Google Cloud

Google を支える推薦モデル「Two-Tower」とベクトル近傍検索技術

佐藤 一憲

Google デベロッパー アドボケイト

スピーカー自己紹介

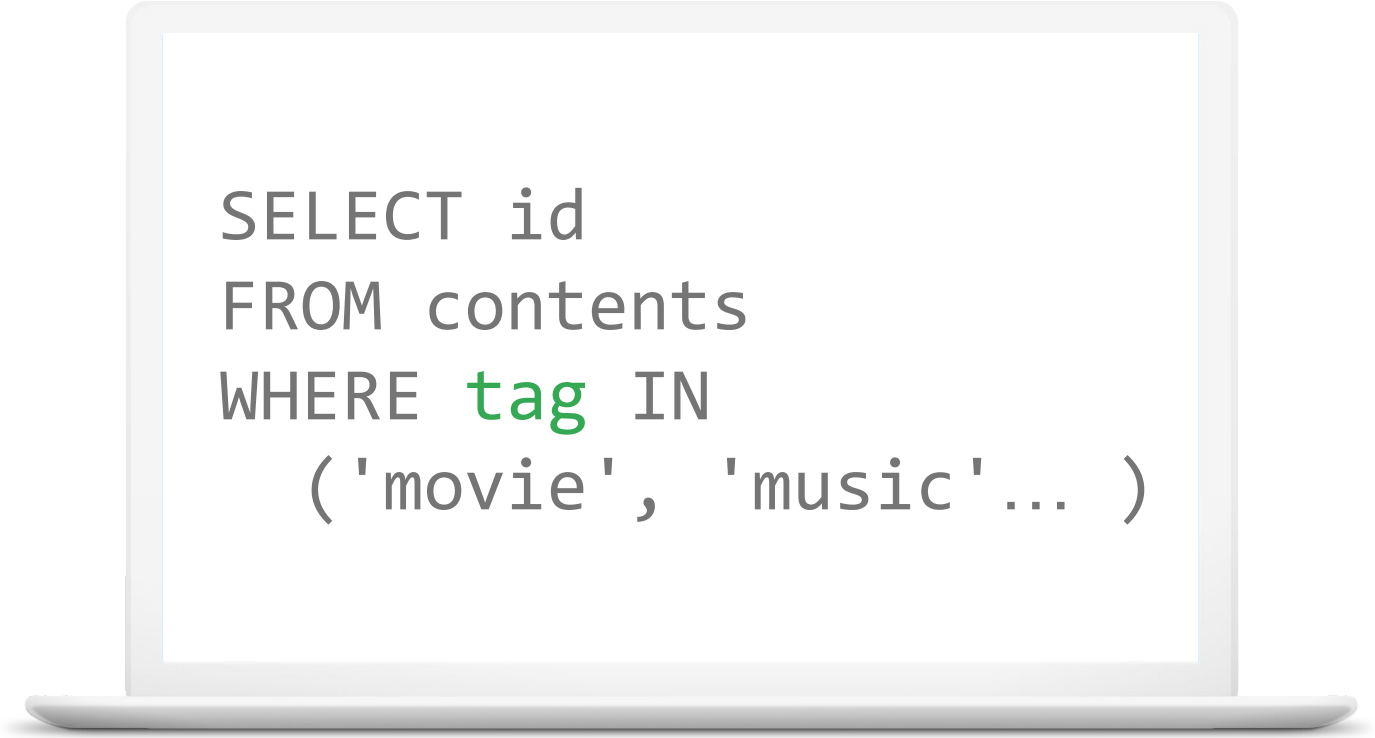
佐藤 一憲
Google
デベロッパーアドボケイト

Google Cloud のデベロッパー アドボケイトとして、
機械学習や AI 系プロダクトの開発者支援を担当。
Google Cloud Next、Google I/O、NVIDIA GTC 等の
主要イベントでスピーカーを務め、Google Cloud 公式ブログに
多数の記事を寄稿。また Google Cloud 開発者コミュニティを
10 年以上にわたり支援している。

