LEARNING MADE EASY

Neo4j Special Edition

Graph Data Science (GDS)

for dummies®
A Wiley Brand

Understand graphs and GDS

Explore your graph opportunities

Discover real-world GDS use cases

Brought to you by:

neo4j

Amy Hodler
Mark Needham

About Neo4j, Inc.

Neo4j is the leader in graph database technology. As the world's most widely deployed graph database, Neo4j helps global brands — including Comcast, eBay, NASA, UBS, and Volvo — to reveal and predict how people, processes, and systems are interrelated. With this relationships-first approach, applications built using Neo4j graph technology tackle connected data challenges such as analytics and artificial intelligence, fraud detection, real-time recommendations, and knowledge graphs. Find out more at **neo4j.com.**

Dữ liệu đồ thị Khoa học (GDS)

Phiên bản đặc biệt Neo4j

của Amy Hodler và Mark Needham

for dummies®
A Wiley Brand

Khoa học dữ liệu đồ thị (GDS) cho Dummies® , Neo4j Phiên bản đặc biệt

Xuất bản bởi

John Wiley & Sons, Inc.
111 Sông St.

Hoboken, NJ 07030-5774
www.wiley.com

ISBN: 978-1-119-74604-1 (pbk); ISBN: 978-1-119-74605-8 (ebk)

Sản xuất tại Hoa Kỳ

10 9 8 7 6 5 4 3 2 1

Lời cảm ơn của nhà xuất bản

Một số người đã giúp đưa cuốn sách này ra thị trường bao gồm:

Giám đốc dự án:
Carrie Burchfield-Leighton

Sr. Tổng biên tập: Rev Mengle

Biên tập viên mua lại: Ashley Coffey

Biên tập sản xuất: Siddique Shaik

Phát triển kinh doanh
Người đại diện: Molly Daugherty

Mục lục

Giới thiệu

Khả năng kết nối là đặc điểm phổ biến nhất của mạng xã hội, cơ sở hạ tầng thông tin liên lạc đến quyền lực lưới và từ trải nghiệm bán lẻ đến chuỗi cung ứng, các mạng thậm chí có mức độ phức tạp vừa phải không phải là ngẫu nhiên, điều đó có nghĩa là các kết nối không được phân bổ đồng đều và cũng không cố định. Một mình phân tích thống kê đơn giản không thể mô tả đầy đủ, chứ chưa nói đến dự đoán, các hành vi trong các hệ thống được kết nối.

mạng và hệ thống ngày nay. Từ tương tác protein đến mạng xã hội, cơ sở hạ tầng thông tin liên lạc đến quyền lực

Khi thế giới ngày càng trở nên kết nối với nhau và các hệ thống ngày càng phức tạp, việc sử dụng các công nghệ được xây dựng để thúc đẩy các mối quan hệ và các đặc điểm năng động của chúng là bắt buộc. Không có gì đáng ngạc nhiên, sự quan tâm đến khoa học dữ liệu đồ thị (GDS) và phân tích đồ thị đã bùng nổ vì chúng được phát triển rõ ràng để thu được thông tin chi tiết từ dữ liệu được kết nối. GDS và phân tích biểu đồ tiết lộ hoạt động của các hệ thống và mạng phức tạp ở quy mô lớn.

Về cuốn sách này

Chúng tôi say mê về tiện ích và tầm quan trọng của GDS và phân tích biểu đồ, vì vậy chúng tôi đã viết cuốn sách này để giúp các tổ chức tận dụng tốt hơn các biểu đồ để họ có thể khám phá những điều mới và phát triển các giải pháp thông minh nhanh hơn.

Trong cuốn sách này, chúng tôi tập trung vào các ứng dụng thương mại của phân tích đồ thị và máy học nâng cao đồ thị (ML), có dạng GDS. Chúng tôi cũng sử dụng công nghệ đồ thị Neo4j để minh họa nền tảng GDS. Bạn hãy xem nhanh GDS và công dụng của nó trước khi kể về hành trình áp dụng GDS. Bạn cũng xem xét công nghệ Neo4j dưới dạng nền tảng GDS và xem qua một ví dụ về phát hiện gian lận.

Các biểu tượng được sử dụng trong cuốn sách này

Các biểu tượng sau đây được sử dụng trong cuốn sách này:

REMEMBER

Thông tin ở đây có thể được lưu trữ để sử dụng sau này.

TECHNICAL STUFF

Thông tin này có thể không quan trọng đối với hầu hết mọi người, nhưng nếu bạn thích những mẩu tin bổ sung về công nghệ, bạn sẽ thích thú với thông tin chi tiết ở đây. Nếu không, chỉ cần bỏ qua nó!

TIP

Bạn có quan tâm đến việc tiết kiệm thời gian hoặc công sức cho các dự án của mình không? Kiểm tra những lời khuyên để giúp bạn làm điều đó.

Ngoài cuốn sách

Cuốn sách này tập trung vào GDS và dựa trên lý thuyết đồ thị, phân tích đồ thị và cơ sở dữ liệu đồ thị. Nếu bạn muốn các nguồn lực vượt quá những gì chúng tôi có thể cung cấp cho bạn trong cuốn sách ngắn này, chúng tôi đề xuất những điều sau:

- neo4j.com/graph-algorithms-book: Đối với các ví dụ thực hành về thuật toán đồ thị, cuốn sách này cung cấp mã có thể sử dụng được và giải thích để bắt đầu.
- neo4j.com/graph-databases-book: Chi tiết bổ sung về cơ sở dữ liệu đồ thị Neo4j và mô hình đồ thị thuộc tính của nó có thể được tìm thấy tại đây.
- neo4j.com/graph-databases-for-dummies: Nếu bạn chưa quen với cơ sở dữ liệu đồ thị, thì cuốn sách này là một nơi tuyệt vời để bắt đầu hành trình của bạn vì nó không có kinh nghiệm trước đó và hướng dẫn bạn qua việc lập mô hình, truy vấn và nhập dữ liệu đồ thị, xuyên suốt hệ thống sản xuất đầu tiên của bạn.

TRONG CHƯƠNG NÀY

» Định nghĩa một đồ thị

» Hiểu về phân tích đồ thị và GDS

» Sử dụng câu hỏi để khám phá GDS

Chương 1

Hiểu về đồ thị và khoa học dữ liệu đồ thị

Cách tiếp cận đồ thị đối với dữ liệu đang bùng nổ trong thương mại thế giới để tiết lộ tốt hơn ý nghĩa trong dữ liệu cũng như dự báo hành vi của các hệ thống phức tạp. Sự bùng nổ này là do khả năng kết nối ngày càng tăng của dữ liệu, những đột phá trong việc mở rộng quy mô công nghệ đồ thị cho các vấn đề ở quy mô doanh nghiệp, kết quả xuất sắc khi được tích hợp với các giải pháp máy học (ML) và trí tuệ nhân tạo (AI) cũng như các công cụ dễ tiếp cận hơn cho phân tích chung và khoa học dữ liệu đội.

Trong chương này, bạn khám phá cách chúng tôi xác định biểu đồ và mối quan hệ của biểu đồ với phân tích và khoa học dữ liệu. Bạn cũng có được nền tảng về cách sử dụng biểu đồ để trả lời các câu hỏi khó về các hệ thống phức tạp.

Giải thích đồ thị là gì

Mạng là một đại diện, một công cụ để hiểu các hệ thống phức tạp và các kết nối phức tạp vốn có trong dữ liệu ngày nay. Ví dụ, bạn có thể trình bày cách thức hoạt động của một hệ thống xã hội bằng cách suy nghĩ về sự tương tác giữa các cặp người. Bằng cách phân tích các

cấu trúc của biểu diễn này, bạn có thể trả lời các câu hỏi và đưa ra dự đoán về cách hệ thống hoạt động hoặc cách các cá nhân hành xử trong đó. Theo nghĩa này, khoa học mạng là một tập hợp các công cụ kỹ thuật có thể áp dụng cho hầu hết mọi lĩnh vực và đồ thị là các mô hình toán học được sử dụng để thực hiện phân tích. Nói một cách đơn giản, đồ thị là một biểu diễn toán học của các hệ thống phức tạp.

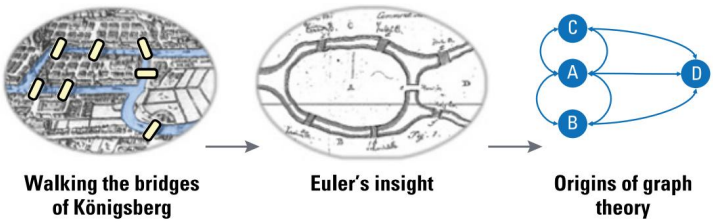
Đồ thị có lịch sử từ năm 1736. Nguồn gốc của lý thuyết đồ thị bắt nguồn từ thành phố Königsberg, bao gồm hai hòn đảo lớn được kết nối với nhau và hai phần đất liền của thành phố bằng bảy cây cầu. Câu đố là tạo ra một lối đi bộ trong thành phố, băng qua mỗi cây cầu một lần và chỉ một lần. Leonhard Euler đã giải câu đố đó bằng cách đặt câu hỏi liệu có thể tham quan cả bốn khu vực của một thành phố được nối với nhau bằng bảy cây cầu mà chỉ đi qua mỗi cây cầu một lần hay không. Không phải vậy.

Với nhận thức sâu sắc rằng chỉ bản thân các liên kết mới có liên quan đến việc giải quyết loại vấn đề này, Euler đã thiết lập nền tảng cho lý thuyết đồ thị và toán học của nó. Là một trong những bản phác thảo ban đầu của Euler, Hình 1-1 mô tả quá trình của Euler:

» Đi bộ qua những cây cầu của Königsberg: Bốn khu vực chính của Königsberg với bảy cây cầu. Bạn có thể đi qua mỗi cây cầu chỉ một lần và quay trở lại điểm xuất phát của mình không?

» Cái nhìn sâu sắc của Euler: Dữ liệu liên quan duy nhất là các khu vực chính và các cây cầu kết nối chúng.

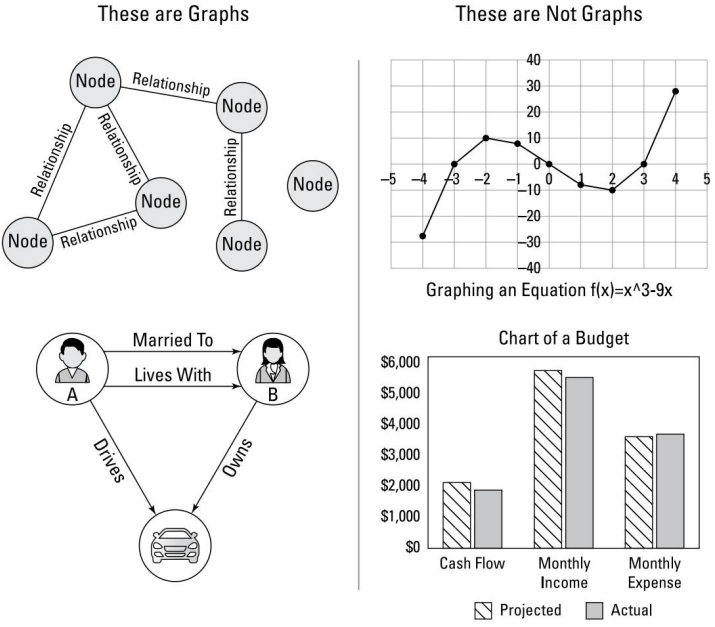
» Nguồn gốc của lý thuyết đồ thị: Euler đã trừu tượng hóa vấn đề và tạo ra các quy tắc tổng quát dựa trên các nút và mối quan hệ áp dụng cho mọi hệ thống được kết nối.



HÌNH 1-1: Nguồn gốc của lý thuyết đồ thị.

REMEMBER

Mặc dù các biểu đồ có nguồn gốc từ toán học, nhưng chúng cũng là một biểu diễn dữ liệu thực tế và trung thực để lập mô hình và phân tích. Biểu đồ là biểu diễn của một mạng, thường được minh họa bằng các vòng tròn để biểu thị các thực thể, còn được gọi là các nút hoặc đỉnh và các đường nối giữa chúng. Những dòng đó được gọi là mối quan hệ, liên kết, hoặc các cạnh. Hãy coi các nút là danh từ trong câu và các mối quan hệ vận chuyển như các động từ cung cấp ngữ cảnh cho các nút. Để tránh nhầm lẫn, các đồ thị mà chúng ta nói đến trong cuốn sách này không liên quan gì đến việc vẽ đồ thị các phương trình hoặc biểu đồ. Hãy xem sự khác biệt trong Hình 1-2.



HÌNH 1-2: Đồ thị là biểu diễn của mạng.

Đồ thị phía dưới bên trái trong Hình 1-2 là đồ thị người.
Khi nhìn vào biểu đồ đó, bạn có thể xây dựng một số câu để mô tả nó. Ví dụ: người A sống với người B sở hữu một chiếc ô tô và người A lái chiếc ô tô mà người B sở hữu. Cách tiếp cận lập mô hình này dễ dàng ánh xạ tới thế giới thực và thân thiện với bảng trắng, giúp căn chỉnh mô hình hóa và phân tích dữ liệu.

TECHNICAL STUFF

Chúng tôi thường sử dụng cụm từ “bảng trắng thân thiện” cho bất kỳ thứ gì dễ mô tả bằng các hình vẽ đơn giản mà bạn có thể minh họa trên bảng trắng.

Xác định phân tích đồ thị và đồ thị Khoa học dữ liệu

Mô hình hóa đồ thị chỉ là một nửa của câu chuyện. Bạn cũng có thể muốn phân tích chúng để tiết lộ cái nhìn sâu sắc không rõ ràng ngay lập tức. Vì vậy, trong phần này, chúng tôi giải thích lĩnh vực khoa học dữ liệu đồ thị (GDS) và phân tích đồ thị.

