

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG & TIN HỌC

—o0o—



ĐỒ ÁN TỐT NGHIỆP
MỘT SỐ PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU ĐỒ THỊ
Ứng dụng trong phân tích mạng giao thông và các đồ thị cỡ lớn

Thầy/Cô hướng dẫn: **TS. LÊ CHÍ NGỌC**
Sinh viên thực hiện: **NGÔ TRƯỜNG GIANG**

Hà Nội - 06/2017

Mục lục

1	Lời mở đầu	3
2	Tập độ lập cực đại và các thuật toán tham lam	4
3	Áp dụng các thuật toán trên dữ liệu mạng xã hội và ứng dụng trong dịch tễ học	5
4	Phân tích những đặc trưng cơ bản của đồ thị	6
5	Phân cụm đồ thị và ứng dụng trong phân tích mạng giao thông thành phố Hà Nội	7
6	Frequency subgraph và ứng dụng	8
7	Kết luận	9

1 Lời mở đầu

Khai phá dữ liệu đồ thị (graph mining) là một lĩnh vực đang phát triển với mục đích khám phá ra những tri thức và hiểu biết về những dữ liệu được biểu diễn dưới dạng đồ thị. Dữ liệu đồ thị có mặt khắp trong những lĩnh vực khác của đời sống hiện đại, từ mạng xã hội, mạng internet đến mạng giao thông, mạng lưới điện,... Trong khoa học dữ liệu thì dữ liệu có thể ở rất nhiều hình dạng khác nhau, như dữ liệu vector, dữ liệu thời gian, dữ liệu không gian, dữ liệu ảnh,... Đồ thị thường được sử dụng để mô hình hóa dữ liệu khi liên kết, quan hệ giữa những đối tượng là trọng tâm của dữ liệu đó. Ví dụ, trong khoa học xã hội, mỗi một node trong đồ thị tương ứng với một người, và liên kết giữa những người đó có thể là quan hệ bạn bè nhưng trên Facebook, hay quan hệ đồng nghiệp như trên LinkedIn. Trích xuất được những thông tin, tri thức mới từ những đồ thị này có thể thúc đẩy quá trình tìm kiếm công việc mới đối với người lao động và quá trình tuyển dụng nhân sự phù hợp của các công ty, như trên mạng xã hội công việc LinkedIn đã phát triển từ lâu.

Trong báo cáo này, tôi sẽ trình bày những kĩ thuật khai phá dữ liệu trên đồ thị, như phân cụm đồ thị, phân tích các tính chất của đồ thị, phân lớp đồ thị; và áp dụng những kĩ thuật này vào một số ví dụ là những đồ thị cỡ lớn.

2 Tập độ lập cực đại và các thuật toán tham lam

3 Áp dụng các thuật toán tìm tập độ lập cực đại trên dữ liệu mạng cỡ lớn và ứng dụng trong dịch tễ học

4 Ứng dụng lý thuyết đồ thị trong phân tích mạng giao thông

4.1 Mô hình hóa dữ liệu GIS đường bộ Việt Nam bằng đồ thị

4.2 Phân tích những đặc trưng cơ bản của đồ thị

4.3 Phân tích đồ thị và ứng dụng trong phân tích mạng đường bộ Việt Nam

5 Kết luận

Trong báo cáo này, chúng em đã nêu được sơ lược về xích Markov liên tục và hàng chờ vòng đóng. Phân phối theo pha cũng được đề cập trong báo cáo này như một cơ sở để có thể mô hình hóa hàng chờ vòng đóng. Bằng cách mô hình các trạm của hàng đợi vòng đóng bằng phân phối theo pha, chúng ta có thể mô hình toàn bộ hàng chờ này bằng một xích Markov thời gian liên tục, từ đó tính được thời gian phục vụ và các độ đo hiệu năng để đánh giá hàng đợi.

Do năng lực và thời gian có hạn, báo cáo còn nhiều sai sót mong cô và các bạn góp ý thêm.

Chúng em xin chân thành cảm ơn!

Tài liệu

- [1] Svenja Lagershausen. *Markov-chain model and algorithmic procedure for the performance analysis of closed cyclic queues*. Performance Evaluation, 2015, Elsevier B.V.
- [2] Philipp Reinecke, Levente Bodrog, Alexandra Danilkina. *Phase-Type Distributions*. Resilience Assessment and Evaluation of Computing Systems, pp 85-113, 2012.