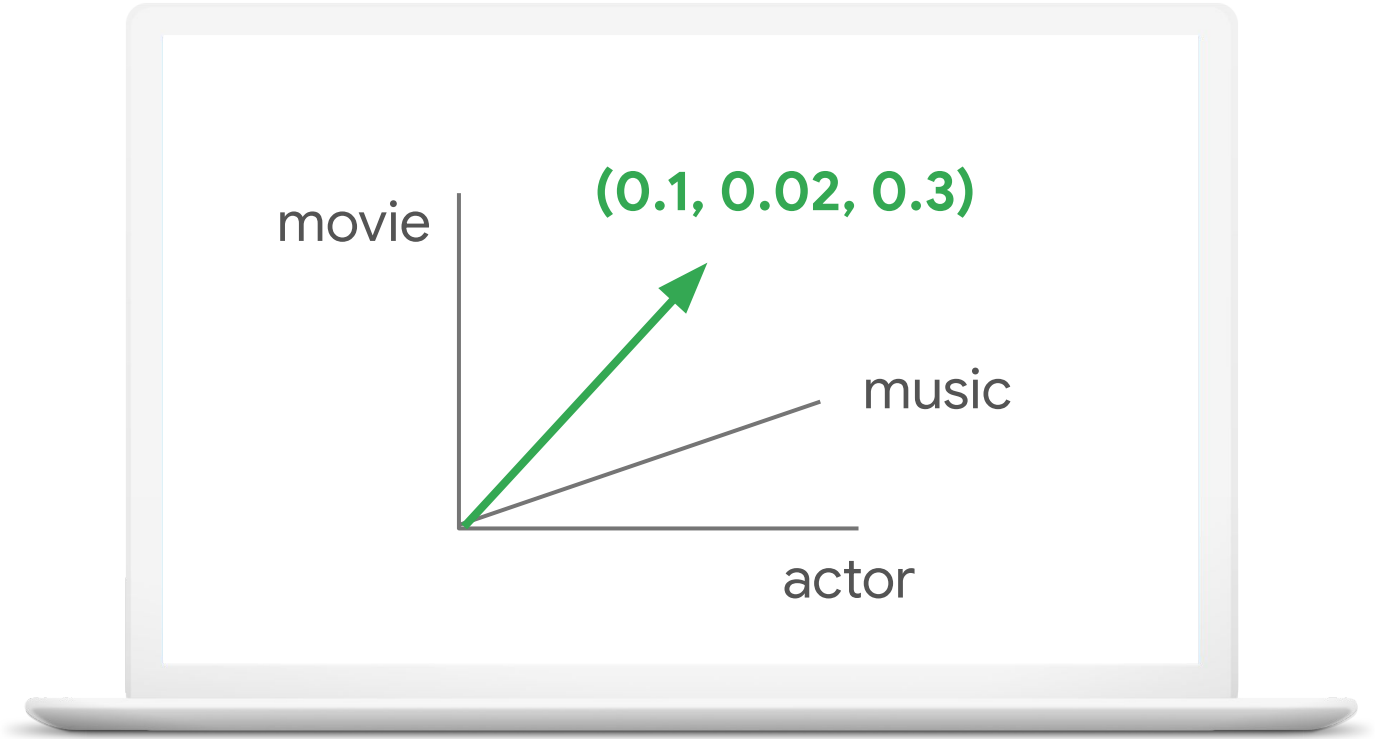
Google を支える
ベクトル検索技術

インターネット上の膨大な情報の中から、
Google はいかにして価値ある情報を見つけているか

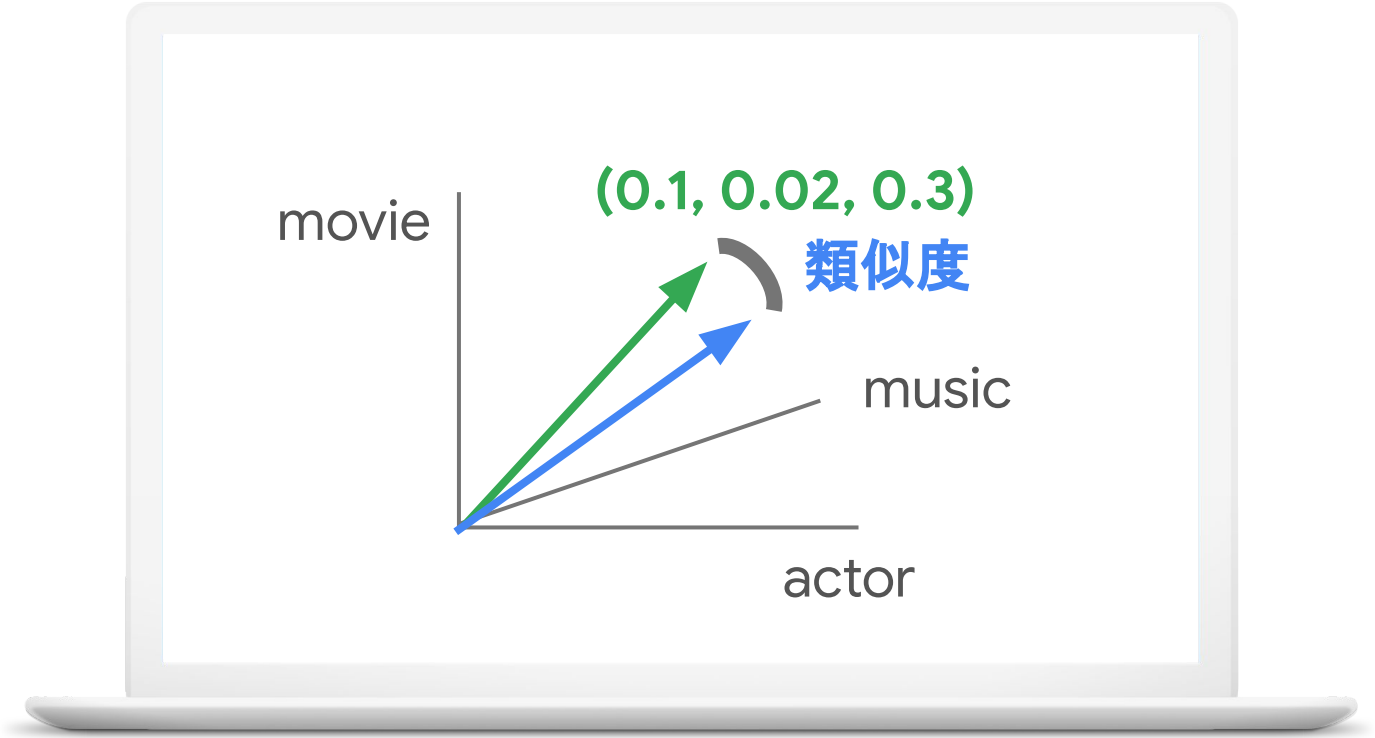
一般的な**キーワード検索**では、文字列やタグ、ラベル等でコンテンツを検索します