GDS là một cách tiếp cận dựa trên cơ sở khoa học để thu thập kiến thức từ các mối quan hệ và cấu trúc trong dữ liệu, điển hình là để tăng sức mạnh cho các dự đoán. Nó sử dụng quy trình công việc đa ngành có thể bao gồm truy vấn, số liệu thống kê, thuật toán và ML.

GDS thường có thể được chia thành ba lĩnh vực:

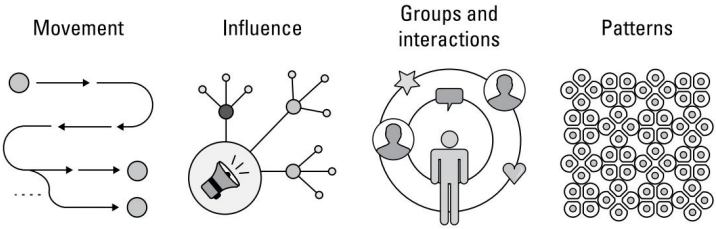
- Thống kê đồ thị cung cấp các biện pháp cơ bản về đồ thị, chẳng hạn như số lượng nút và phân phối quan hệ
  tàu thuyền. Những hiểu biết sâu sắc này có thể ảnh hưởng đến cách bạn định cấu hình và thực hiện phân tích phức tạp hơn cũng như diễn giải kết quả.
- Phân tích biểu đồ dựa trên thống kê biểu đồ bằng cách trả lời
  câu hỏi cụ thể và thu được thông tin chi tiết từ các kết nối trong dữ liệu hiện có hoặc lịch sử. Các thuật toán và truy vấn đồ thị thường được áp dụng cùng nhau trong "công thức" trong quá trình phân tích đồ thị và kết quả được sử dụng trực tiếp để phân tích.
- ML và AI nâng cao đồ thị là ứng dụng của dữ liệu đồ thị và kết quả phân tích để đào tạo các mô hình ML hoặc hỗ trợ các quyết định danh sách có thể xảy ra trong một hệ thống AI.

Số liệu thống kê và phân tích biểu đồ thường được sử dụng kết hợp để trả lời một số loại câu hỏi nhất định về các hệ thống phức tạp và thông tin chi tiết tiếp theo, được áp dụng để cải thiện ML.

Xem xét các loại câu hỏi cho GDS

REMEMBER

Các nhà khoa học dữ liệu cố gắng giải quyết nhiều loại câu hỏi khi sử dụng GDS để đánh giá sự phụ thuộc lẫn nhau, suy luận ý nghĩa và dự đoán hành vi. Ở mức độ trừu tượng nhất, những câu hỏi này rơi vào một vài lĩnh vực rộng lớn: chuyển động, ảnh hưởng, các nhóm và tương tác, và các mẫu, như trong Hình 1-3.



HÌNH 1-3: Các câu hỏi GDS thuộc bốn lĩnh vực khác nhau.

Các khu vực trong Hình 1-3 trả lời các câu hỏi sau:

» Mọi thứ di chuyển (di chuyển) qua mạng như thế nào?
Hiểu cách mọi thứ di chuyển qua mạng liên quan đến phân tích đường dẫn sâu để tìm ra các đường lan truyền, chẳng hạn như đường truyền bệnh hoặc lỗi mạng. Nó cũng có thể được sử dụng để tối ưu hóa cho tuyến đường tốt nhất có thể hoặc cho các giới hạn dòng chảy. Chúng tôi đề cập đến những cách sử dụng cổ điển này cho các thuật toán định tuyến nhiều hơn trong Chương 3.

» Điểm ảnh hưởng nhất là gì? Việc xác định những người có ảnh hưởng liên quan đến việc phát hiện ra các điểm nút được sắp xếp hợp lý về mặt cấu trúc đại diện cho các điểm kiểm soát trong mạng. Những người có ảnh hưởng này có thể đóng vai trò là điểm phổ biến nhanh, cầu nối giữa các nhóm ít kết nối hơn hoặc nút thắt cổ chai. Những người có ảnh hưởng có thể tăng tốc hoặc làm chậm dòng chảy của các mục thông qua các mạng từ tài chính đến ý kiến. Khái niệm về các nút có ảnh hưởng và kết nối cao trong biểu đồ được gọi là tính trung tâm. Các thuật toán trung tâm rất cần thiết để hiểu ảnh hưởng trong mạng.

» Các nhóm và tương tác là gì? Việc phát hiện các cộng đồng yêu cầu phải nhóm và phân vùng các nút dựa trên số lượng và cường độ của các tương tác. Phương pháp này là cách chính để giả định mối quan hệ nhóm, mặc dù sự giống nhau của hàng xóm cũng có thể là một yếu tố. Dự đoán liên kết là dự đoán các kết nối trong tương lai (hoặc chưa nhìn thấy) dựa trên cấu trúc mạng. Các thuật toán dự đoán liên kết heuristic thường được sử dụng để dự đoán hành vi. Ngoài các thuật toán phát hiện cộng đồng, các thuật toán tương tự cũng được sử dụng để hiểu các nhóm.

» Những mô hình nào là quan trọng? Khám phá mạng các mẫu tiết lộ những điểm tương đồng và cũng có thể được sử dụng để khám phá chung.

> Ví dụ: bạn có thể tìm kiếm một mẫu mối quan hệ đã biết giữa một vài nút hoặc so sánh các thuộc tính của tất cả các nút của bạn để tìm điểm tương đồng. Hoặc có lẽ bạn muốn đánh giá toàn bộ cấu trúc của mạng, với các thứ bậc phức tạp của nó, để tương quan các mẫu với hành vi xã hội nhất định cần điều tra. Tổng hợp thông tin có liên quan nhưng không rõ ràng trong các bộ dữ liệu lớn là một hoạt động phổ biến dựa trên việc tìm kiếm thông tin tương tự và có liên quan. Tìm các mẫu có thể sử dụng các truy vấn đơn giản hoặc các loại thuật toán khác nhau được tìm thấy trong Chương 3.

Nhiều loại truy vấn đồ thị và thuật toán thường được áp dụng theo kiểu công thức như một phần của quy trình làm việc GDS. Ví dụ: truy vấn để hiểu mật độ của các mối quan hệ trong biểu đồ có thể giúp xác định thuật toán phát hiện cộng đồng phù hợp để có kết quả phù hợp nhất. Về mặt chiến thuật, các thuật toán và truy vấn đồ thị là những công cụ để hiểu bản chất tổng thể của một hệ thống được kết nối và để sử dụng các mối quan hệ trong các quy trình khoa học dữ liệu khác nhau.

SỰ NỔI BẬT CỦA KHOA HỌC DỮ LIỆU ĐỒ THỊ

Sự phát triển của khoa học dữ liệu đồ thị (GDS) là kết quả của các công nghệ dễ tiếp cận hơn, tăng khả năng tính toán trên các tập dữ liệu đồ thị lớn và nhận thức về sức mạnh của đồ thị trong việc suy luận ý nghĩa và cải thiện dự báo. Các nhà nghiên cứu đóng một vai trò thiết yếu trong việc phát triển nhận thức và ủng hộ các kỹ thuật tốt nhất. Khi các nhà khoa học dữ liệu nhìn thấy tiềm năng của thông tin cấu trúc, họ ngày càng tích hợp các biểu đồ vào các phương pháp thống kê, phân tích và ML của họ. Trên thực tế, theo hệ thống Kiến thức thứ nguyên cho các ấn phẩm nghiên cứu, việc sử dụng công nghệ đồ thị trong nghiên cứu AI đang tăng tốc. Trong 10 năm qua, số lượng tài liệu nghiên cứu về AI có sử dụng công nghệ đồ thị đã tăng hơn 700%.

TRONG CHƯƠNG NÀY

» Xem biểu đồ giúp ngành chăm sóc sức khỏe như thế nào

» Sử dụng đồ thị trong marketing

» Đưa đồ thị vào hoạt động để ngăn chặn gian lận

chương 2

Sử dụng dữ liệu đồ thị Khoa học trong thế giới thực

Những thách thức dữ liệu cấp bách nhất hiện nay xoay quanh khám phá và tận dụng cấu trúc mạng của khoa học dữ liệu đồ thị (GDS) thúc đẩy một loạt các trường hợp sử dụng từ phòng chống gian lận và đề xuất có mục tiêu đến trải nghiệm được cá nhân hóa và tái sử dụng thuốc.

các mối quan hệ, không chỉ lập bảng dữ liệu rời rạc. Khả năng

REMEMBER

Chúng ta không thể phóng đại tác động của các kỹ thuật đồ thị được cải thiện như thuật toán mới hoặc nỗ lực của các nhà khoa học mạng ứng dụng như trong sinh học tính toán. Chúng tôi cũng không muốn bạn bỏ qua các dự án xã hội sử dụng biểu đồ. Tuy nhiên, chúng tôi tin rằng sự bùng nổ gần đây của biểu đồ trong thế giới kinh doanh thể hiện sự thay đổi về khả năng tiếp cận và cơ hội thúc đẩy quá trình dân chủ hóa biểu đồ cho mọi người.

Công nghệ đồ thị giúp các tổ chức có nhiều trường hợp sử dụng thực tế trong các ngành và lĩnh vực. Trước đây, nhiều doanh nghiệp đã bắt đầu khám phá công nghệ đồ thị để tạo ra cái nhìn 360 độ về khách hàng của họ hoặc để thống nhất dữ liệu tổng thể, bao gồm thông tin về khách hàng, sản phẩm, nhà cung cấp và hậu cần. Họ có thể sử dụng loại theo dõi này để cải thiện trải nghiệm của khách hàng hoặc để đáp ứng các quy định tuân thủ của các đạo luật về quyền riêng tư gần đây như của EU

Quy định chung về bảo vệ dữ liệu (GDPR) và Đạo luật về quyền riêng tư của người tiêu dùng California (CCPA). Loại chế độ xem hoàn chỉnh và dòng dữ liệu trong biểu đồ này hiện cũng được sử dụng để hiểu và theo dõi dữ liệu được sử dụng trong máy học (ML) cho các ứng dụng trí tuệ nhân tạo (AI) có trách nhiệm hơn.

Ngày nay, các doanh nghiệp có xu hướng xem xét sử dụng biểu đồ đặc biệt cho khoa học dữ liệu khi họ nhận ra sức mạnh dự đoán của các mối quan hệ, khả năng sử dụng cấu trúc mạng để cải thiện ML và nhu cầu đổi mới của chính họ. Các phần trong chương này nêu bật một số trường hợp sử dụng GDS trong các lĩnh vực thúc đẩy tăng trưởng và lợi ích thương mại đáng kể.

Nhìn vào đồ thị trong chăm sóc sức khỏe

Thật dễ dàng để thấy bất kỳ ngành công nghiệp nào có nguồn gốc sinh học sẽ tự nhiên hiểu được tầm quan trọng của các hệ thống kết nối với nhau như thế nào. Bạn có thể thấy mối quan hệ này trong sinh học tính toán cũng như chăm sóc sức khỏe và khoa học đời sống trong cách họ nhìn nhận những thách thức như một phần của quy trình lớn hơn. Hai ví dụ nổi bật để phục vụ lợi ích sức khỏe và thương mại: khám phá thuốc hiệu quả hơn và kết quả bệnh nhân tốt hơn.

Khám phá các loại thuốc hiệu quả hơn

An toàn, tốc độ và chi phí là tối quan trọng trong việc tạo ra các giải pháp thuốc mới có thể tiếp cận được. Đồ thị có thể giúp giải quyết sự phức tạp của các mối quan hệ đan xen giữa bệnh tật, gen, thuốc, tác dụng phụ và nhân khẩu học – chỉ kể tên một số vấn đề cần cân nhắc.

Một biểu đồ tri thức ấn tượng trong ngành khoa học đời sống tích hợp hơn 50 năm dữ liệu y sinh bao gồm gen, hợp chất, bệnh tật và các thông tin khác như triệu chứng và tác dụng phụ. Một trong những dự án từ biểu đồ dự đoán cách sử dụng thuốc mới bằng cách sử dụng cấu trúc liên kết biểu đồ. Biểu đồ giúp dự đoán cách sử dụng mới cho các loại thuốc hiện đã được phê duyệt bằng cách đánh giá các mối quan hệ, cấu trúc mạng lưới và sự tương đồng. Việc tái sử dụng thuốc giúp giảm đáng kể chi phí và thời gian đưa ra thị trường so với việc phát triển và thử nghiệm các loại thuốc mới – chưa kể đến lợi ích của việc có sẵn nhiều thông tin thực tế hơn về tác dụng phụ và kết quả không mong muốn khi thuốc đã được sử dụng.

Cải thiện hành trình của bệnh nhân

Một lĩnh vực khác đang được quan tâm là việc sử dụng đồ thị để lập bản đồ, đánh giá và cải thiện hành trình của bệnh nhân. Khi một bệnh nhân cảm thấy không khỏe, có nhiều yếu tố tác động có thể đã phát triển trong một khoảng thời gian. Tương tự như vậy, các phương pháp điều trị hiếm khi là một sự kiện đơn lẻ, đặc biệt là đối với các bệnh mãn tính hoặc nghiêm trọng. Cây các triệu chứng có thể xảy ra, các lần thăm khám, xét nghiệm, người chăm sóc, kế hoạch điều trị, kết quả, sau đó là các xét nghiệm và điều trị thứ cấp, v.v. có thể phân nhánh thành vô số con đường khả thi. Hãy tưởng tượng các tùy chọn điều trị bệnh nhân có thể được ánh xạ bằng biểu đồ để thấy rõ hơn các lựa chọn thay thế theo trình tự và phân chia đường dẫn sau mỗi và mọi kết quả xét nghiệm hoặc lần khám. Trên thực tế, các nhà nghiên cứu và nhà cung cấp dịch vụ chăm sóc sức khỏe đã sử dụng biểu đồ để hiểu rõ hơn điều gì ảnh hưởng đến hành trình của bệnh nhân để họ có thể cải thiện kết quả cá nhân cũng như tạo và so sánh với các lộ trình tối ưu.

