neo4j

HELPING THE WORLD MAKE SENSE OF DATA

GRAPH DATA SCIENCE USE CASES: ENTITY RESOLUTION

Zachary Blumenfeld

Product Specialist, Graph Data Science

Jaimie Chung

Product Manager, Graph Data Science

Problem

Whether you're a retailer, media conglomerate, or bank, having a consistent and accurate view of the people, places, and organizations in your data is central to understanding your business. Companies contend with multiple data streams of varying quality and have to consolidate and disentangle the data in order to glean actionable insights.

Entity resolution, or disambiguation, is a widely applicable approach to resolve data into unique and valuable entity profiles. Without this crucial process, organizations are left making key decisions based on incomplete, misleading data.

This paper offers a brief overview of ways to perform entity resolution using Neo4j's Graph Data Science framework, which includes the Neo4j Graph Database for graph persistence and the Neo4j Graph Data Science Library for graph analytics.

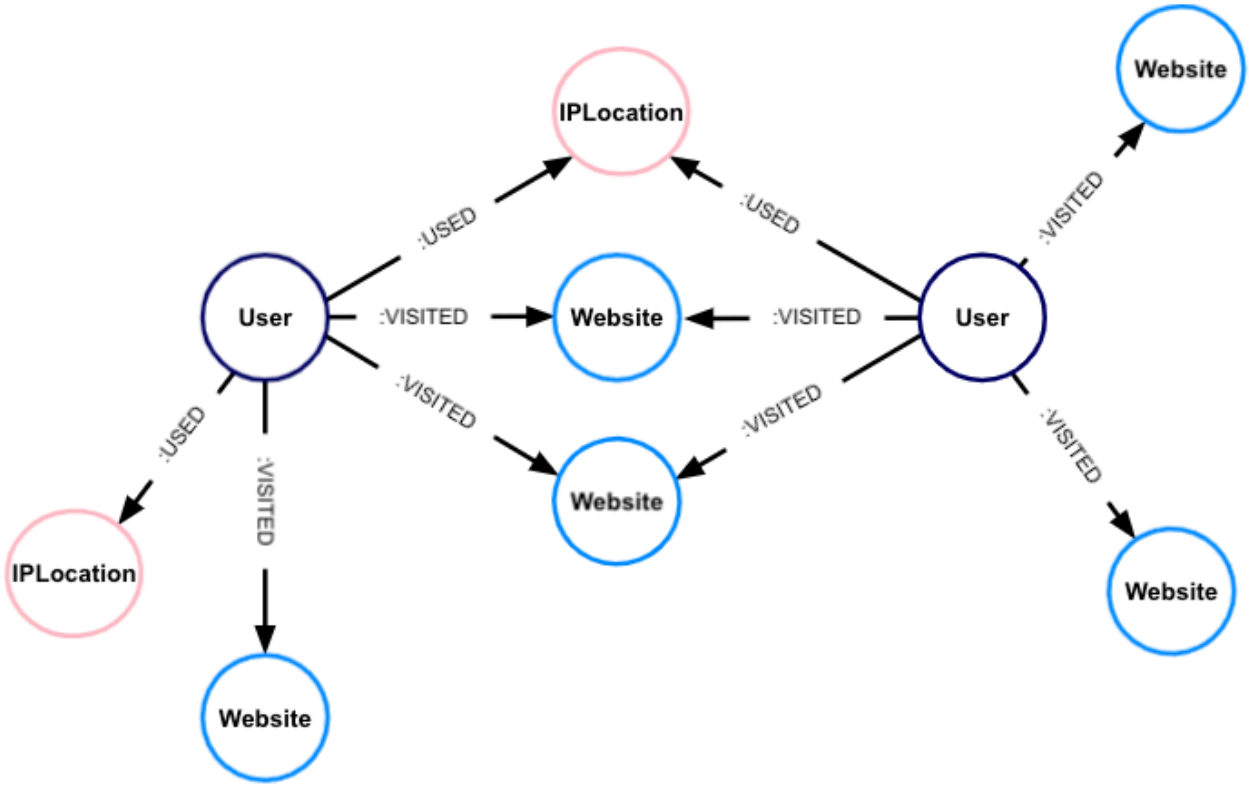
Neo4j enables you to ingest massive streams of data and resolve them to unique, targeted profiles by leveraging the power of graph analytics.

Data Model

One common use case for entity resolution is the disambiguation of website users, but this example easily extends to any scenario where there are various identifiers and user signals.

Users often own multiple devices and multiple accounts and browse on the go. An advertiser would have a limited – and incorrect – view of that user's history unless they were able to resolve all instances of activity and correctly attribute them. Entity resolution accomplishes exactly that.

Here is a graph data model for this example.



A graph data model for entity resolution of website users.

While many entity resolution processes focus on user or customer disambiguation, entity resolution is not limited to resolving questions about people. The same technique works for resolving any type of entity, including places, organizations, and more.