一方、**ベクトル検索**では、
ベクトルの類似度でコンテンツを探します

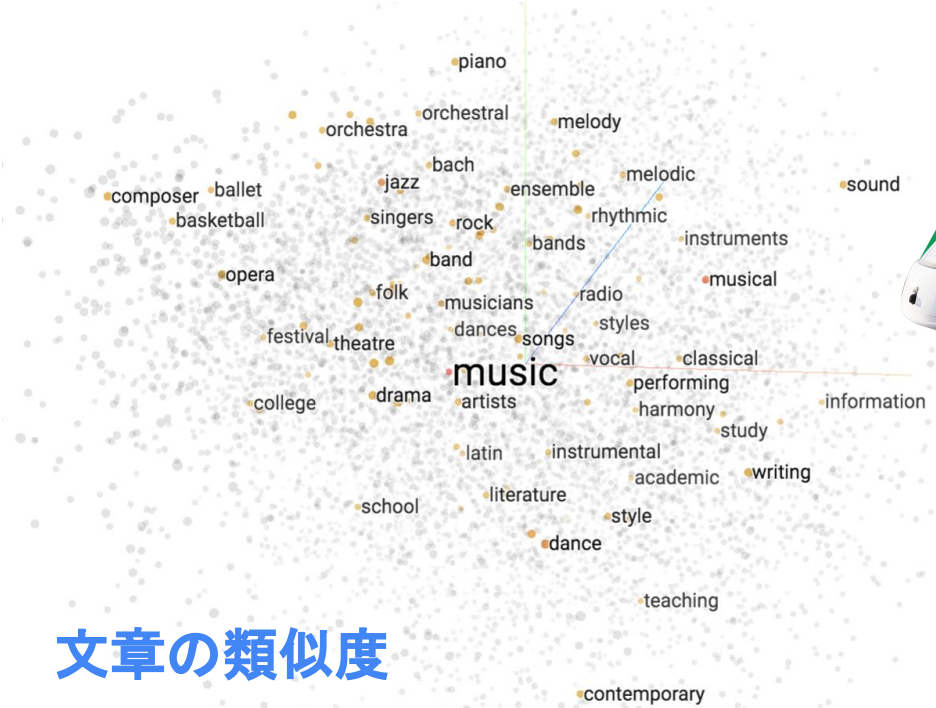


一方、**ベクトル検索**では、
ベクトルの類似度でコンテンツを探します

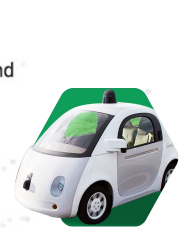


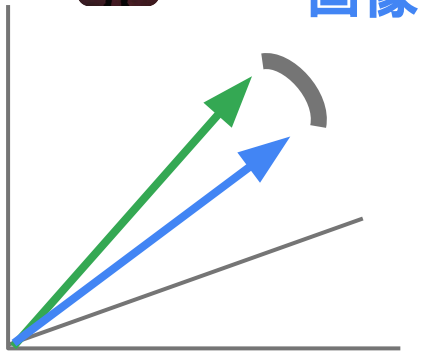
Top-25 matches from 2 million images.