Khuyến nghị và Tiếp thị được cá nhân hóa

Việc đưa ra các đề xuất về sản phẩm và dịch vụ có liên quan đòi hỏi phải có mối tương quan giữa sản phẩm, thông tin khách hàng, hành vi lịch sử, hàng tồn kho, nhà cung cấp, hậu cần và thậm chí cả dữ liệu tình cảm xã hội. Các đề xuất dựa trên biểu đồ và hoạt động tiếp thị có mục tiêu giúp các công ty cung cấp các dịch vụ và trải nghiệm phù hợp hơn cho nhiều người dùng hơn. Ví dụ: thuật toán phát hiện cộng đồng đồ thị được sử dụng để nhóm các khách hàng có tương tác hoặc hành vi tương tự để đưa ra các đề xuất phù hợp hơn. Nghiên cứu cho thấy rằng ML nâng cao đồ thị có thể dự đoán tỷ lệ khách hàng rời bỏ, chẳng hạn như đối với các mục đích sử dụng như phòng ngừa hoặc tiếp thị có mục tiêu.

Phân tích biểu đồ cũng được sử dụng để giúp nhắm mục tiêu ưu đãi cho người dùng trực tuyến ẩn danh về tên và nhân khẩu học nhưng không ẩn danh trong hành vi trang web. Thông tin chi tiết từ phân tích được thực hiện ngoại tuyến thường được đưa vào các mô hình quyết định được sử dụng trong sản xuất để đưa ra các đề xuất theo thời gian thực, có thể bao gồm các đề xuất cho các sản phẩm vận chuyển nhanh hơn dựa trên việc thay đổi mức độ tồn kho hoặc dữ liệu tích hợp tức thời từ lượt truy cập hiện tại của khách hàng.

Phát hiện gian lận

Số tiền bị mất do gian lận mỗi năm ngày càng tăng, mặc dù việc sử dụng AI và ML để phát hiện và ngăn chặn gian lận ngày càng tăng. Để phát hiện ra nhiều hành vi gian lận hơn trong khi tránh các kết quả sai gây tốn kém, các tổ chức nhìn xa hơn các điểm dữ liệu riêng lẻ đến các kết nối và mẫu liên kết chúng. Các tổ chức sử dụng cấu trúc mạng để tăng cường các quy trình ML hiện có như một cách tiếp cận thực tế để tăng số lượng gian lận được phát hiện và khôi phục.

Kỹ thuật tính năng biểu đồ cho phép doanh nghiệp trích xuất các yếu tố dự đoán dựa trên các truy vấn hoặc thuật toán biểu đồ và sử dụng thông tin đó để huấn luyện các mô hình ML. Cải thiện độ chính xác dự đoán trong việc phát hiện gian lận, ngay cả những điểm phần trăm nhỏ cũng có thể giúp tiết kiệm hàng chục triệu đô la chỉ sau vài tháng. GDS cho phép các công ty vượt qua các mô hình gian lận luôn thay đổi cũng như thu hồi nhiều khoản lỗ hơn.

Hãy chuyển sang Chương 5, nơi chúng tôi cung cấp cho bạn một ví dụ chi tiết về cách phát hiện gian lận với GDS.

TRONG CHƯƠNG NÀY

» Tổng hợp thông tin đa dạng

» Sử dụng phân tích biểu đồ để hiểu mạng của bạn

» Tìm kiếm, kết hợp và trích xuất các yếu tố dự đoán

» Đơn giản hóa đồ thị bằng cách nhúng

» Tiếp cận mới với mạng lưới đồ thị

Chương 3

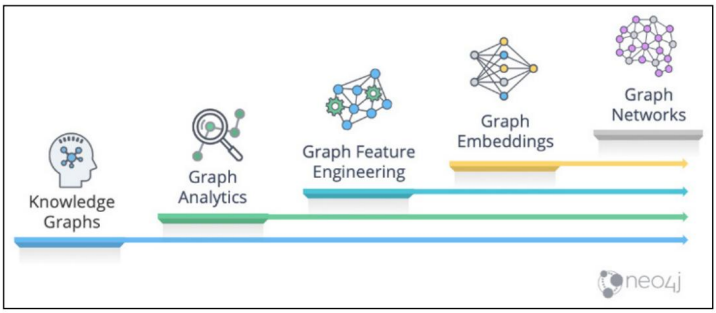
Phát triển của bạn

Ứng dụng của GDS Công nghệ

Ngày nay, khoa học dữ liệu đồ thị (GDS) thường được áp dụng trong kinh doanh với một hoặc nhiều mục tiêu chính trong tâm trí: quyết định tốt hơn, mang lại chất lượng dự đoán và tạo ra những cách thức mới để đổi mới và học hỏi. Những mục tiêu này ngày càng gắn liền với những lợi ích hữu hình, chẳng hạn như giảm tổn thất tài chính, thời gian đạt được kết quả nhanh hơn, tăng sự hài lòng của khách hàng và mức tăng dự đoán. Bạn có thể đang cố gắng cải thiện hoặc tự động hóa quá trình ra quyết định của những người và chuyên gia trong lĩnh vực cần thêm bối cảnh. Hoặc có lẽ mục tiêu của bạn là cải thiện độ chính xác của dự đoán bằng cách sử dụng các mối quan hệ và cấu trúc công việc mạng trong phân tích và học máy (ML).

Đồ thị cung cấp một cấu trúc độc đáo cho việc học giúp phát triển các kỹ thuật ML thông qua khả năng diễn giải và trừu tượng hóa tốt hơn. Các mục tiêu kinh doanh này ánh xạ mạnh mẽ đến cách các tổ chức tích hợp công nghệ đồ thị vào các hoạt động khoa học dữ liệu của họ. Hình 3-1 sơ đồ các giai đoạn chính của hành trình GDS điển hình. Chúng tôi bao gồm mỗi

của các giai đoạn này trong chương này. Ba giai đoạn đầu tiên của hành trình GDS phổ biến nhất trong thế giới thương mại ngày nay và hai giai đoạn cuối cùng là những giai đoạn mới nổi trên hành trình GDS của bạn.



HÌNH 3-1: Hành trình GDS.

REMEMBER

Tổ chức của bạn có thể sử dụng các bước thực tế để đạt được giá trị ngay lập tức và sau đó xếp lớp các kỹ thuật phức tạp hơn theo cách liên tục làm tăng hiệu quả nỗ lực của bạn.

Sơ đồ tri thức

Biểu đồ tri thức là nền tảng của GDS và cung cấp một cách hợp lý hóa quy trình công việc, tự động hóa phản hồi và mở rộng các quyết định thông minh. Ở cấp độ cao, biểu đồ tri thức là tập hợp các điểm dữ liệu được liên kết với nhau và mô tả các thực thể, sự kiện hoặc sự vật trong thế giới thực và mối quan hệ của chúng với nhau ở dạng con người có thể hiểu được. Không giống như một cơ sở tri thức đơn giản với cấu trúc phẳng và nội dung tĩnh, biểu đồ tri thức thu thập và tích hợp thông tin liền kề bằng cách sử dụng các mối quan hệ dữ liệu để lấy tri thức mới.

Là giai đoạn đầu tiên trong GDS, các biểu đồ tri thức thường được triển khai để tập hợp các thông tin đa dạng nhằm giúp các chuyên gia miền tìm thấy nội dung liên quan cũng như khám phá các kết nối trong dữ liệu của họ. Biểu đồ tri thức cũng có thể thêm ngữ cảnh vào các ứng dụng, chẳng hạn như các ứng dụng trong hệ thống trí tuệ nhân tạo (AI), để chúng có thể đưa ra các quyết định gần đúng tốt hơn và nhanh hơn. Cách tiếp cận này được sử dụng trong các hệ thống AI, chẳng hạn như chatbot, chẳng hạn như sử dụng biểu đồ tri thức để định tuyến tốt hơn yêu cầu “con dơi cho sinh nhật của ban nhạc chồng tôi”. Trong trường hợp này, biểu đồ nắm bắt rằng yêu cầu rất có thể không phải là động vật có vú biết bay mà ai đó đang tìm kiếm mà là

thay vào đó là đồ thể thao chất lượng cao hơn cho một dịp đặc biệt. Chatbot cũng có thể tính đến những gì có trong kho, thời gian vận chuyển và các sản phẩm đặc biệt kết hợp bối cảnh không chỉ của người yêu cầu mà còn của nguồn cung và hậu cần khác.

Phân tích đồ thị

Sau khi triển khai biểu đồ tri thức (xem phần trước), các doanh nghiệp thường bắt đầu sử dụng phân tích biểu đồ để hiểu rõ hơn về mạng của họ và trả lời các câu hỏi cụ thể dựa trên các mối quan hệ và cấu trúc liên kết. Bạn thường cố gắng suy luận ý nghĩa dựa trên cấu trúc mạng: tìm các cụm, xác định các nút có ảnh hưởng, đánh giá các lộ trình khác nhau. Phân tích biểu đồ thường đề cập đến việc sử dụng các truy vấn và thuật toán toàn cầu xem xét toàn bộ biểu đồ để phân tích ngoại tuyến dữ liệu lịch sử. Quá trình này trái ngược với các giao dịch nhỏ, thời gian thực và các truy vấn cục bộ tập trung vào các khu vực nhỏ xung quanh một vài nút.

Truy vấn đồ thị được sử dụng khi bạn biết chính xác những gì bạn đang tìm kiếm, chẳng hạn như đặt câu hỏi như “Mia có bao nhiêu mối quan hệ?” hoặc “Có bao nhiêu kẻ lừa đảo hoặc tài khoản bị gắn cờ cách đây bốn bước?” (Một bước nhảy là một cấp độ hoặc một lớp của mối quan hệ.) Những loại truy vấn này có vẻ đơn giản vì chúng ta có thể tưởng tượng việc đứng lên và nhìn vào những thứ ở gần chúng ta. Tuy nhiên, các giải pháp không lưu trữ các mối quan hệ cùng với dữ liệu của chúng phải thực hiện các quy trình bổ sung để tra cứu và kết hợp thông tin liên quan nà Đồ thị lưu trữ các mối quan hệ cùng với dữ liệu nên việc theo dõi đường dẫn của các mối quan hệ rất đơn giản và nhanh chóng. Cơ sở dữ liệu đồ thị gốc đặc biệt tốt trong nhiều truy vấn hop vì chúng tránh tra cứu chỉ mục tốn kém và nối dữ liệu bằng cách lưu trữ và xử lý thông tin liên quan một cách liền kề và coi các mối quan hệ là công dân hạng nhất.

Các thuật toán đồ thị là một tập hợp con của các thuật toán khoa học dữ liệu có nguồn gốc từ khoa học mạng để cho phép lập luận về cấu trúc theo kiểu không giám sát hơn. Chúng được sử dụng khi bạn biết mẫu hoặc chỉ báo mà bạn đang tìm kiếm nhưng không chính xác những gì bạn sẽ tìm thấy. Ví dụ: bạn có thể đang tìm kiếm các cộng đồng chặt chẽ bất thường nơi các nút có nhiều mối quan hệ với nhau hơn bạn mong đợi trong phân phối ngẫu nhiên hoặc bình thường. Để tìm các cộng đồng này, bạn có thể sử dụng thuật toán đồ thị có tên Louvain Modularity để khám phá các cụm có mức độ tương tác cao hơn

mật độ bên trong, giữa các thành viên trong nhóm khi so sánh với các hành động tương tác bên ngoài nhóm.

Các thuật toán đồ thị phổ biến nhất trong các ứng dụng thương mại rơi vào khoảng sáu loại:

» Tìm đường và tìm kiếm: Các thuật toán này là nền tảng để phân tích biểu đồ và khám phá các đường dẫn giữa các nút. Họ đánh giá các tuyến đường để sử dụng như hậu cần vật lý và định tuyến cuộc gọi hoặc giao thức Internet (IP) với chi phí thấp nhất.

» Tính trung tâm (tầm quan trọng): Thuật toán tính trung tâm giúp bạn khám phá vai trò của các nút riêng lẻ và tác động của chúng. Họ xác định các nút có ảnh hưởng dựa trên vị trí của chúng trong mạng. Các thuật toán này suy ra các động lực của nhóm, chẳng hạn như độ tin cậy, tính dễ bị tổn thương và cầu nối giữa các nhóm.

» Phát hiện cộng đồng: Các thuật toán này tìm các cộng đồng nơi các thành viên có nhiều tương tác quan trọng hơn. Những kết nối này tiết lộ các cụm chặt chẽ, các nhóm bị cô lập và cấu trúc. Thông tin này giúp dự đoán hành vi hoặc sở thích tương tự, ước tính khả năng phục hồi, tìm các thực thể trùng lặp hoặc đơn giản là chuẩn bị dữ liệu cho các phân tích khác.

» Tính tương tự: Các thuật toán này sử dụng phép so sánh tập hợp để xem các nút riêng lẻ giống nhau như thế nào. Các thuộc tính và thuộc tính của các nút được sử dụng để ghi điểm giống nhau giữa điểm giao. Cách tiếp cận này được sử dụng trong các ứng dụng như đề xuất được cá nhân hóa cũng như phát triển hệ thống phân cấp theo phân loại.

» Dự đoán liên kết theo kinh nghiệm: Các thuật toán này xem xét khoảng cách gần của các nút trong mạng cũng như các yếu tố cấu trúc, chẳng hạn như hình tam giác có thể có giữa các nút, để ước tính khả năng hình thành một mối quan hệ mới hoặc tồn tại các kết nối không có tài liệu. Loại thuật toán này có nhiều ứng dụng từ tái sử dụng ma túy đến điều tra tội phạm.

