Introducing Pixie, an advanced graph-based recommendation system

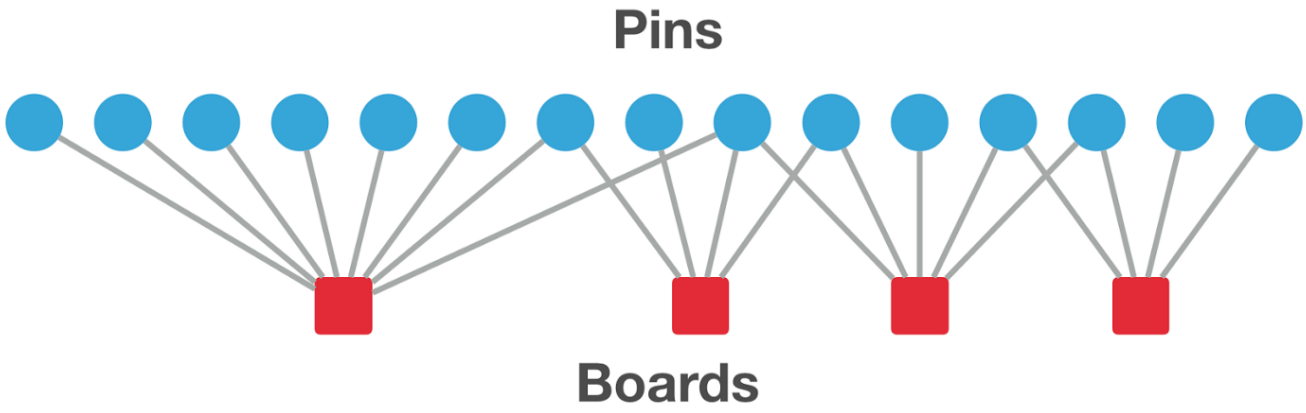
Pinterest Engineering
Mar 31, 2017 · 3 min read

By: Pong Eksombatchai & Mark Ulrich | Pinterest engineers, Discovery

At Pinterest, a primary engineering challenge is helping people discover and do things every day, which means serving the right idea to the right person at the right time. While most other recommender systems have a small pool of possible candidates, for instance 100,000 movies, Pinterest has to recommend from more than 100 billion ideas saved by 150 million people around the world in real-time. We set a performance goal of 60 milliseconds p99 latency, and to achieve it, we built Pixie, a flexible, graph-based system for making personalized recommendations in real-time. Pixie now powers recommendations across Pinterest in Related Pins, home feed and Explore, and accounts for about half of all Pins saved.

Pixie

We started from a bipartite graph where each edge shows that a person saved a Pin to a board.



This graph captures a huge amount of rich data from our users, and is quite large, with more than 100 billion edges and several billion nodes. Thankfully, RAM today is incredibly cheap, and big data like this is small enough to fit on readily available AWS machines. Before terabyte-scale RAM machines were available, complex distributed systems like Hadoop or Spark were needed to compute algorithms for data of this scale. Fortunately, in a way big data is actually getting smaller! Now we can load the entire graph into a single machine and traverse all of it without making any network calls. This makes real-time algorithms on densely connected graphs much easier to develop and deploy at scale, and allows us to make recommendations in real-time the moment a Pinner opens our app (instead of computing them in batch jobs the night before).

While we've developed advanced machine-learning systems like Pinnability that predict how relevant an idea is to a Pinner, the first challenge is figuring out which of the more than 100 billion Pins to even consider since we can't possibly score them all at once. This is why having a graph-based recommendation system is valuable to us. Pixie solves the candidate generation problem by starting graph traversal from a set of nodes we already know are currently relevant to the Pinner. Then, it only examines the portion of the graph nearest to these nodes by using a biased random walk algorithm to estimate the Personalized PageRank. We start the walk from multiple Pins and find

recommendations at the intersection of all of them. For instance, say a Pinner recently interacted with the following three Pins.

We could send them all to Pixie and get back thousands of similar Pins.

In fact, we could send Pixie hundreds of different Pins each with its own custom weight in a single query.

Pixie has successfully replaced multiple candidate generators at Pinterest. We've seen the system improve user engagement by up to 50 percent and also improve ecosystem health by recommending previously undiscovered content. Today we have a large farm of Pixie servers each supporting 1,000 queries per second with a p99 latency of 60 ms for our recommendations products including Related Pins, home feed, email, Explore and more!

Acknowledgements: This system was built and improved by many engineers at Pinterest Labs lead by Jure Leskovec, including the authors, Rahul Sharma, Jerry Zitao Liu, Pranav Jindal, Yuchen Liu and Charles Sugnet. People across the whole company helped launch this to individual products and improve it with their insights and feedback.

Data Machine Learning AI Engineering Pinterest