Google 画像検索、YouTube、Google Play 等では、この仕組みでコンテンツを数ミリ秒で検索します



文章の類似度



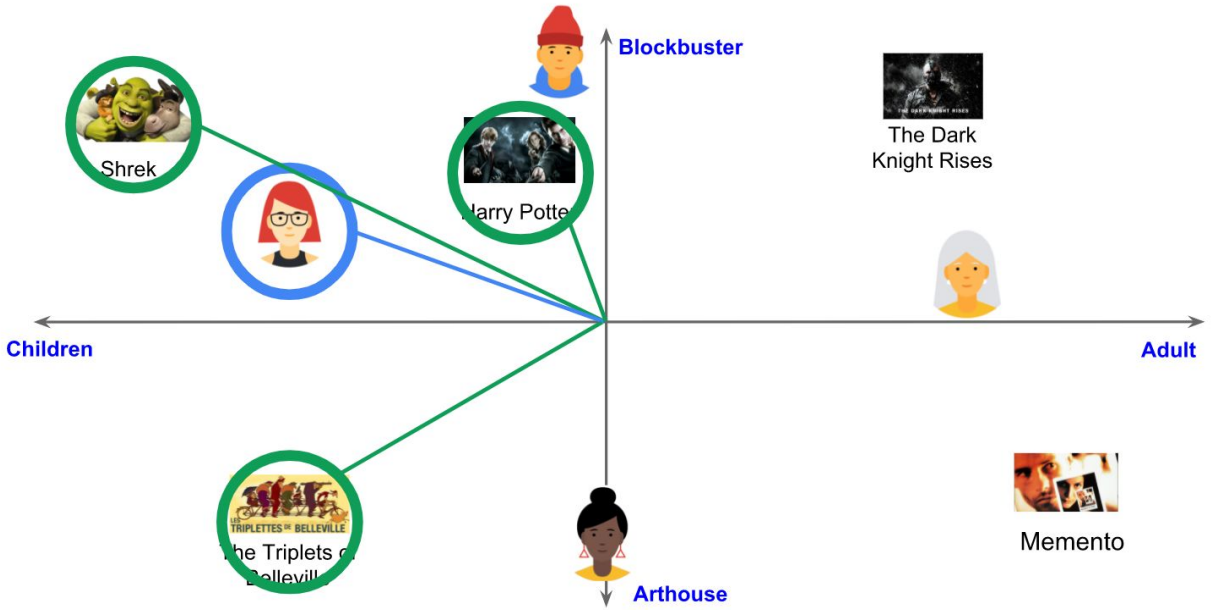


画像の類似度

ベクトル検索で
ビジネスの課題を解決する

Embeddings: ML で作る「かしこい」ベクトル

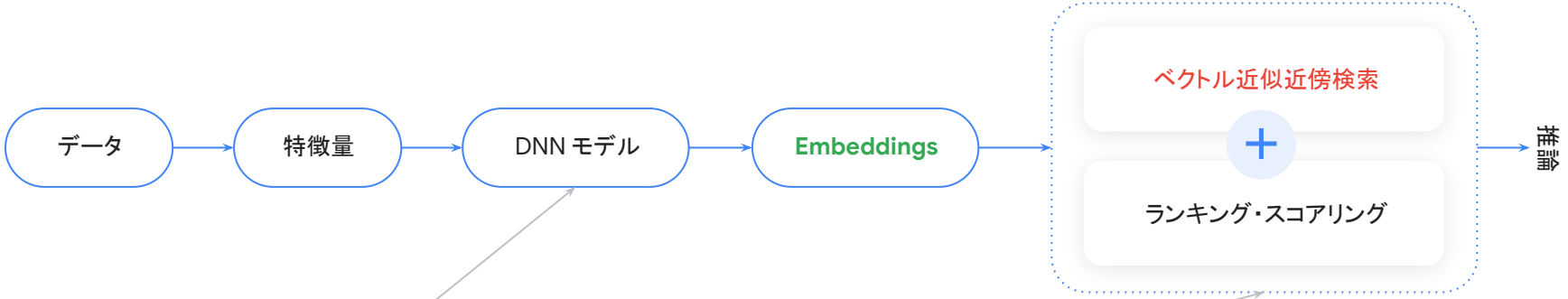
ビジネスやユーザーの要件にフィットするベクトル空間を　ML で作る



From: Machine Learning Crash Course

Embeddings: AI で作る「かしこい」ベクトル

ビジネスやユーザーの要件にフィットするベクトル空間を AI で作る



Step 1: Embeddings 抽出モデルの学習

特徴量から embeddings を抽出する DNN モデル

Step 2: ベクトル検索

ベクトル近似近傍検索による
高速なレコメンデーション

ベクトル検索の応用範囲:

