fit@hcmus

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

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

Dự đoán liên kết trong đồ thị phức

(Knowledge Graph Embedding for Link Prediction)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– Ths. Lê Ngọc Thành (Khoa Công nghệ Thông tin)

Nhóm Sinh viên thực hiện:

1. Phan Minh Tâm (18424059)
2. Hoàng Minh Thanh (18424062)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *04/2020* đến *10/2020*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Đồ thị tri thức (Knowledge Graphs-KG) là các biểu diễn cấu trúc của thông tin thế giới thực. Do khả năng mô hình hóa dữ liệu có cấu trúc, phức tạp theo cách máy tính có thể dễ dàng "hiểu được", KG hiện đang được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau, từ trả lời câu hỏi đến truy xuất thông tin và các hệ thống có thể suy luận dựa trên nội dung đã có. Việc phát triển một KG có thể

được thực hiện bằng cách trích xuất các sự kiện mới từ các nguồn bên ngoài hoặc bằng cách suy ra các sự kiện còn thiếu từ những sự kiện đã có trong KG. Phương pháp tiếp cận, được gọi là Dự đoán liên kết (Link Prediction-LP).

2.2 Mục tiêu đề tài

Cùng với nhiều kỹ thuật trí tuệ nhân tạo phát triển mạnh gần đây, đề tài tập trung nghiên cứu vào các khía cạnh của bài toán LP trên KG như đặc trưng tập dữ liệu, thuật toán, thực nghiệm đánnh giá các phương pháp cũng như các kỹ thuật khác nhau cùng tìm hiểu xem liệu những đặc trưng gì của tập dữ liệu hoặc các thuật toán khác nhau ảnh hưởng tới khả năng khái quát hóa của mô hình.

2.3 Phạm vi của đề tài

LP là một lĩnh vực nghiên cứu ngày càng sôi nổi gần đây đã phát triển mạnh mẽ từ sự bùng nổ của các kỹ thuật trong trí tuệ nhân tạo như: máy học (Machine Learning) và kỹ thuật học sâu (Deep Learning). Đề tài sẽ tập trung nghiên cứu các mô hình LP sử dụng KG làm nền tảng để tìm hiểu các biểu diễn dữ liệu với số chiều thấp còn được gọi là Knowledge Graph Embeddings, sau đó sử dụng chúng để suy ra các sự kiện, quan hệ mới.

2.4 Cách tiếp cận dự kiến

Hiện tại có 3 hướng nghiên cứu chính về lĩnh vực dự đoán liên kết bao gồm : Ma trận hóa (Matrix Factorization), Biến đổi hình học (Geometric) và kỹ thuật Học sâu (Deep Learning). Với mong muốn được tìm hiểu và nghiên cứu các phương pháp trí tuệ nhân tạo được áp dụng vào KG từ cổ điển đến hiện đại, nhóm dự kiến tìm hiểu theo hai nhóm chính gồm : Phương pháp cổ điển dựa trên luật (Rule Base) với mô hình AnyBURL [1] và nhóm phương pháp phương pháp học sâu (Deep Learning) với phương pháp KBGATs [2]. Đối với phương pháp cổ điển thuật toán cố gắng đưa các luật Horn [**?**] sau đó khái quát hóa chúng thành các luật tổng quát trên tập dữ liệu huấn luyện. Phương pháp học sâu dựa vào kỹ thuật mạng chú ý để tổng hợp thông tin từ các thực thể (entity) kế cận để chuyển vector

đặc trưng của mỗi thực thể và mỗi quan hệ đầu vào thành vector đặc trưng đầu ra với số chiều lớn hơn và giá trị được tổng hợp từ các thực thể lân cận. Sau khi biểu diễn các thực thể và quan hệ trong KGs lên số chiều đặc trưng mới, phương pháp tiến hành dự đoán các cạnh của đồ thị .

2.5 Kết quả dự kiến của đề tài

Tìm hiểu rõ các thuật toán cơ sở và tìm hiểu các thông tin liên quan để tiến hành cài đặt và chạy trên một số tập dữ liệu chuẩn rất phổ biến cho KGs là : FB15K ([3]), và WN18RR ([4]). Sau khi tìm hiểu và đọc các tài liệu liên quan có thể tiến hành cải tiến mô hình. Với phương pháp AnyBURL, nhóm dự kiến cải tiến theo hướng sử dụng một hàm heuristic để tìm kiếm hướng đi trong quá trình phát sinh luật. Với phương pháp KBGAT, nhóm dự kiến thay layer mạng chú ý (GAT [5]) bằng layer mới từ cải tiến mới nhất của mô hình Attention [6] trên việc xử lý ngôn ngữ tự nhiên - Collaborate Attention [7] .

2.6 Kế hoạch thực hiện

- 4/2020 - 6/2020: Tìm hiểu các kiến thức liên quan về mạng nơ-ron, KG. Đọc các tài liệu bài báo liên quan tới đề tài.
- 6/2020 - 8/2020: Hiện thực các phương pháp, thuât toán đã đề xuất và tìm hiểu.
- 8/2020 - 9/2020: Tìm hiểu các phương pháp khác thực nghiệm và so sánh kết quả với thuật toán gốc.
- 9/2020 - 10/2020: Báo cáo và bảo vệ khóa luận tốt nghiệp.

Tài liệu

[1] C. Meilicke, M. W. Chekol, D. Ruffinelli, and H. Stuckenschmidt, "Anytime bottom-up rule learning for knowledge graph completion.," in *IJCAI*, pp. 3137–3143, 2019.

[2] A. Rossi, D. Firmani, A. Matinata, P. Merialdo, and D. Barbosa, "Knowledge graph embedding for link prediction: A comparative analysis," *arXiv preprint arXiv:2002.00819*, 2020.

[3] K. Toutanova and D. Chen, "Observed versus latent features for knowledge base and text inference," in *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015.

[4] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, pp. 2787–2795, 2013.

[5] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[7] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "Multi-head attention: Collaborate instead of concatenate," *arXiv preprint arXiv:2006.16362*, 2020.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày 20 tháng 09 năm 2020
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)