» Nhúng biểu đồ: Các thuật toán này dịch cấu trúc liên kết và thuộc tính của biểu đồ thành một biểu diễn số duy nhất có thể được sử dụng cho kỹ thuật tính năng (xem phần tiếp theo “Kỹ thuật tính năng biểu đồ” để biết thêm thông tin), tính toán tương tự hoặc trực quan hóa. Không giống như các thuật toán đồ thị truyền thống sử dụng các công thức được tính toán trước, các nhúng tìm hiểu biểu diễn từ đồ thị của bạn dựa trên các mô hình mạng thần kinh (học sâu) hoặc đại số tuyến tính. Xem phần sau “Nhúng biểu đồ” trong chương này để biết thêm về nhúng biểu đồ.

REMEMBER

Trong phân tích biểu đồ, bạn đang hỏi một câu hỏi được nhắm mục tiêu hoặc xem toàn bộ biểu đồ để suy ra ý nghĩa hoặc đưa ra dự đoán về hành vi trong tương lai.

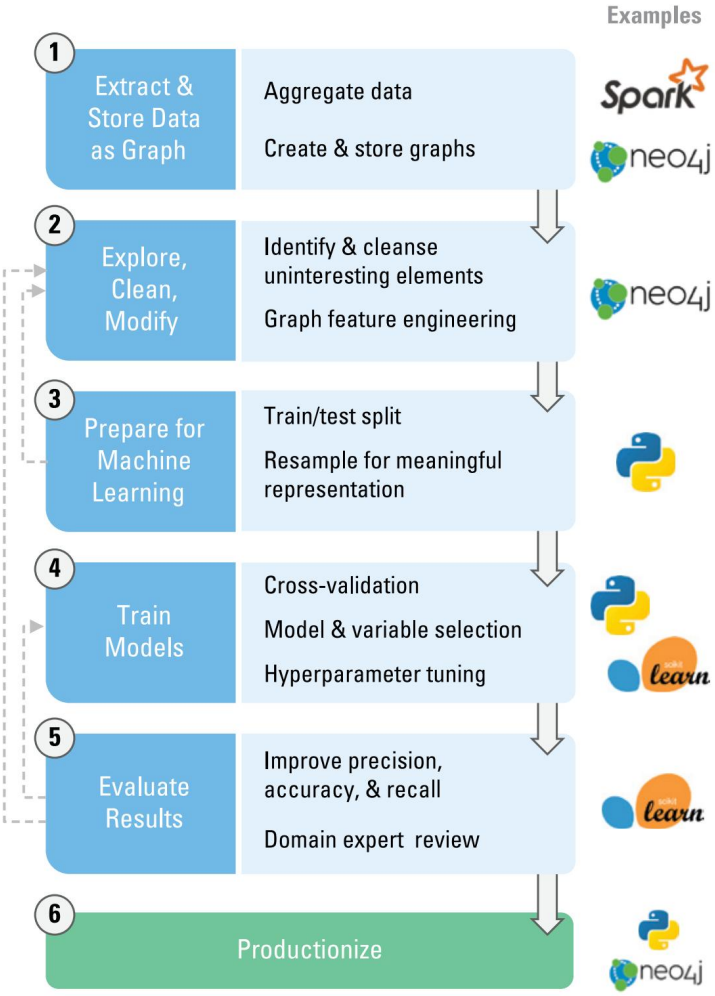
Kỹ thuật tính năng đồ thị

Kỹ thuật tính năng biểu đồ là quá trình tìm kiếm, kết hợp và trích xuất các phần tử dự đoán từ dữ liệu biểu đồ thô để sử dụng trong các tác vụ ML. Nhiều thông tin hơn thường làm cho các mô hình ML chính xác hơn, nhưng các nhà khoa học dữ liệu hiếm khi có nhiều dữ liệu như họ muốn. Vì các mối quan hệ có tính dự đoán cao về hành vi và chúng vốn tồn tại bên trong dữ liệu hiện tại, nên bạn có thể sử dụng kỹ thuật tính năng biểu đồ để cải thiện các dự đoán và tăng độ chính xác của mô hình ML — với dữ liệu bạn đã có.

Kỹ thuật tính năng đồ thị sử dụng các mối quan hệ và cấu trúc mạng để tạo ra các tính năng mới, có ý nghĩa hơn. Đây là bước tiếp theo để áp dụng những gì bạn học được từ phân tích biểu đồ vào ML. Ví dụ: bạn có thể chấm điểm các nút dựa trên truy vấn tính toán có bao nhiêu kẻ gian lận trong bốn bước nhảy hoặc thuật toán trung tâm để đo lường tầm quan trọng. Bạn cũng có thể gắn nhãn các nút dựa trên ID cộng đồng của chúng. (ID cộng đồng được chỉ định bởi thuật toán phát hiện cộng đồng.) Sau đó, các điểm số và nhãn này có thể được trích xuất thành danh sách hoặc bảng số và số nhận dạng (còn được gọi là vectơ đặc trưng) để huấn luyện các mô hình ML. Các tính năng biểu đồ và chỉ số ML kết quả thường được ghi lại vào cơ sở dữ liệu biểu đồ để sử dụng liên tục và trong tương lai.

Hình 3-2 cho thấy cách sử dụng các tính năng biểu đồ để nâng cao ML là một phần của quy trình công việc lớn hơn với một số công nghệ ví dụ để minh họa.

Đối với ML nâng cao đồ thị, thông thường bạn sẽ tổng hợp, khám phá và làm sạch dữ liệu, sau đó sử dụng các truy vấn hoặc thuật toán đồ thị cho kỹ thuật tính năng. Sau đó, bạn sẽ chuẩn bị dữ liệu cho ML và chia dữ liệu đó thành tập dữ liệu huấn luyện và thử nghiệm. Mặc dù quy trình này không hoàn toàn tuyến tính, nhưng sau khi bạn đã đào tạo một mô hình và hài lòng với kết quả, thì mô hình đó có thể được sử dụng trong sản xuất. Mặc dù mô hình có thể cung cấp giao dịch thời gian thực trong quá trình sản xuất, chẳng hạn như phê duyệt các ứng dụng tín dụng trực tuyến, nhưng kỹ thuật tính năng biểu đồ và ML được thực hiện ngoại tuyến và được cập nhật định kỳ theo quy trình theo chu kỳ. Kỹ thuật tính năng đồ thị cung cấp cho các tổ chức những cải tiến mô hình có thể đạt được mà không cần thay đổi quy trình ML của họ.



HÌNH 3-2: Kỹ thuật tính năng đồ thị là một phần của quy trình ML lớn hơn.

Nhúng đồ thị

Nhúng biểu đồ đơn giản hóa các biểu đồ hoặc tập hợp con của biểu đồ thành một vectơ đặc trưng hoặc tập hợp các vectơ ở dạng chiều thấp hơn, chẳng hạn như một danh sách các số. Mục tiêu là tạo dữ liệu có thể tiêu thụ dễ dàng cho các tác vụ như ML vẫn mô tả phức tạp hơn thuộc tính cấu trúc liên kết, kết nối hoặc nút. Ví dụ: bạn có thể biểu thị toàn bộ biểu đồ hoặc đường dẫn dưới dạng nhúng và sau đó

học dựa trên biểu đồ hoặc đường dẫn. Có ba loại nhúng đồ thị:

- Nút nhúng mô tả khả năng kết nối của từng nút.
- Nhúng đường dẫn bao gồm các đường đi ngang qua biểu đồ.
- Nhúng đồ thị mã hóa toàn bộ đồ thị thành một đồ thị duy nhất véc tơ.


REMEMBER

Nhúng biểu đồ thường được sử dụng cho kỹ thuật tính năng nâng cao hơn kết hợp nhiều thông tin phức tạp hơn, đó là lý do tại sao giai đoạn này thường đến sau trong hành trình GDS. Các lệnh nhúng cũng có thể hữu ích cho việc khám phá dữ liệu, tính toán sự tương đồng giữa các thực thể và giảm kích thước để hỗ trợ phân tích thống kê. Nhúng biểu đồ cung cấp khả năng sử dụng rộng rãi hơn các cấu trúc phong phú tạo nên biểu đồ trong các tác vụ khoa học dữ liệu khác nhau và tìm hiểu dựa trên thông tin sắc thái.

Mạng đồ thị

Mạng đồ thị là một lĩnh vực nghiên cứu thú vị đại diện cho một cách tiếp cận mới đối với ML có thể cải thiện đáng kể kết quả với ít dữ liệu hơn, đưa ra dự đoán dễ hiểu hơn và dẫn đến các kiểu học tập mới. Mạng đồ thị và học tập bản địa đồ thị là những thuật ngữ được đặt ra bởi Peter Battaglia và một nhóm các nhà nghiên cứu. Họ kết luận rằng việc sử dụng biểu đồ cho ML là bước tiến lớn tiếp theo trong chính ML vì khả năng trừu tượng hóa topol ogy của biểu đồ. Suy nghĩ của họ theo cách tiếp cận này:

1. Học đồ thị gốc lấy đồ thị làm đầu vào, mỗi hình thành các tính toán học tập trong khi vẫn duy trì các trạng thái tạm thời, sau đó trả về một biểu đồ.
2. Quá trình học đồ thị gốc này cho phép chuyên gia miền xem xét và xác thực lộ trình học dẫn đến các dự đoán dễ giải thích hơn.
3. Quá trình này trở nên phong phú và chính xác hơn dự đoán sử dụng ít dữ liệu và chu kỳ đào tạo.

REMEMBER

Tính năng học tập riêng của đồ thị cho phép học toàn bộ đồ thị và dự đoán nhiều tác vụ giúp giảm yêu cầu dữ liệu và tự động hóa việc xác định các tính năng có liên quan. Ngày nay, thời gian quý giá của các nhà khoa học dữ liệu và chuyên gia trong lĩnh vực thường được sử dụng để lựa chọn và kiểm tra dữ liệu có khả năng dự đoán một cách tẻ nhạt và thu thập các đặc điểm đó vào các mô hình tối ưu. Cải thiện độ chính xác của mô hình trong khi hợp lý hóa quy trình tác động tích cực đến các quy trình và kết quả ML trên tất cả các ứng dụng. Chúng tôi rất phấn khích trước những tiến bộ ban đầu và mong muốn được thấy ML phát triển để trở nên cực kỳ hiệu quả và linh hoạt cũng như chính xác và minh bạch hơn.

TRONG CHƯƠNG NÀY

Chương 4

Sử dụng Neo4j làm Biểu đồ Nền tảng khoa học dữ liệu

Nếu bạn đã sử dụng khoa học dữ liệu đồ thị, GDS là một phần mềm mà Neo4j cung cấp để giúp bạn. Neo4j là một công ty công nghệ đồ thị cung cấp nền tảng GDS cấp doanh nghiệp bao gồm bốn thành phần.

REMEMBER

Neo4j hỗ trợ xử lý giao dịch và xử lý phân tích dữ liệu biểu đồ cũng như trực quan hóa. Nó cũng bao gồm lưu trữ đồ thị và tính toán với công cụ phân tích và quản lý dữ liệu. Bộ công cụ tích hợp bao gồm một giao thức chung, API và ngôn ngữ truy vấn (Cypher) để cung cấp quyền truy cập hiệu quả cho các mục đích sử dụng khác nhau. Trong chương này, chúng tôi sẽ trình bày chi tiết hơn một chút về từng lĩnh vực trong số bốn lĩnh vực của nền tảng Neo4j để giúp bạn thấy giải pháp GDS của mình phù hợp với nhau như thế nào.

Thư viện Neo4j GDS

Thư viện Neo4j GDS cung cấp một cách tiếp cận sẵn sàng cho doanh nghiệp để chạy các thuật toán đồ thị phức tạp trên dữ liệu được kết nối ở quy mô lớn. Phân tích biểu đồ và kỹ thuật tính năng bổ sung các mối quan hệ có tính dự đoán cao vào quá trình học máy (ML) của bạn để tốt hơn

kết quả. Các thuật toán được thực thi trong một không gian làm việc phân tích giúp chia tỷ lệ tính toán để xử lý các biểu đồ chứa hàng chục tỷ nút và mối quan hệ. Để biết ví dụ, đào tạo và chi tiết về cách sử dụng Thư viện Neo4j GDS, hãy truy cập neo4j.com/developer/
đồ thị-thuật toán. Bạn cũng có thể truy cập trực tiếp vào Thư viện Neo4j GDS tại neo4j.com/graph-data-science-library.

Cơ sở dữ liệu đồ thị Neo4j Hệ thống quản lý

Hệ thống quản lý cơ sở dữ liệu Neo4j (DBMS) hỗ trợ nhiều cơ sở dữ liệu nhỏ có thể chạy trong các cài đặt độc lập hoặc theo cụm và hỗ trợ truy cập phân mảnh và liên kết vào cơ sở dữ liệu.

Cơ sở dữ liệu đồ thị Neo4j được thiết kế để coi mối quan hệ giữa dữ liệu cũng quan trọng như chính dữ liệu đó. Nó được coi là cơ sở dữ liệu đồ thị gốc vì dữ liệu được lưu trữ cùng với cách mỗi thực thể riêng lẻ kết nối với hoặc có liên quan với những thực thể khác.

Bạn có thể tìm thêm thông tin về Neo4j Graph DBMS tại neo4j.com/developer/graph-database.

TIP

Để khám phá thêm về mô hình biểu đồ thuộc tính được DBMS và các công cụ khác sử dụng, hãy xem Cơ sở dữ liệu đồ thị cho người mới bắt đầu, Phiên bản đặc biệt Neo4j, tại neo4j.com/graph-databases-for-dummies.

CÂU HỎI TUYÊN BỐ CỦA CYPHER NGÔN NGỮ

Cypher là ngôn ngữ truy vấn mở, được xác định đầy đủ và được áp dụng rộng rãi nhất cho cơ sở dữ liệu đồ thị thuộc tính. Nó là một ngôn ngữ khai báo, lấy cảm hứng từ SQL để mô tả các mẫu trực quan trong biểu đồ bằng cách sử dụng cú pháp ASCII-Art. Bạn có thể nêu những gì bạn muốn chọn, chèn, cập nhật hoặc xóa khỏi dữ liệu biểu đồ của mình mà không cần mô tả cách thực hiện.