ベクトルを定義できるあらゆる用途で利用可能

文書や**画像**の内容で探す

似ている**製品**を探す

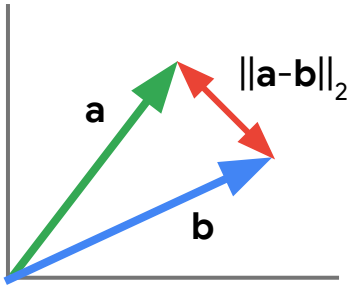
似ている**ユーザー**を探す

おすすめの**音楽**や**動画**を探す

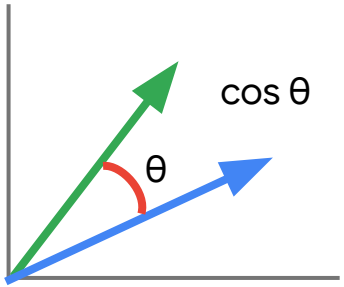
故障しそうな **IoT デバイス**を探す

Vertex AI Matching Engine による高速近似近傍検索

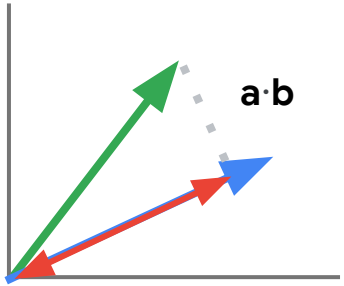
ベクトル検索の難しさ：類似度の比較が重い



L2 distance



cosine similarity



inner product

2 dimensions x

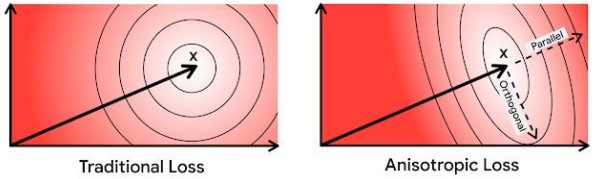
1M items

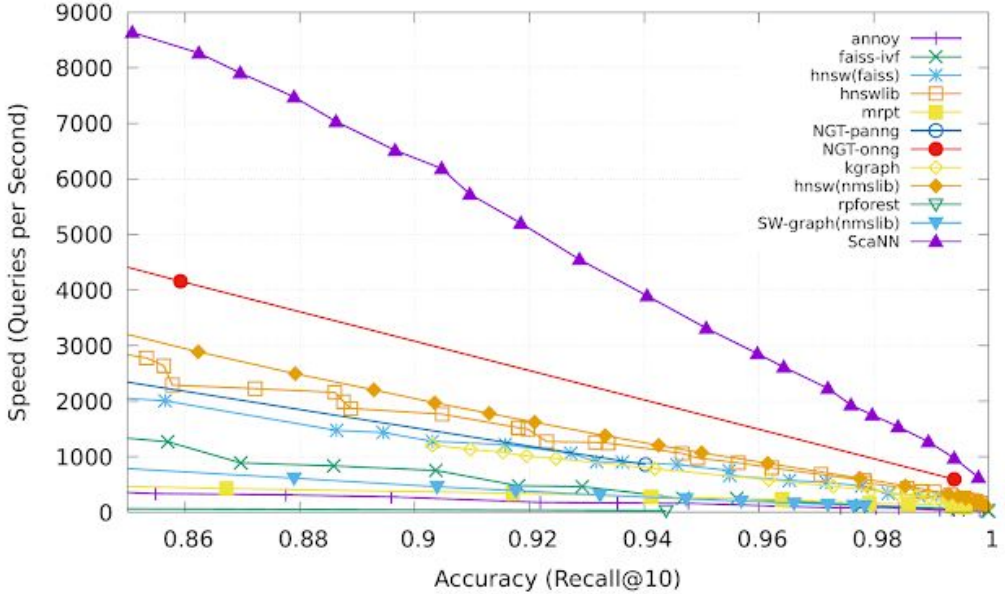
= O(1M x 2)

