên đồ thị và lựa chọn bài toán.
01/04/2020	15/04/2020	Tìm hiểu, xây dựng các cấu trúc dữ liệu thích hợp để giải quyết bài toán.
16/04/2020	15/05/2020	Tìm hiểu, xây dựng thuật toán Khai thác Mẫu đồ thị phổ biến trên đồ thị có hướng.
16/05/2020	31/05/2020	Tìm hiểu, xây dựng thuật toán Sinh Luật kết hợp từ các tập Mẫu đồ thị phổ biến.
01/06/2020	30/06/2020	Tìm hiểu, tối ưu các thuật toán đã cài đặt.
15/06/2020	10/08/2020	Viết báo cáo khóa luận.
01/07/2020	10/08/2020	Tìm hiểu, triển khai các thuật toán Khai thác Mẫu đồ thị phổ biến và Luật kết hợp cho đồ thị vô hướng. Tiếp tục tối ưu thuật toán về không gian và thời gian.

2.6.2 Phân công công việc

STT	Người thực hiện	Nội dung công việc
1	Cả hai	Tìm hiểu về các đề tài trong lĩnh vực Khai thác dữ liệu và lựa chọn đề tài.
2		Tìm hiểu các bài toán trong lĩnh vực Khai thác Dữ liệu trên đồ thị và lựa chọn bài toán.
3		Xây dựng các cấu trúc dữ liệu cần thiết.
4		Viết báo cáo khóa luận.
5		Làm slides trình bày.
6	Nguyễn Thanh Tuấn	Thu thập dữ liệu thực tế có sẵn và tiền xử lý.
7		Tự tạo dữ liệu dựa trên mẫu có sẵn.
8		Tiến hành thực nghiệm và kiểm chứng kết quả.
9	Nguyễn Thị Tình	Cài đặt thuật toán Khai thác Mẫu đồ thị con phổ biến.
10		Cài đặt thuật toán Sinh Luật Kết hợp Mẫu đồ thị.

Tài liệu

[1] X. Wang, Y. Xu, and H. Zhan, "Extending association rules with graph patterns," *Expert Systems with Applications*, vol. 141, p. 112897, 2020.

[2] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 721–724, IEEE, 2002.

[3] GraMi. `https://github.com/ehab-abdelhamid/GraMi`, 2015.

[4] Zaki and M. Jr. `http://www.dataminingbook.info/uploads/Main/BookPathUploads/slides-chap11.pdf`, 2014.

[5] S. network P. `https://snap.stanford.edu/data/`, 2012.

[6] W. Fan, X. Wang, Y. Wu, and J. Xu, "Association rules with graph patterns," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1502–1513, 2015.

[7] X. Wang and Y. Xu, "Mining graph pattern association rules," in *International Conference on Database and Expert Systems Applications*, pp. 223–235, Springer, 2018.

[8] M. Elseidy, E. Abdelhamid, S. Skiadopoulos, and P. Kalnis, "Grami: Frequent subgraph and pattern mining in a single large graph," *Proceedings of the VLDB Endowment*, vol. 7, no. 7, pp. 517–528, 2014.

[9] A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," in *European conference on principles of data mining and knowledge discovery*, pp. 13–23, Springer, 2000.

[10] M. Kuramochi and G. Karypis, "Frequent subgraph discovery," in *Proceedings 2001 IEEE International Conference on Data Mining*, pp. 313–320, IEEE, 2001.

XÁC NHẬN	***TP.HCM, ngày.25...tháng.07.năm 2020***
CỦA GVHD	**NHÓM SINH VIÊN THỰC HIỆN**
(Ký và ghi rõ họ tên)	***(Ký và ghi rõ họ tên)***
GS.TS Lê Hoài Bắc	**Nguyễn Thị Tình - Nguyễn Thanh Tuấn**