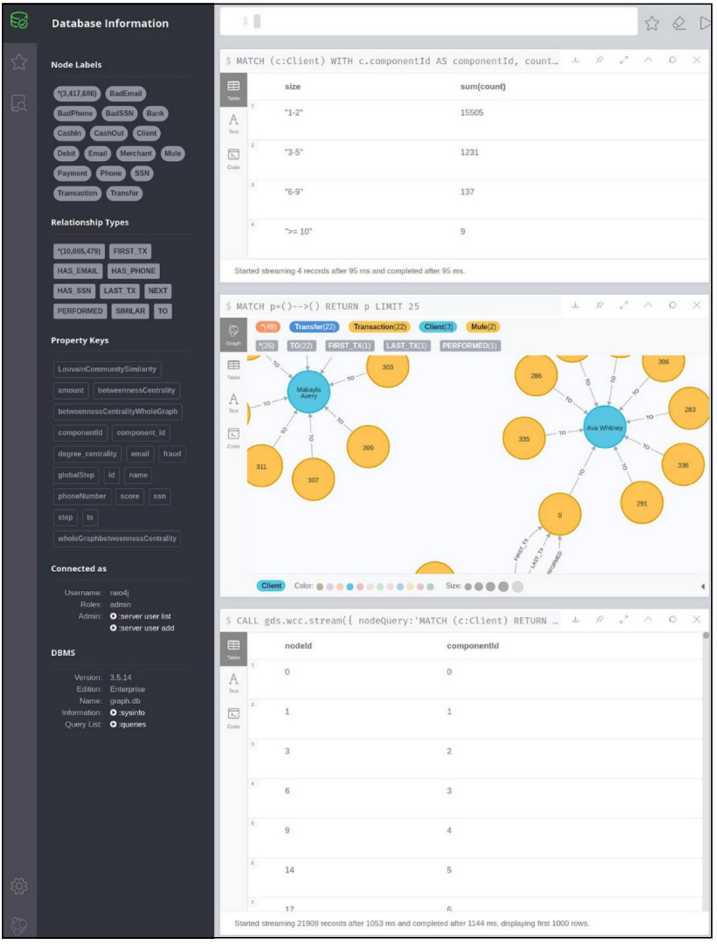
Cypher được thiết kế để có thể đọc được. Ví dụ: cụm từ "Jennifer thích công nghệ đồ thị" sẽ được viết là

```
(p:Person {name: "Jennifer"})-[rel:LIKES]->(g:Technology {type: "Graphs"})
```

Bạn có thể tìm thấy các tài nguyên học tập và thông tin cơ bản về Cypher trên trang Cypher dành cho nhà phát triển Neo4j tại neo4j.com/developer/cypher query-language.

Máy tính để bàn và trình duyệt Neo4j

Neo4j Desktop là giao diện người dùng để vận hành cơ sở dữ liệu cục bộ. Neo4j Browser là giao diện người dùng có mục đích chung để làm việc với cơ sở dữ liệu Neo4j và là thành phần cốt lõi của Neo4j Desktop. Các nhà phát triển và nhà khoa học dữ liệu có thể sử dụng công cụ này để truy vấn, trực quan hóa, quản trị và giám sát cơ sở dữ liệu của họ. Sơ đồ trong Hình 4-1 cho thấy Trình duyệt Neo4j đang được sử dụng để chống lại biểu đồ gian lận.



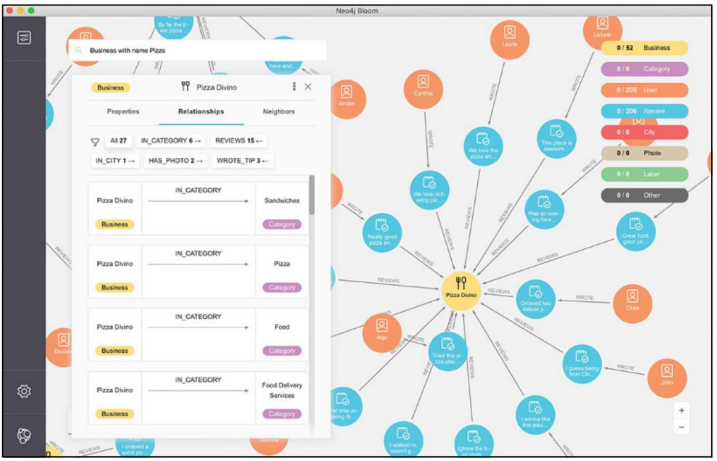
HÌNH 4-1: Trình duyệt Neo4j là một giao diện dành cho các nhà phát triển quản trị và tương tác với cơ sở dữ liệu Neo4j.

Nở hoa Neo4j

Neo4j Bloom là một công cụ khám phá và trực quan hóa biểu đồ cho phép bạn tìm các mẫu trong biểu đồ Neo4j bằng cách sử dụng mô hình tìm kiếm không dùng mã. Nó sử dụng giao diện trỏ và nhấp tương tác để mở rộng và tinh chỉnh kết quả, tìm các đường dẫn thú vị và chia sẻ thông tin chi tiết với những người khác.

Bloom được thiết kế để khám phá trực quan, đặc biệt và gõ nhanh chuyên nghiệp với các đề xuất tìm kiếm nhập trước và chỉnh sửa trực tiếp các nút và mối quan hệ. Bản trình bày trực quan có bảng phối màu, kích thước và biểu tượng linh hoạt để giúp phân biệt các mục có ảnh hưởng với kiểu dáng có thể dựa trên kết quả chạy thuật toán từ Thư viện GDS (xem phần trước trong chương này có tiêu đề “Thư viện GDS Neo4j”).

Hình 4-2 hiển thị giao diện Bloom cho một ví dụ về đánh giá nhà hàng có thể được xuất và chia sẻ.



HÌNH 4-2: Khám phá trực quan các biểu đồ Neo4j Bloom với tính năng tìm kiếm không dùng mã.

TRONG CHƯƠNG NÀY

» Chuẩn bị bộ dữ liệu tốt

» Khám phá các nhóm khả nghi

» Dự đoán kẻ lừa đảo

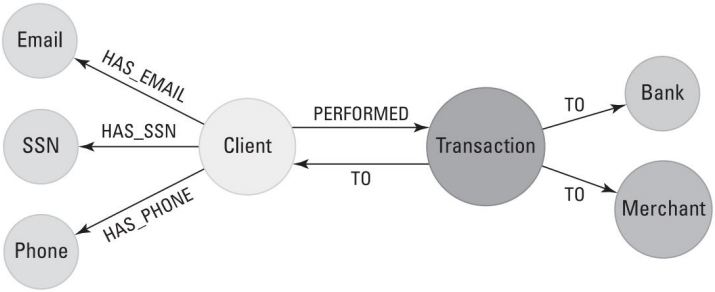
Chương 5

Phát hiện gian lận với Khoa học dữ liệu đồ thị

Trong chương này, chúng tôi hướng dẫn bạn một ví dụ về việc áp dụng kỹ thuật khoa học dữ liệu đồ thị (GDS) để điều tra và dự đoán gian lận tài chính. Sau khi chúng tôi giúp bạn làm quen với bộ dữ liệu giao dịch tài chính mẫu, chúng tôi sẽ xóa thông tin ngoại lệ có thể làm sai lệch kết quả của bạn và xác định các nhóm khách hàng đáng ngờ. Sau đó, bạn khám phá trực quan một trong các cụm cho các chỉ báo gian lận dựa trên biểu đồ và xem cách các tính năng dựa trên biểu đồ có thể giúp dự đoán hành vi gian lận trong tập dữ liệu lớn hơn.

Tìm một bộ dữ liệu gian lận tốt

Để mô phỏng bộ dữ liệu gian lận tốt, bạn muốn tạo dữ liệu tổng hợp, thực tế để mô tả các giao dịch gian lận, vì vậy, trong phần này, chúng tôi cung cấp cho bạn mô hình mạng tài chính, nơi người dùng thực hiện giao dịch với người bán và với nhau thông qua thiết bị di động. Điều này có các mẫu tương tự như các mạng thẻ tín dụng truyền thống phổ biến hơn ở Hoa Kỳ, Canada và Châu Âu. Hình 5-1 là một ví dụ về biểu đồ sử dụng một tập hợp con các nút có sẵn và các mối quan hệ vận chuyển từ dữ liệu mà chúng tôi đã sửa đổi bằng các mã định danh bổ sung.



HÌNH 5-1: Bộ dữ liệu gian lận.

Ví dụ này sử dụng các nhãn nút sau:

» Khách hàng: Những người có thông tin nhận dạng cá nhân (PII) chẳng hạn như Số An sinh Xã hội (SSN), số điện thoại và địa chỉ email

» Mules: Những khách hàng được cho là đã chuyển tiền lừa đảo

» PII của khách hàng:

- SSN
- Điện thoại: Số điện thoại
- Email: Địa chỉ email

Các nút này được kết nối bởi các loại mối quan hệ sau:

» (Khách hàng)-[:HAS_PHONE]->(Điện thoại)

» (Khách hàng)-[:HAS_SSN]->(SSN)

» (Khách hàng)-[:HAS_EMAIL]->(Email)

REMEMBER

Phân tích được thực hiện ở đây tập trung vào thông tin trên, nhưng bộ dữ liệu cũng chứa thông tin bổ sung, chẳng hạn như các giao dịch được thực hiện cho ngân hàng, người bán và khách hàng.

Loại bỏ ngoại lệ

Bước đầu tiên quan trọng khi thực hiện phân tích gian lận là kiểm tra chất lượng của dữ liệu. Trong bộ dữ liệu gian lận, bạn có thể có các giá trị ngoại lệ không liên quan đến phân tích của mình. Ngoại lệ là những sự kiện hiếm gặp hoặc

các mục gây nghi ngờ bằng cách khác biệt đáng kể so với phần lớn dữ liệu. Các ngoại lệ trong biểu đồ dựa trên khả năng kết nối của chúng và cấu trúc liên kết của biểu đồ thay vì giá trị thuộc tính.

TIP

Thuật toán Độ trung tâm đo lường số lượng mối quan hệ mà một nút có. Do đó, chạy thuật toán Mức độ trung tâm là một cách tốt để tìm ra các ngoại lệ tiềm năng.

Bởi vì bạn mong đợi các loại nút khác nhau có khả năng kết nối khác nhau, nên bạn cần chiếu một biểu đồ bao gồm một loại nút duy nhất để bạn có thể tìm kiếm các ngoại lệ liên quan đến một loại nút tại một thời điểm. Ví dụ, mỗi ngân hàng sẽ nhận được nhiều khoản tiền gửi (vì vậy nó sẽ có mức độ tập trung cao), nhưng một chủ tài khoản cá nhân sẽ nhận được ít hơn nhiều. Bạn muốn so sánh ngân hàng với ngân hàng và tài khoản với tài khoản.

Để tìm các ngoại lệ tiềm năng, bạn có thể chạy thuật toán Mức độ trung tâm đối với tập dữ liệu gian lận với truy vấn sau:

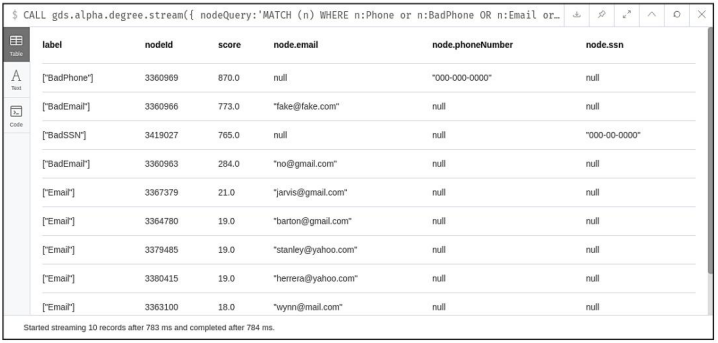
```
GỌI gds.alpha.degree.stream({

   nodeQuery:'MATCH (n) WHERE n:Phone OR n:Email OR n:SSN RETURN
    id(n) as id',
   mối quan hệTruy vấn:'MATCH (n1)<-
    [:HAS_PHONE|HAS_EMAIL|HAS_SSN]-(c:Client),

    (n2)<-[:HAS_PHONE|HAS_EMAIL|HAS_SSN]->(c)
                            RETURN id(n1) làm nguồn,
    id(n2) làm mục tiêu'
})
YIELD nodeId, điểm số
VỚI nút gds.util.asNode(nodeId) AS, nútId,
    điểm
TRẢ LẠI nhãn (nút) dưới dạng nhãn, nútId, điểm, nút.
    email, nút.phoneNumber, nút.ssn
ĐẶT HÀNG THEO ĐIỂM DESC
GIỚI HẠN 10
```

Khi bạn thực hiện truy vấn này, bạn sẽ nhận được nhiều kết nối đến số nhận dạng giả mạo. Đầu ra được hiển thị trong Hình 5-2. Xem phụ lục để có cái nhìn đầy đủ về hình này.

Bốn ngoại lệ lớn thể hiện điểm số cao (cột ba) cho số lượng kết nối. Những ngoại lệ này đại diện cho các nút có tài khoản email, SSN và số điện thoại giả mạo.



label	nodeId	score	node.email	node.phoneNumber	node.ssn
["BadPhone"]	3360969	870.0	null	"000-000-0000"	null
["BadEmail"]	3360966	773.0	"fake@fake.com"	null	null
["BadSSN"]	3419027	765.0	null	null	"000-00-0000"
["BadEmail"]	3360963	284.0	"no@gmail.com"	null	null
["Email"]	3367379	21.0	"jarvis@gmail.com"	null	null
["Email"]	3364780	19.0	"barton@gmail.com"	null	null
["Email"]	3379485	19.0	"stanley@yahoo.com"	null	null
["Email"]	3380415	19.0	"herrera@yahoo.com"	null	null
["Email"]	3363100	18.0	"wynn@mail.com"	null	null

Started streaming 10 records after 783 ms and completed after 784 ms.