ScaNN: Google 画像検索、YouTube、Play 等を支える、Google Research 開発の高速近似近傍検索

コードブック作成の新手法

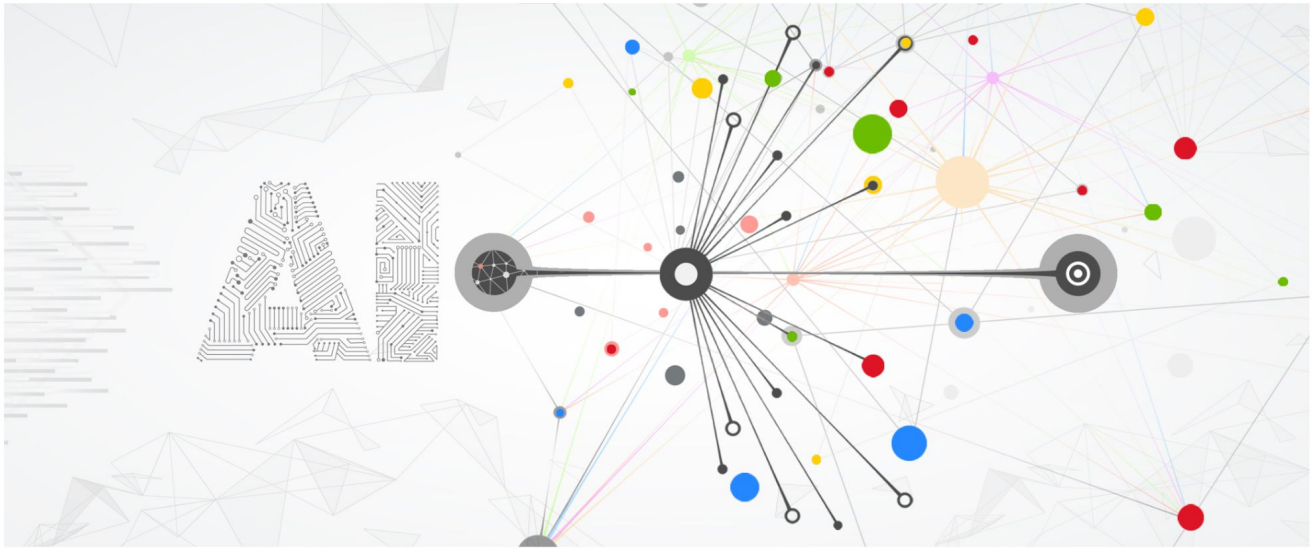
精度と速度のトレードオフを大幅に改善



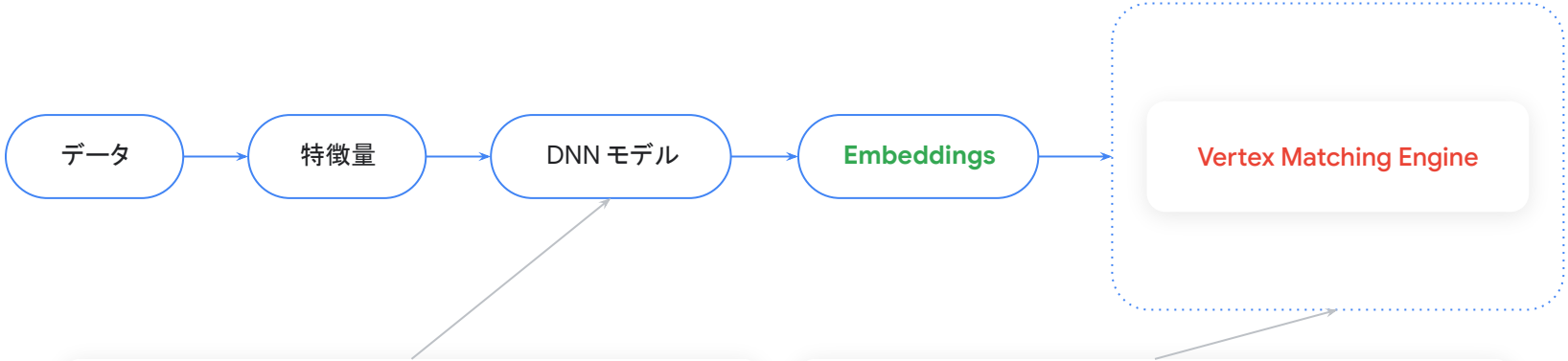


AI & 機械学習

Vertex Matching Engine: 非常に高速かつスケーラブルな最近傍探索



Matching Engine の使い方



Step 1: Embeddings 抽出モデルの学習

- 既存の CNN や BERT モデル等
- Two-tower サンプル
- TensorFlow Similarity

Step 2: ベクトル検索

- embeddings を Matching Engine に登録

リコメンドのための embeddings の作り方

embeddings の作り方

Vertex Matching Engine documentation

https://cloud.google.com/vertex-ai/docs/matching-engine

Resources

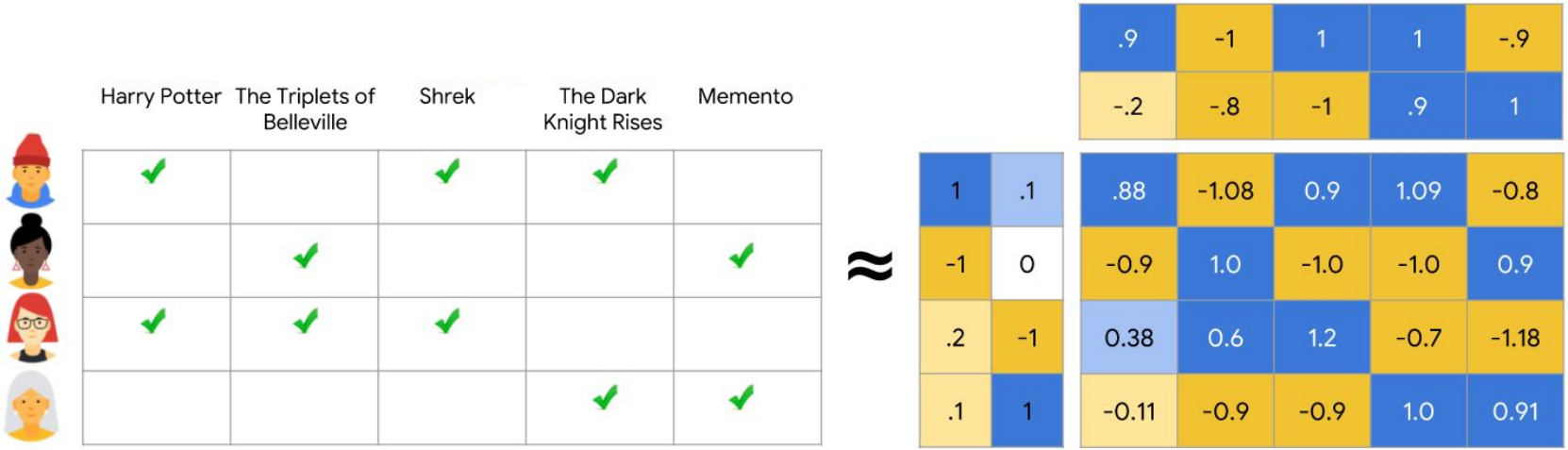
Two-Tower モデル：クエリと候補のペアから embedding 抽出

Swivel モデル：購買履歴などの共起行列から embedding 抽出

TensorFlow Similarity：距離学習による embedding 抽出

https://blog.tensorflow.org/2021/09/introducing-tensorflow-similarity.html

協調フィルタリングによるレコメンデーション



From: [Machine Learning Crash Course]