Graph-based approaches also allow you to use not only the traditional identifiers of an entity – names, addresses, and other personal identifiable information,), but also actions and behavior – allowing you to triangulate entities with even more accuracy. In all use cases, there are a set of entities and a set of details about these entities, such as actions or identifiers, to analyze in order to resolve them.

Going back to the earlier example of website or app users, businesses could group users based on similar behavior to create audience segments for marketing campaigns, content personalization, recommendations, and more.

Solution

Identifiers and behavioral patterns can be explicitly modeled as nodes and relationships in a knowledge graph, allowing you to literally "connect the dots" between entities.

Neo4j's Graph Data Science framework offers a variety of analytical approaches to make actionable entity resolution inferences based on graph structure, ranging from localized deterministic query patterns to global supervised machine learning (ML) methods.

Queries

Cypher is a powerful, intuitive, graph-optimized query language. Cypher queries are a great way to get started and explore entity resolution, and they're especially powerful when the criteria for entity resolution is clear and the graph model is relatively simple.

A trivial example is deduplication of user records based on common identifiers: If a system allows only unique email addresses, a query to find all User nodes connected with an Email Address node would quickly show which nodes are duplicates. Cypher queries enable manual resolution and also help uncover patterns that reveal duplicates in the rest of the graph.

Graph Algorithms

The Neo4j Graph Data Science Library's out-of-the-box graph algorithms for similarity and community detection are useful on larger datasets with entities that possess clear similarities. These techniques often identify entity linkages that are not easy to see with manually constructed queries.

The Node Similarity algorithm computes pairwise similarity between nodes based on weighted connections to common identifier nodes. Setting a threshold of similarity based on business requirements yields the resolved entities. While it is a simple and straightforward approach for computing similar nodes in a graph, keep in mind that Node Similarity is computationally intensive.

As an alternative, Weakly Connected Components (WCC), a community detection algorithm, can identify disjointed sections of a graph, each of which can be used to infer entities. WCC is extremely fast and scalable, but it works best where identifier values can be used to cleanly segment entities.

Of course, these algorithms can also be used together to create more sophisticated resolution techniques. For example, Node Similarity can be applied to make linkage predictions between users in the graph, and WCC can leverage those linkages to create identities for consolidated person views.

Supervised Machine Learning

A combination of graph embedding techniques with Link Prediction can reveal entity linkages where identifying overlap is challenging or where linkages cannot be easily defined with deterministic business rules.

As a pre-processing step, running an embedding algorithm like Node2Vec reduces the dimensionality of the graph so that each entity is embedded with knowledge of its connections. Link Prediction then uses the embedded nodes and known, already resolved entities to illuminate linkages in the graph that are not obvious.

The Graph Data Science Library's Link Prediction Pipeline automates this process by letting users define the features they want to use and how they want to sample and split their data. The pipeline automatically builds and selects the best model for the data, expediting the ML process and ultimately leading to value sooner.

Results

Unique entities and segments allow businesses across verticals to better understand the massive amounts of data they possess about users. Identifying unique users, accounts, and devices provides clean, consistent data and allows companies to rapidly target users, detect fraudulent activity, and dynamically merge duplicate records.

Maturing from hand-crafted queries based on business logic to AI-powered, dynamic entity resolution enables businesses to promptly respond and adapt to heterogeneous data streams. Rapid and effective entity resolution can mean the difference between wasting money on incorrectly targeted advertisements and selling to the consumer who needs your product, or letting millions of dollars of fraud go unnoticed instead of proactively stopping fraudsters before they start.

Customer Spotlight: Meredith Corporation

Meredith Corporation is a media conglomerate with a digital presence reaching 180 million users monthly across multimedia platforms, amounting to over $3 billion in annual revenue. Through 40+ publications, from Better Homes & Gardens to People, InStyle, and Allrecipes, Meredith is dedicated to giving users specific and personalized experiences, delivering the right content at the right time.

Entity resolution enables Meredith to create descriptive audience segments leading to personalized experiences and boosting traffic to their media properties. Using a community detection algorithm, Weakly Connected Components (WCC), Meredith was able to disambiguate anonymous traffic on their media properties, turning 14 billion anonymous users into 163 million unique user profiles. As a result, they were able to create a better user experience, which led to a **612% increase in visits per profile.**

Today, Meredith is adding tens of millions of customer data points to their 40 billion node graph daily. Utilizing Neo4j GDS Enterprise Edition's powerful compression feature and low-memory analytics format, Meredith is able to run WCC at scale on a daily basis, resolving new user profiles and enriching existing ones. These results are used both internally to inform business decisions and externally by advertisers for targeting.

Conclusion

Neo4j's Graph Data Science framework enables you to make sense of your data — at scale.

No matter whether your data challenge is finding fraud, audience targeting, or simply ensuring that you have the correct data for each of your customers, using the Neo4j Graph Data Science Library for entity resolution provides fast and valuable results. Entity resolution is just one of the many use cases enabled by graph data science.

Learn more about Neo4j's Graph Data Science framework at neo4j.com/graph-data-science or get started right away with a Neo4j Sandbox.