HÌNH 5-2: Kết quả của thuật toán Degree Centrality.

TIP

Loại trừ các nút kết quả giả mạo này khỏi phân tích của bạn vì nhiều khả năng những người này đã chọn không điền thông tin của biểu mẫu một cách chính xác thay vì họ đại diện cho hoạt động gian lận. Nếu không bị loại trừ, bạn sẽ tìm thấy nhiều thông tin xác thực sai dựa trên việc mọi người chia sẻ thông tin bổ sung không có thật phổ biến, chẳng hạn như email "fake@fake.com."

Tiếp theo, hãy cập nhật nhãn trên các nút này để dễ dàng loại trừ chúng khỏi phân tích trong tương lai. Các truy vấn sau đây sẽ xóa nhãn gốc trong khi thêm nhãn "Xấu" mới:

```
MATCH (n:Email)
Ở ĐÂU n.email='fake@fake.com' hoặc n.email='no@
    gmail.com'
SET n:BadEmail XÓA n:Email;
TRẬN ĐẤU (n:SSN)
WHERE n.ssn='000-00-0000'
ĐẶT n:BadSSN XÓA n:SSN;
MATCH (n:Điện thoại)
WHERE n.phoneNumber='000-000-0000'
SET n:BadPhone REMOVE n:Phone;
```

Tìm các cụm đáng ngờ

Bạn muốn tìm một số kẻ lừa đảo thực sự? Bây giờ là thời gian của bạn! Trong trường hợp gian lận của bên thứ nhất, các tài khoản giả mạo được tạo ra mà không có ý định hoàn trả các khoản vay hoặc nợ. Một cách phổ biến để tìm ra những kẻ giả mạo này là tìm những tài khoản chia sẻ thông tin nhận dạng, như SSN, số điện thoại và địa chỉ email.

Quần đảo của các nút tương tác có ít kết nối với biểu đồ lớn hơn không đại diện cho hành vi tài chính điển hình.
Bạn có thể sử dụng thông tin này và thuật toán Thành phần được kết nối yếu để tìm các sơ đồ con rời rạc chia sẻ các mã định danh chung một cách đáng ngờ.

Thuật toán Các thành phần được kết nối yếu là một thuật toán phát hiện cộng đồng tìm các tập hợp các nút được kết nối trong một biểu đồ không có hướng trong đó mỗi nút có thể truy cập được từ bất kỳ nút nào khác trong cùng một tập hợp.

Truy vấn sau đây chạy các Thành phần được kết nối yếu trên biểu đồ máy khách được chiếu:

```
GỌI gds.wcc.write({
                nodeQuery:'MATCH (c:Client) TRẢ LẠI id(c) dưới
    dạng id',
mối quan hệTruy vấn:'MATCH
    (c1:Client)-[:HAS_PHONE|HAS_EMAIL|HAS_SSN]-
    >(trung cấp)<-[:HAS_PHONE|HAS_EMAIL|HAS_SSN]-
    (c2:Khách hàng)
WHERE không (trung gian: BadSSN)
VÀ không (trung gian:BadEmail)
VÀ không (trung gian: BadPhone)
TRẢ LẠI id(c1) làm nguồn, id(c2) làm đích',
writeProperty:'thành phầnId'
});
```

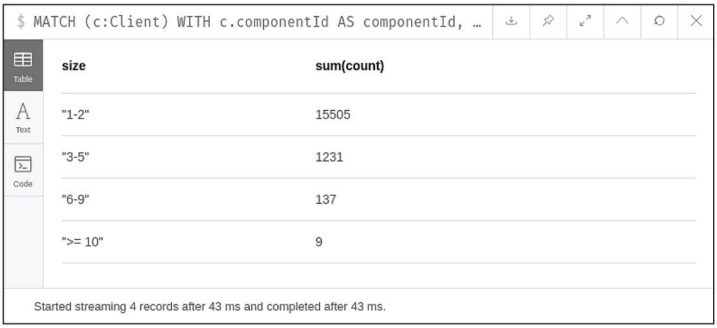
Trong phần “Xóa điểm ngoại lệ” ở đầu chương này, bạn đã giả định rằng nhiều kẻ không lừa đảo sử dụng thông tin dạng giả mạo tương tự, vì vậy, bây giờ khi bạn đang xây dựng biểu đồ dự kiến đó, bạn cần loại trừ các SSN, địa chỉ email và số điện thoại xấu. số được xác định trong phần trước. Nếu không, bạn sẽ chỉ có một cụm lớn do tính phổ biến của dữ liệu biểu mẫu không có thật.

Thuật toán Thành phần được kết nối yếu này so khớp các ứng dụng khách chia sẻ email, số điện thoại hoặc SSN và gán nhãn cho thuộc tính componentId cho mỗi nút ứng dụng khách. Các nút có cùng giá trị ID thành phần được coi là trong cùng một cụm.

Sau đó, bạn có thể sử dụng truy vấn sau để xem phân phối kích thước cụm được thuật toán này trả về:

```
MATCH (c:Khách hàng)
VỚI c.componentId NHƯ componentId, đếm (*) NHƯ
    kích cỡ
VỚI kích thước, số lượng (*) AS số lượng
TRẢ LẠI TRƯỜNG HỢP KHI 1 <= size <= 2 THÌ "1-2"
                  KHI 3 <= kích thước <= 5 THÌ "3-5"
                  KHI 6 <= size <= 9 THÌ "6-9"
                  ELSE ">= 10" END AS kích thước,
          tổng (đếm)
ĐẶT HÀNG THEO kích thước
```

Kết quả của truy vấn này cho thấy hầu hết các máy khách nằm trong các cụm nhỏ chỉ có 1 hoặc 2 máy khách và được minh họa trong Hình 5-3. Xem phụ lục để có cái nhìn đầy đủ về hình này.



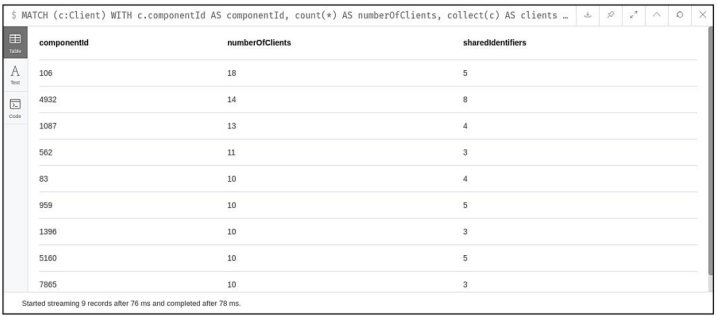
HÌNH 5-3: Hầu hết khách hàng nằm trong các nhóm nhỏ.

Phần lớn khách hàng thuộc cụm 1 hoặc 2 khách hàng, nhiều hơn so với khách hàng thuộc cụm 3 đến 5 và 6 đến 9, và sau đó rất ít khách hàng thuộc cụm 10 khách hàng hoặc lớn hơn. Bạn có thể thấy điều này; hầu hết mọi người không chia sẻ số nhận dạng cá nhân, nhưng có chín cụm với mười khách hàng trở lên chia sẻ ít nhất một số nhận dạng. Nếu bạn chỉ phóng to những cụm có mười khách hàng trở lên và khám phá thêm những cụm đó, bạn sẽ nhận được truy vấn sau:

```
MATCH (c:Khách hàng)
VỚI c.componentId AS componentId, đếm (*) AS numberOfClients, thu
    thập (c) AS client
WHERE numberOfClients >= 10
```

```
VỚI componentId, numberOfClients,
       // Tìm tất cả các định danh của khách hàng trong một
   cụm
       apoc.coll.toSet(apoc.coll.flatten(
          [khách hàng trong khách hàng | [(khách hàng)-[:HAS_
   SSN|HAS_EMAIL|HAS_PHONE]->(id) | id]])) AS id,
       khách hàng
trả về componentId, numberOfClients,
          // Tìm xem có bao nhiêu định danh được chia sẻ


          // Chỉ trả về các định danh được chia sẻ bởi > 1
   Khách hàng trong cụm
          kích thước ([bản ghi trong [id trong id | {
             tôi đã làm,
             sharedClients: size([(id)<-
   (khách hàng:Khách hàng) Ở đâu khách hàng trong khách hàng |
   khách hàng])
          }] WHERE record.sharedClients > 1 | bản ghi]) AS
   sharedIdentifiers
ĐẶT HÀNG THEO numberOfClients DESC
```



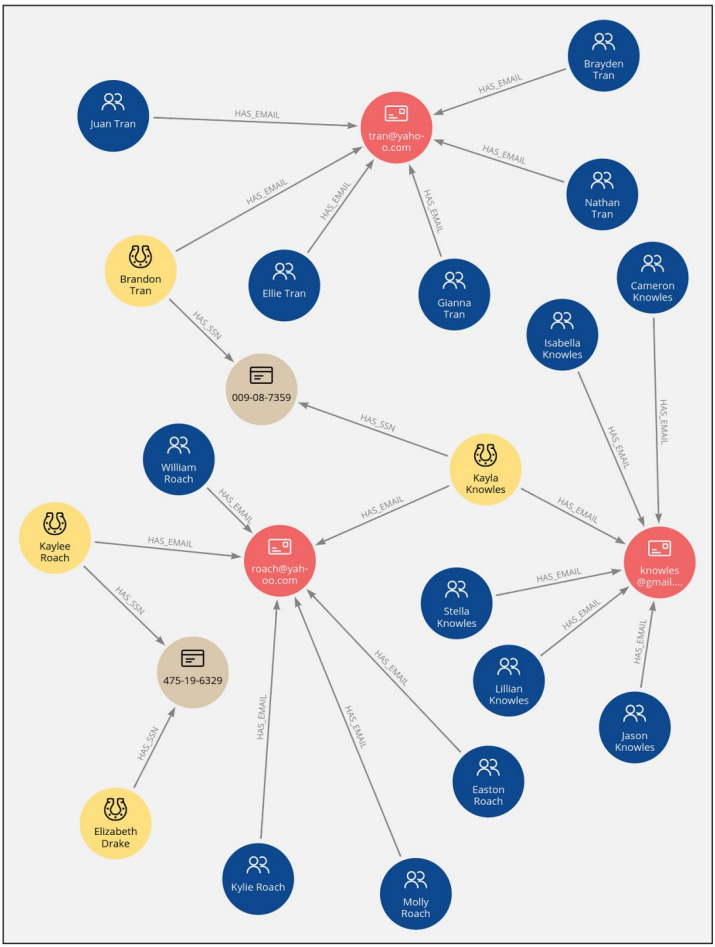
HÌNH 5-4: Kết quả của các cụm có mười khách hàng trở lên.

Các kết quả truy vấn này được hiển thị trong Hình 5-4. Xem phụ lục để có cái nhìn đầy đủ về hình này.

Cụm 106 có vẻ là một cụm thú vị để khám phá thêm vì cụm này có một số lượng lớn khách hàng và năm mã định danh được chia sẻ giữa họ. Trong phần tiếp theo, bạn sẽ điều tra trực quan cụm này.

Khám phá trực quan một cụm đáng ngờ

Khám phá cụm 106 trong một công cụ như Neo4j Bloom, mà chúng tôi đề cập trong Chương 4, có thể giúp bạn hiểu rõ hơn về nhóm này. Bạn có thể hình dung mối quan hệ giữa các khách hàng trong cụm đó bằng cụm từ tìm kiếm Bloom. Cụm từ tìm kiếm Bloom là một cách mà bạn có thể xác định cấu trúc ngôn ngữ tự nhiên thực thi truy vấn đối với cơ sở dữ liệu cho bạn. Cụm từ tìm kiếm "khám phá cụm 106" tìm các mối quan hệ trong Hình 5-5 giữa các khách hàng trong cụm 106.



HÌNH 5-5: Biểu đồ kết quả của cụm từ tìm kiếm "khám phá cụm 106."

Các nút có biểu tượng móng ngựa đại diện cho con la, các nút có biểu tượng người là khách hàng, các nút có biểu tượng thư là địa chỉ email và các nút khác là SSN.

Trong cụm này, bạn có bốn con la và bạn cũng có thể thấy ba địa chỉ email được chia sẻ bởi 13 khách hàng. Tại thời điểm này, bạn có thể muốn gửi danh sách những người trong cụm này cho một chuyên gia tên miền để khám phá thêm.

Từ hình ảnh trực quan này, bạn có thể thấy rằng hầu hết khách hàng trong cụm này chỉ chia sẻ ba tài khoản email. Chúng ta có thể tưởng tượng một vài người chia sẻ địa chỉ email nhưng có nhiều hơn thế có thể là điều cần khám phá thêm.