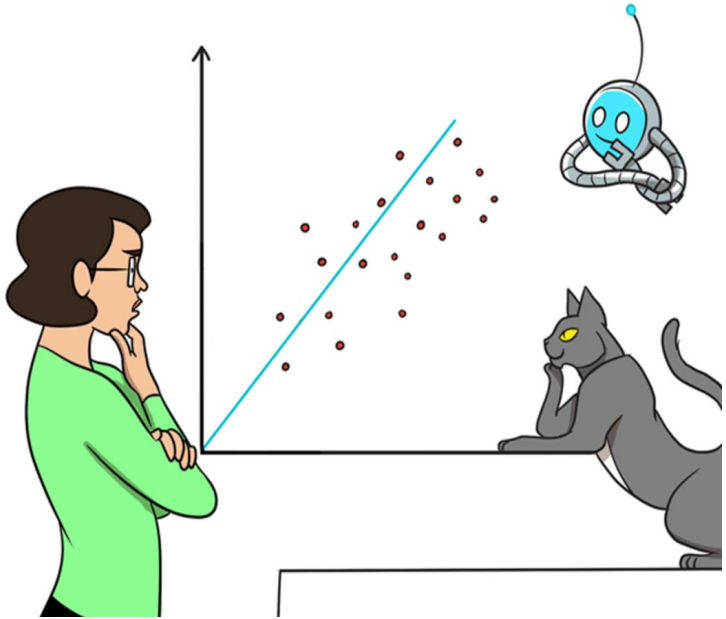
協調フィルタリングの課題

よい点

- ドメイン知識が不要
- 偶然性（serendipity）が期待できる
- 実装例や実績が豊富

いまひとつな点

- コールドスタート問題
  - 未知のクエリに対応にしくい
- 多彩な特徴量を加味できない



Google

Google を支えるレコメンデーション：
Two-Tower モデル

YouTube

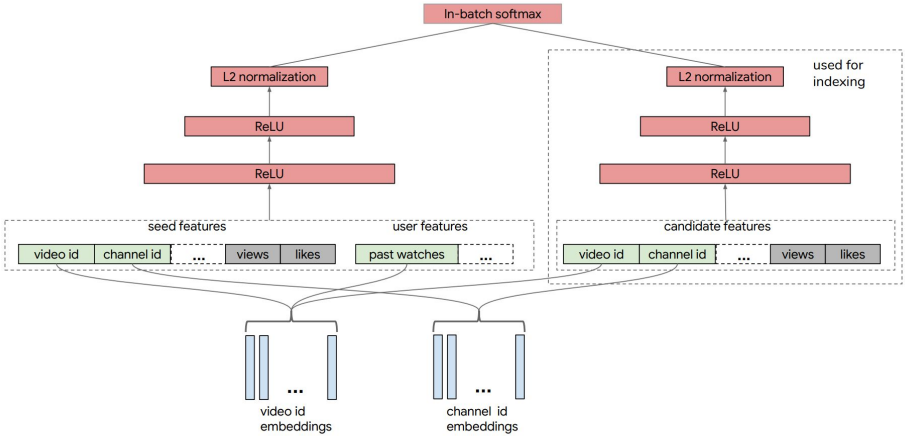


Figure 2: Illustration of the Neural Retrieval Model for YouTube.

Google Play

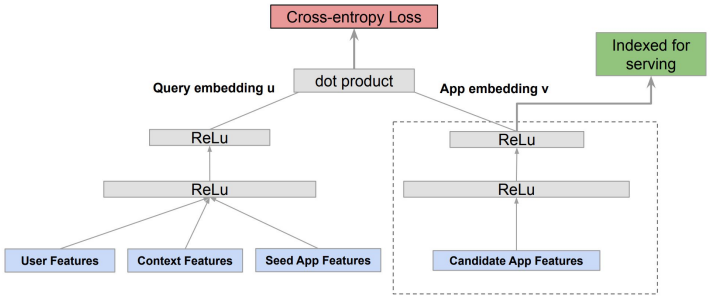
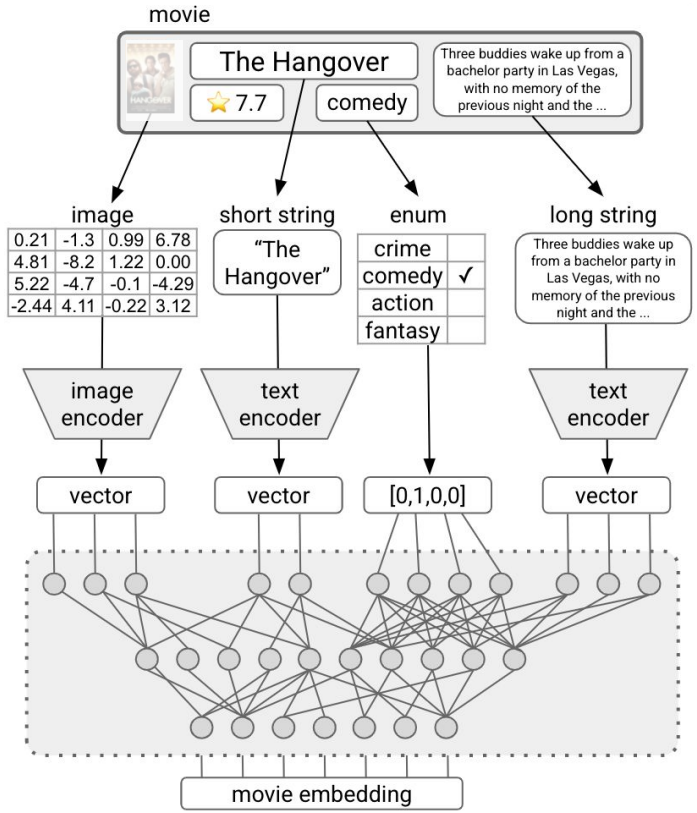


Figure 5: Two-tower model architecture for Google Play app recommendation.