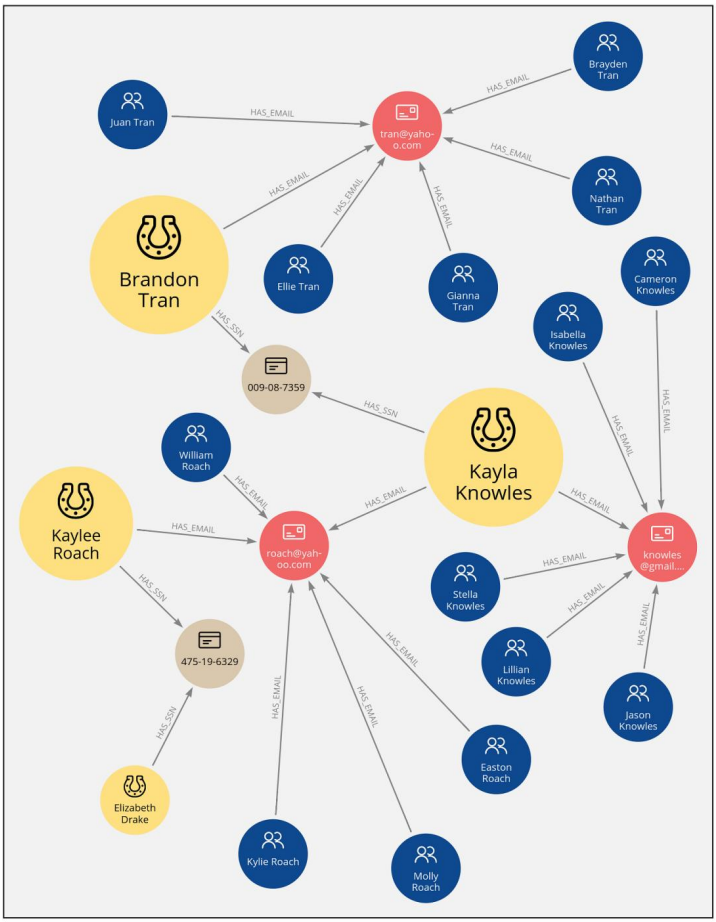
Trong cụm, một số nút này có vẻ quan trọng hơn, đóng vai trò là cầu nối cục bộ giữa các máy khách ở các khu vực khác nhau của Hình 5-5. Sau đó, bạn có thể sử dụng thuật toán Betweenness Centrality để xác nhận những nghi ngờ của mình.

Thuật toán Betweenness Centrality ước tính đường đi ngắn nhất giữa mỗi cặp nút và sau đó mỗi nút nhận được một điểm, dựa trên số đường đi ngắn nhất đi qua nút. Các nút thường xuyên nằm trên các đường đi ngắn nhất này sẽ có điểm trung tâm giữa các điểm cao hơn.

Chạy thuật toán này bằng cách thực hiện truy vấn sau:

```
GỌI gds.betweenness.write({
   nodeQuery: 'MATCH (c:Client) WHERE
    c.componentId=106 TRẢ LẠI id(c) dưới dạng id',
   mối quan hệTruy vấn: 'MATCH
    (c1:Client)-[:HAS_PHONE|HAS_EMAIL|HAS_SSN]-
    >(trung cấp)<-[:HAS_PHONE|HAS_EMAIL|HAS_SSN]-
    (c2:Khách hàng)
WHERE không (trung gian: BadSSN)
VÀ không (trung gian:BadEmail)
VÀ không (trung gian: BadPhone)
TRẢ LẠI id(c1) làm nguồn, id(c2) làm đích',
writeProperty:'betweennessCentrality'
})
```

Truy vấn này lưu trữ điểm trung tâm giữa trong thuộc tính giữa nessCentrality trên mỗi nút máy khách cho cụm này. Sau đó, bạn có thể cập nhật các quy tắc tạo kiểu trong Neo4j Bloom để kiểm tra kết quả trong Hình 5-6.

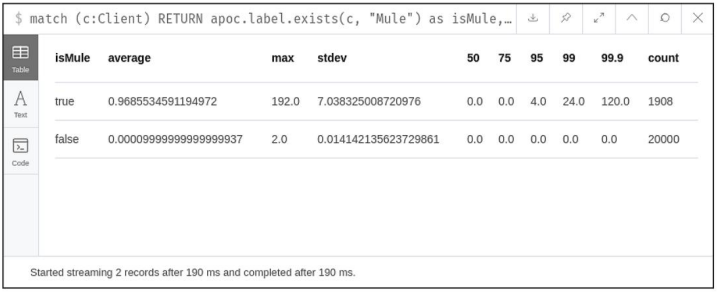


HÌNH 5-6: Kết quả của việc sử dụng điểm số Độ trung tâm cho kích thước nút trong Neo4j Bloom.

Các nút lớn nhất là các nút có ảnh hưởng nhất trong cụm.
Các nút này đại diện cho các con la được biết là có hành vi gian lận.
Tại thời điểm này, bạn đã xác định được các hành vi và cụm đáng ngờ.
Sau khi các nhà phân tích gian lận của bạn xác nhận hoạt động bất chính có thể xảy ra
này, bạn có thể sử dụng thông tin này để dự đoán các con la trong tập dữ liệu lớn hơn.

Dự đoán kẻ lừa đảo bằng đồ thị Đặc trưng

Trong tập dữ liệu thực, bạn sẽ không thực sự biết ai là con la, nhưng trong tập dữ liệu chúng tôi sử dụng, chúng được xác định. Việc nhận dạng này cho phép bạn kiểm tra dự đoán của mình rằng điểm số trung tâm giữa các mức độ trung bình cao hơn có thể dự đoán gian lận bằng cách sử dụng toàn bộ biểu đồ. Kiểm tra nhanh lý thuyết của bạn cho thấy rằng các khách hàng có nhãn la có trung bình 0,9685 điểm trung tâm giữa các điểm, cao hơn đáng kể so với các điểm không phải la như trong Hình 5-7.
Xem phụ lục để có cái nhìn đầy đủ về hình này.



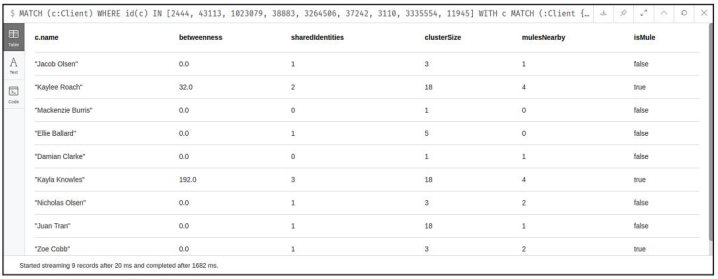
HÌNH 5-7: So sánh điểm giữa Độ trung tâm cho toàn bộ biểu đồ.

Mặc dù đây là một chỉ số quan trọng, nhưng độ lệch và phân phối điểm số có nghĩa là có sự chồng chéo có thể dẫn đến kết quả dương tính và âm tính giả. Trong tình huống này, bạn muốn kết hợp điểm trung tâm giữa mức độ trung bình này với các yếu tố dự đoán khác và làm việc với nhà khoa học dữ liệu để tạo mô hình ML.

Một kịch bản ML mà bạn có thể sử dụng là một cách tiếp cận trích xuất các tính năng biểu đồ để sử dụng trong bộ phân loại nhị phân để dự đoán các con la. Ví dụ về các tính năng đồ thị bao gồm

- Điểm trung tâm giữa
- Số lượng khách hàng chia sẻ định danh
- Trọng số của định danh dùng chung
- Số lượng con la đã biết trong <n> bước nhảy
- Kích thước của cụm

Hình 5-8 cho thấy một số tính năng biểu đồ mà chúng ta có thể trích xuất cho những người khác nhau trong biểu đồ. Xem phụ lục để có cái nhìn đầy đủ về hình này.



$ MATCH (c:Client) WHERE id(c) IN [2444, 43113, 1023079, 38883, 3264506, 37242, 3110, 3335554, 11945] WITH c MATCH (:Client {...

c.name	betweenness	sharedIdentities	clusterSize	mulesNearby	isMule
"Jacob Olsen"	0.0	1	3	1	false
"Kaylee Roach"	32.0	2	18	4	true
"Mackenzie Burris"	0.0	0	1	0	false
"Ellie Ballard"	0.0	1	5	0	false
"Damian Clarke"	0.0	0	1	1	false
"Kayla Knowles"	192.0	3	18	4	true
"Nicholas Olsen"	0.0	1	3	2	false
"Juan Tran"	0.0	1	18	1	false
"Zoe Cobb"	0.0	1	3	2	true

Started streaming 9 records after 20 ms and completed after 1682 ms.

HÌNH 5-8: Một ma trận các tính năng được thiết kế bằng đồ thị và phân loại con la.

Các tính năng này có thể được trích xuất thành định dạng bảng để đào tạo mô hình ML.

REMEMBER

Sau khi hài lòng với mô hình phát hiện gian lận của mình, bạn có thể sử dụng mô hình này trong sản xuất để xác định các bộ đếm khác khi biểu đồ của bạn phát triển. Khi thông tin mới được thêm vào biểu đồ trong thế giới thực, thông thường sẽ lặp lại quy trình này và tạo các tính năng biểu đồ mới cũng như mô hình cập nhật.

TRONG CHƯƠNG NÀY

Chương 6

Mười mẹo với tài nguyên để vẽ đồ thị thành công Khoa học dữ liệu

Nếu bạn mới bắt đầu với khoa học dữ liệu đồ thị (GDS), chương này có thể giúp ích. Chúng tôi cung cấp cho bạn một số tài nguyên Neo4j để hướng dẫn bạn biết thêm thông tin và để giúp bạn khám phá cơ hội cho dự án của mình cũng như chuyển tiếp thành công từ ý tưởng sang sản xuất, chúng tôi bao gồm mười mẹo sau:

» Điều tra các trường hợp sử dụng và cảm thấy thoải mái với lừa đảo chấp nhận. Bởi vì công nghệ đồ thị được áp dụng trong các ngành và trong nhiều trường hợp sử dụng khác nhau nên khó có thể biết bắt đầu từ đâu. Để mở rộng kiến thức của bạn và giúp bạn cảm thấy thoải mái với GDS, hãy xem lại các ví dụ sau:

- Xem xét các trường hợp sử dụng. Bắt kịp tốc độ giải quyết các vấn đề công nghệ đồ thị có thể giải quyết. Truy cập neo4j.com/use-cases để đọc một số trường hợp sử dụng.
- Xem các cuộc nói chuyện. Tìm hiểu cách mọi người sử dụng GDS. Đồng hồ các bài thuyết trình từ sự kiện kỹ thuật số Kết nối cho GDS của Neo4j: go.neo4j.com/connections-graph-data science-lp.html.

- Mở rộng kiến thức của bạn về các khái niệm chính. Ôn tập tài liệu đặt GDS trong một bối cảnh lớn hơn. Truy cập neo4j.com/whitepapers/ công nghệ đồ thị trí tuệ nhân tạo để biết cách đồ thị nâng cao AI.

» Xác định và thu hút đội ngũ mũi nhọn. Sử dụng đồ thị công nghệ trong sản xuất có thể mới đối với nhiều người, vì vậy đừng mong đợi các nhóm hiểu cách đánh giá hoặc so sánh các tùy chọn biểu đồ với các giải pháp khác. Tập hợp một nhóm nhỏ có thể trở thành chuyên gia của bạn trong việc chuyển các nhu cầu kinh doanh thành các yêu cầu kỹ thuật và ứng dụng GDS. Đảm bảo có đại diện từ các tổ chức chính, bao gồm các nhóm kinh doanh, CNTT và khoa học dữ liệu.


TIP

Cung cấp thêm thông tin kỹ thuật cho các nhà phát triển và nhà khoa học dữ liệu của bạn. Nhóm của bạn có thể sẽ cần thời gian để làm quen với công nghệ, vì vậy hãy tìm kiếm các tài nguyên cho phép bắt đầu dễ dàng. Một số ví dụ bao gồm

- neo4j.com/graph-algorithms-book
- neo4j.com/graph-databases-book
- neo4j.com/graph-databases-for-dummies
- neo4j.com/sandbox

» Đánh giá vấn đề "đồ họa" của bạn. Công nghệ đồ thị hữu ích ở bất cứ nơi nào bạn có nhiều thông tin kết nối, phụ thuộc lẫn nhau. Nhưng tại một thời điểm nào đó, bạn cần xem xét nên tập trung vào lĩnh vực kinh doanh nào và loại dự án nào để bắt đầu.


TIP

Bắt đầu với sự giao thoa ý tưởng giữa người dùng, doanh nghiệp và công nghệ. Cân nhắc tổ chức các phiên đổi mới ảo hoặc ngoại vi với nhóm đa chức năng của bạn để xác định nhu cầu của các bên liên quan, tạo các câu hỏi liên quan đến kết nối, các giải pháp khả thi trên bảng phân cảnh cũng như xác định các thách thức và cơ hội chính. Sự hợp tác này có thể dẫn đến một nguyên mẫu mà bạn có thể chia sẻ với các giám đốc điều hành để nhận phản hồi, nhưng mục tiêu là khám phá các trường hợp sử dụng mục tiêu đầy hứa hẹn.

» Đánh giá hiện trạng. Sau khi bạn có một trường hợp sử dụng mục tiêu lưu ý, hãy bắt đầu bằng việc ghi lại trạng thái hiện tại của bạn. Xem xét các vấn đề hiện có cũng như các bộ phận khác nhau trong tổ chức của bạn sẽ có những trải nghiệm và vấn đề khác nhau như thế nào. Tìm hiểu xem các nhà tài trợ doanh nghiệp của bạn xem trường hợp sử dụng này như thế nào và bất kỳ vấn đề hoặc cơ hội nào. Hãy càng cụ thể càng tốt. Ví dụ: tác động đối với mỗi khách hàng của việc cải thiện

hồ sơ trực tuyến? Ý nghĩa doanh thu của việc tăng một nửa phần trăm gian lận được phục hồi là gì? Ngoài ra, hãy nhớ xem xét các yếu tố thị trường bên ngoài như khách hàng hoặc tăng trưởng giao dịch, các yếu tố cạnh tranh, cơ hội mới nổi như nền tảng giao hàng mới hoặc cơ hội sản xuất.

» Bản đồ giá trị của trạng thái được đề xuất. Mặc dù dự án biểu đồ đầu tiên của bạn có thể tạo ra nhiều ý tưởng mới và các dự án trong tương lai, nhưng hãy lập bản đồ rõ ràng và trực tiếp các đặc điểm của dự án biểu đồ ngắn hạn với các giá trị kinh doanh. Xem xét trạng thái hiện tại và các điểm khó khăn cũng như cách trường hợp sử dụng mục tiêu biểu đồ của bạn có thể giúp giải quyết các vấn đề kinh doanh như tiết kiệm chi phí, tăng doanh thu, cơ hội thị trường mới, thời gian đưa ra thị trường, giảm thiểu rủi ro, v.v. Ví dụ: khám phá các hành trình tương tự của khách hàng và sử dụng thông tin đó trong mô hình máy học (ML) có thể tăng độ chính xác của dự đoán rời bỏ để doanh nghiệp có thể thực hiện hành động phòng ngừa sớm và giảm tổn thất doanh thu.

» Đo lường ROI. Đối với từng lĩnh vực giá trị của bạn, hãy xác định cách bạn dự định đo lường lợi tức đầu tư (ROI) hoặc thành công của mình. Ví dụ: bạn sẽ sử dụng độ chính xác dự đoán hoặc giảm tổn thất tài chính để ước tính tác động của trạng thái kết thúc của mình chứ? So sánh chi phí mềm và chi phí cứng của việc duy trì các quy trình hiện có với dự án đồ thị của bạn. Nếu bạn không thể kiểm tra trạng thái hiện tại của mình, hãy thận trọng hơn khi ước tính các cơ hội tiết kiệm hoặc doanh thu gia tăng. Tương tự như vậy, có thể khó đo lường giá trị của các khả năng mới, chẳng hạn như trả lời các câu hỏi khó trước đây, vì vậy bạn có thể cần phải sáng tạo hoặc thêm phân tích định tính.

REMEMBER

» Gắn kết các bên liên quan. Cuối cùng, bạn cần có sự thống nhất giữa các chức năng về các mục tiêu và yêu cầu của dự án đồ thị của bạn. Quá trình này lặp đi lặp lại, không phải là thứ bạn giải quyết tại một thời điểm. Các nhóm khác nhau có thể có quan điểm khác nhau về tầm nhìn dự án, ROI chính và thậm chí cả vai trò của công nghệ đồ thị. Có được sự thống nhất về các mục tiêu của dự án và cách đo lường thành công là điều cần thiết — và bạn có thể muốn xem xét một quy trình để giải quyết các ý kiến xung đột hoặc bất đồng.

» Phê duyệt dự án của bạn. Việc tận dụng các công nghệ mới như GDS yêu cầu các bên liên quan và người phê duyệt của bạn cảm thấy thoải mái khi thử điều gì đó không quen thuộc, vì vậy, công việc của bạn là nhấm mục tiêu đúng trường hợp sử dụng, giá trị bản đồ và

ước tính ROI cần kết hợp với nhau trong một câu chuyện ngắn gọn phù hợp với động lực của công ty bạn.


TIP

Ghi lại các giả định của các bên liên quan về giá trị kinh doanh. Ví dụ: sự rời bỏ của khách hàng có thể là một vấn đề, nhưng nó có phải là ưu tiên hàng đầu không và tại sao? Bạn có thể được hỏi về bối cảnh cạnh tranh cũng như các lựa chọn thay thế và chi phí hoặc cơ hội bị mất nếu bạn không tiếp tục. Ghi lại rõ ràng các điểm tiếp xúc hệ thống phụ thuộc lẫn nhau là một phần của các quy trình hiện tại và tác động của giải pháp biểu đồ của bạn.

» Tiến hành POC và lập kế hoạch sản xuất. Các dự án lớn hơn, đặc biệt nếu công nghệ là mới đối với một nhóm, thường yêu cầu bằng chứng về khái niệm (POC) trước khi phê duyệt và triển khai. POC có thể chuẩn bị cho nhóm của bạn sản xuất và xác định bất kỳ lỗ hổng nào. Quá trình này có thể liên quan đến việc lặp lại các nguyên mẫu trước đó trước khi bạn chuyển sang mô hình hóa dữ liệu và thử nghiệm các quy trình công việc cụ thể.


REMEMBER

Trong GDS, các lựa chọn thuật toán và mô hình dữ liệu của bạn phụ thuộc nhiều vào các câu hỏi mà bạn đang cố gắng trả lời. Các nhà khoa học dữ liệu và chuyên gia về chủ đề của bạn nên tham gia để đảm bảo đưa ra các giả định đúng. Ngoài ra, hãy đảm bảo rằng các nhóm CNTT của bạn tham gia để đưa ra bất kỳ cảnh báo nguy hiểm nào và người dùng cuối của bạn luôn sẵn sàng đánh giá mọi mối lo ngại về khả năng sử dụng.


TIP

Các nhà cung cấp dịch vụ POC có thể giúp đẩy nhanh dự án của bạn bằng trải nghiệm đồ thị của họ. Truy cập neo4j.com/ những dịch vụ chuyên nghiệp để biết thêm thông tin.

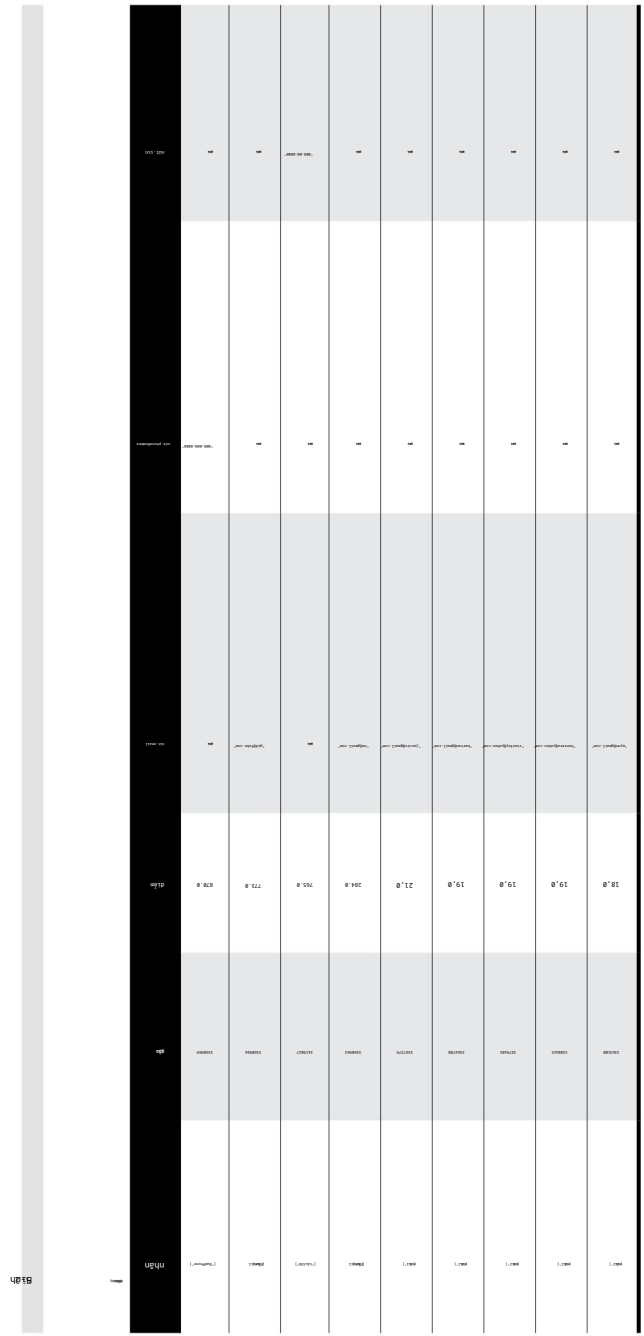
» Kết nối và tiếp tục cuộc hành trình của bạn. Áp dụng GDS là một hành trình. Bạn có thể bắt đầu với một dự án tập trung và thấy mình đang trả lời những câu hỏi mà bạn chưa từng biết. Chúng tôi thực sự khuyên nhóm của bạn nên kết nối và tương tác với cộng đồng đồ thị. Cộng đồng đồ thị bao gồm các nhóm người dùng tích cực chia sẻ ý tưởng mới và trợ giúp với các câu hỏi cụ thể và đôi khi bất thường. Việc tham gia vào một cộng đồng năng động phong phú với sự hỗ trợ về giáo dục và các chứng chỉ sẽ giúp nhóm của bạn thành công với dự án biểu đồ đầu tiên và mở rộng giá trị của các biểu đồ của bạn theo thời gian.


TIP

Ghé thăm cộng đồng Neo4j tại community.neo4j.com, và xem tài nguyên của nó tại neo4j.com/graphacademy.

ruột thừa

Trong phụ lục này, chúng tôi đã định dạng một số hình trong Chương 5 thành các bảng đầy đủ, để bạn có thể xem chi tiết rõ hơn trong từng hình.



Hình 5-3

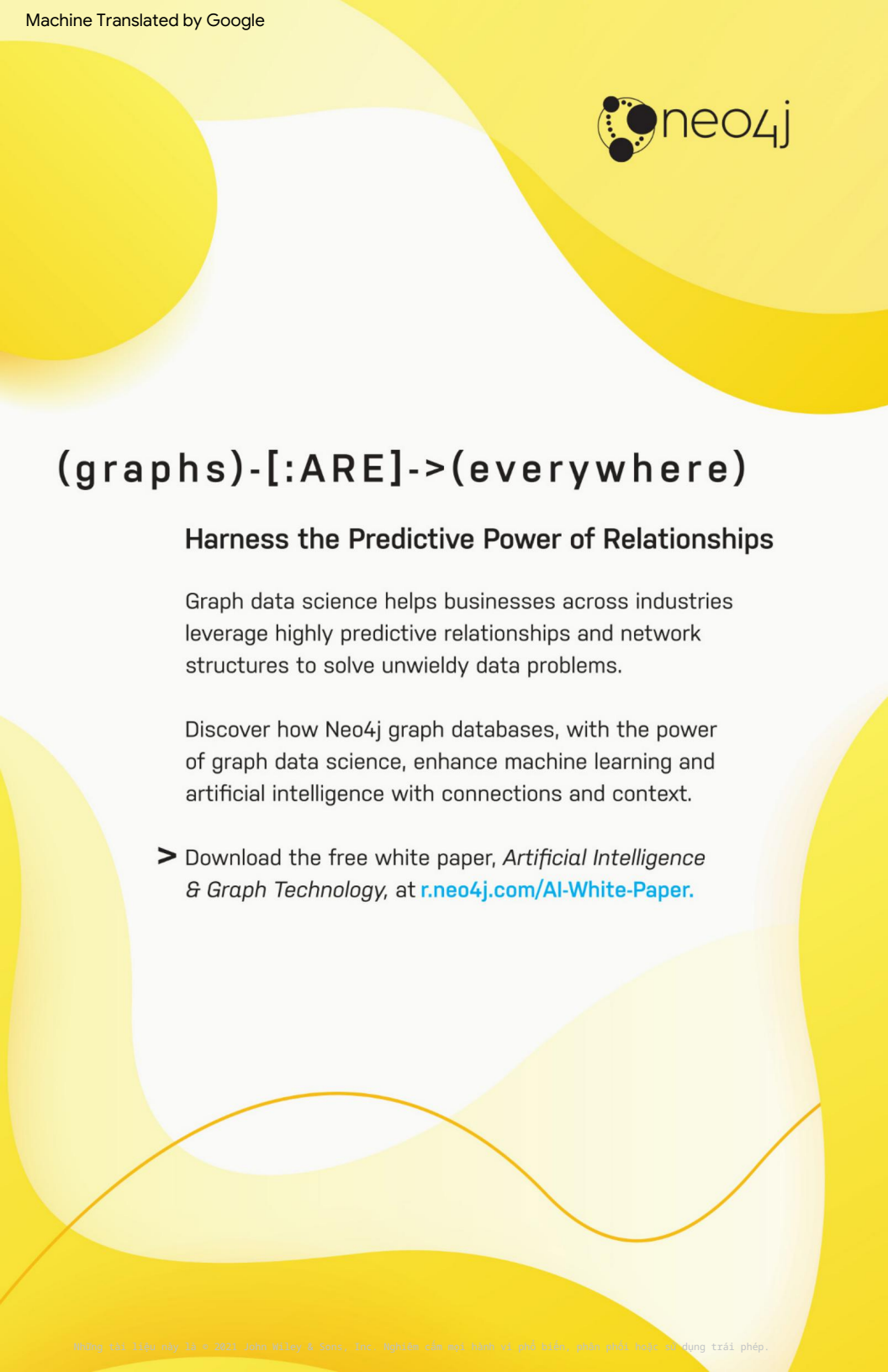
Hầu hết các khách hàng đều ở trong các cụm nhỏ

kích cỡ	tổng (đếm)
“1-2”	15505
“3-5”	1231
“6-9”	137
“>=10”	9

Hình 5-4

Kết quả của các cụm có mười khách hàng trở lên

thành phần	số lượng khách hàng	định danh được chia sẻ
106	18	5
4932	14	“
1087	13	4
562	11	3
83	10	4
959	10	5
1396	10	3
5160	10	5
7865	10	3



Improve predictions with graphs

Graph Data Science (GDS) For Dummies, Neo4j Special Edition, focuses on the applications of graph analysis and graph-enhanced machine learning, which both take the form of GDS. You discover the GDS basics and learn about its adoption. We use the Neo4j database technology to help illustrate our points about the GDS platform. We also supply you with plenty of resources to guide you outside of what this introductory book provides you.

Inside...

- Understanding graph analytics and GDS
- Using questions to explore GDS
- Put graphs to work in the real world
- Applying GDS technology
- Tips and resources for successful GDS

Amy Hodler is a network science enthusiast and AI and graph analytics program director at Neo4j. She promotes using graphs for predicting complex behavior. **Mark Needham** is a graph advocate and developer relations engineer at Neo4j. He helps users embrace graphs with sophisticated data solutions.

Cover image provided by Neo4j, Inc.

ISBN: 978-1-119-74604-1
Not For Resale

for dummies®
A Wiley Brand

THỎA THUẬN GIẤY PHÉP NGƯỜI DÙNG CUỐI WILEY

Truy cập www.wiley.com/go/eula để truy cập EULA sách điện tử của Wiley.