例 : 映画のレコメンデーション

特徴量の抽出

映画の各特徴量から vector をつくる



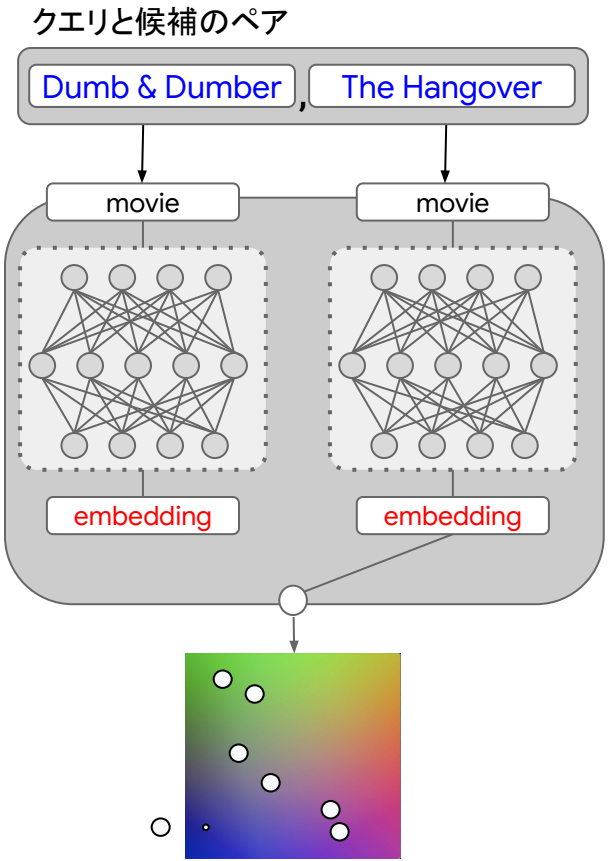
例 : 映画のレコメンデーション

Two-Tower モデルの学習

クエリ（問い合わせ）の映画と
検索結果候補の映画それぞれの
特徴量のペアでモデルを学習

→ クエリの映画と候補の映画の
距離が近い embedding が得られる

→ 個々の映画間の関係ではなく
特徴量と特徴量の関係を学習、
コールドスタート問題を軽減



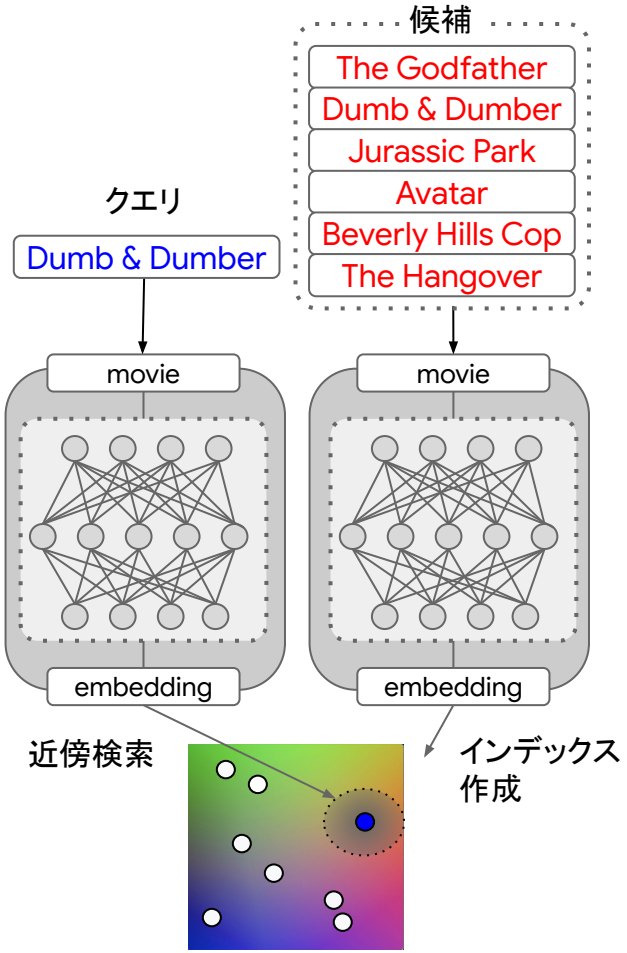
例 : 映画のレコメンデーション

インデックス作成と近傍検索

embedding 空間のインデックスをつくる

Matching Engine 等の近似近傍検索で

クエリから候補を取得



クエリと候補の組み合わせは自由につくれる

クエリと候補それぞれの特徴量を結びつける embedding 空間をつくれる

例:

ユーザ属性 → 製品属性

視聴履歴 → 動画

質問文 → 回答文

文 → 画像（マルチモーダル）

... and more

Two-Tower モデルのメリットとデメリット

よい点

- 特徴量と特徴量の関係を学習
  - コールドスタート問題の解消
- 多彩な特徴量を加味できる
  - マルチモーダルにも対応

いまひとつな点

- まだ広く利用されていない
- モデルの設計と学習が必要

Two-Tower モデル作成の選択肢

TensorFlow Recommenders

TensorFlow のレコメンデーション ライブラリ
https://www.tensorflow.org/recommenders

Universal Sentence Encoders for Q&A

TensorFlow の NLP モデル
https://tfhub.dev/google/universal-sentence-encoder-qa/3

Two-Tower model built-in algorithm

Matching Engine 付属の組み込み機能
https://cloud.google.com/vertex-ai/docs/matching-engine/train-embeddings-two-tower

Recommendations AI

Google Cloud のリコメンデーションソリューション
https://cloud.google.com/recommendations

Two-Tower モデル作成の選択肢

	メリット	考慮点
TensorFlow Recommenders	オープンソース実装 任意の特徴量に対応可能 自由にカスタマイズできる	低レベル実装が必要 技術サポートなし
Universal Sentence Encoders for Q&A	実装の労力が少ない 学習済みモデル オープンソース実装	テキスト Q&A にのみ対応 技術サポートなし
Two-Tower built-in algorithm	実装の労力が少ない 任意の特徴量に対応可能	ソースコードは非公開
Recommendations AI	実装の労力が少ない 技術コンサルあり	ソースコードは非公開 EC サイト等の特定用途に対応

デモ

Key Takeaways

Google の「虎の子」であるベクトル検索技術と

Two-Tower モデルによるレコメンデーションで

ビジネス課題に向けたソリューションを構築

Thank you.

Google Cloud