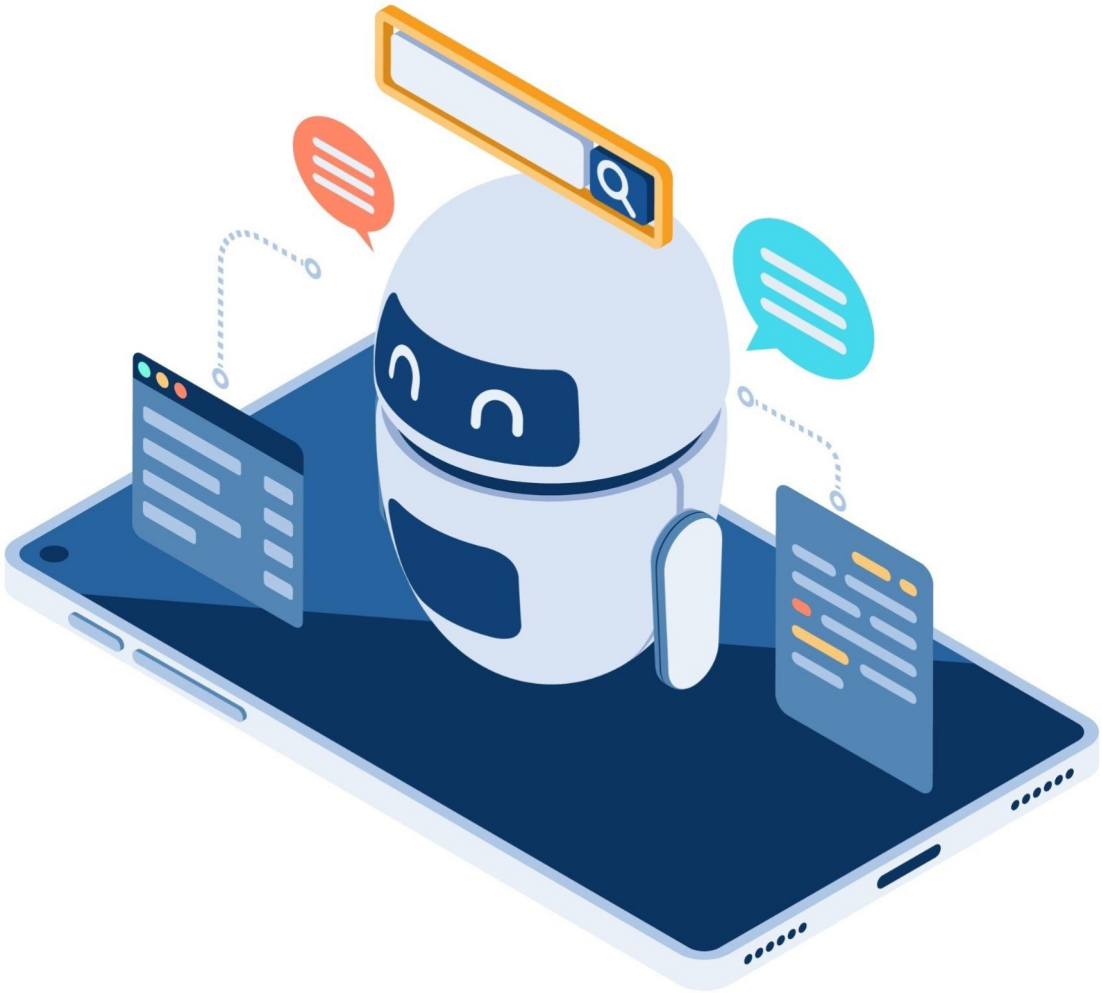


Session Report

既存の RDB を生かしたままで
より利便性が高い検索システムが可能に

今さら聞けない！ベクトル検索超入門 ～データベース ユーザーは 何をおさえておけばいいのか？～

Google Cloud ソリューション & テクノロジー グループ
データベース スペシャリスト　佐藤 貴彦

Google Cloud ソリューション & テクノロジー グループ
データ アナリティクス カスタマー エンジニア　木村 拓仁

Google Cloud

Google Cloud

セッション レポート概要

生成 AI の発展に伴い、画像やテキストから互いに似たものを検索するセマンティック検索のニーズが高まっています。それに関連して「エンベディング」や「ベクトル検索」への関心が高まっており、データベースでもこれらを扱う方法が求められています。本セッションでは、ベクトル検索とは何かについて紹介した後、AlloyDB AI など Google Cloud のマネージド データベースに格納された情報とベクトル検索を組み合わせる方法について紹介します。

プレゼンター紹介

Google Cloud ソリューション & テクノロジー グループ
データベース スペシャリスト
佐藤 貴彦

これまで複数の製品ベンダーにて、ネットワークやデータベースを中心としたインフラ周りの支援に従事。Google Cloud では、お客様が Google Cloud 上のデータベース系サービスをうまく活用できるよう、お客様の話を聞き一緒に議論していくことが仕事。

Google Cloud ソリューション & テクノロジー グループ
データ アナリティクス カスタマー エンジニア
木村 拓仁

Google Cloud 機械学習、データ基盤専門のカスタマー エンジニア。大学、大学院で深層学習の研究に明け暮れ、機械学習エンジニアやフリーランスを経て Google Cloud に入社。専門分野は凸最適化と自然言語処理。

目次

生成 AI がセマンティック検索の進化を加速

昨今、ニュースなどで頻繁に耳にするようになった「生成 AI」は、大きく 2 つの側面でビジネスに価値をもたらすと期待されています。

1 つはコンテンツ生成です。スライドやメールの作成など、コミュニケーションのために使っていた時間を削減することができます。エンジニアであれはプログラムのソースコード生成など、業務時間を短縮化できます。もう 1 つは新しいユーザー体験です。Webサイトやモバイルデバイスを通して、ユーザーは欲しい情報をより素早く、より直感的な形で見つけ出すことができます。

このような生成 AI の価値を踏まえ、Google では Google 検索をはじめとした主力製品を、大胆かつ責任あるアプローチで再構築しています。数あるプロジェクトの中で、特に Google 検索は生成 AI 技術の飛躍により大きく変わりつつあります。

Google では、2015 年頃から、キーワードをスペースで区切って行っていた従来の検索から、例えば「Google Cloud はどのような会社か」という 1 つの文章で質問でき、検索エンジンはよりユーザーの検索意図を理解して検索結果を返すようになりました。このような検索技術は「セマンティック検索」と呼ばれ、生成 AI 技術によってさらに進化を遂げています。

直近の具体的な機能強化例が「Search Generative Experience（SGE）」機能です。SGE 機能を用いると、先ほどの質問文に対して、「Google Cloud は Google が提供するクラウドコンピューティングサービスです」とAI によって要約された検索結果を表示できます。さらに複雑な質問にも対応できます。

これらの機能を企業が活用すれば、ユーザー体験は大きく変わるでしょう。欲しい情報を素早くわかりやすいインターフェースで取得できるため、顧客満足度の向上につながります。検索されたデータを分析することで、企業は自社の顧客の理解を深めることができ、マーケティングの側面でも役立ちます。さらに、セマンティック検索を実装するために必要な AI の知識は、その他の業務にも転用できるため、技術的な優位性につながってくると言えます。

企業がセマンティック検索を扱うことで得られる 3 つのビジネス価値

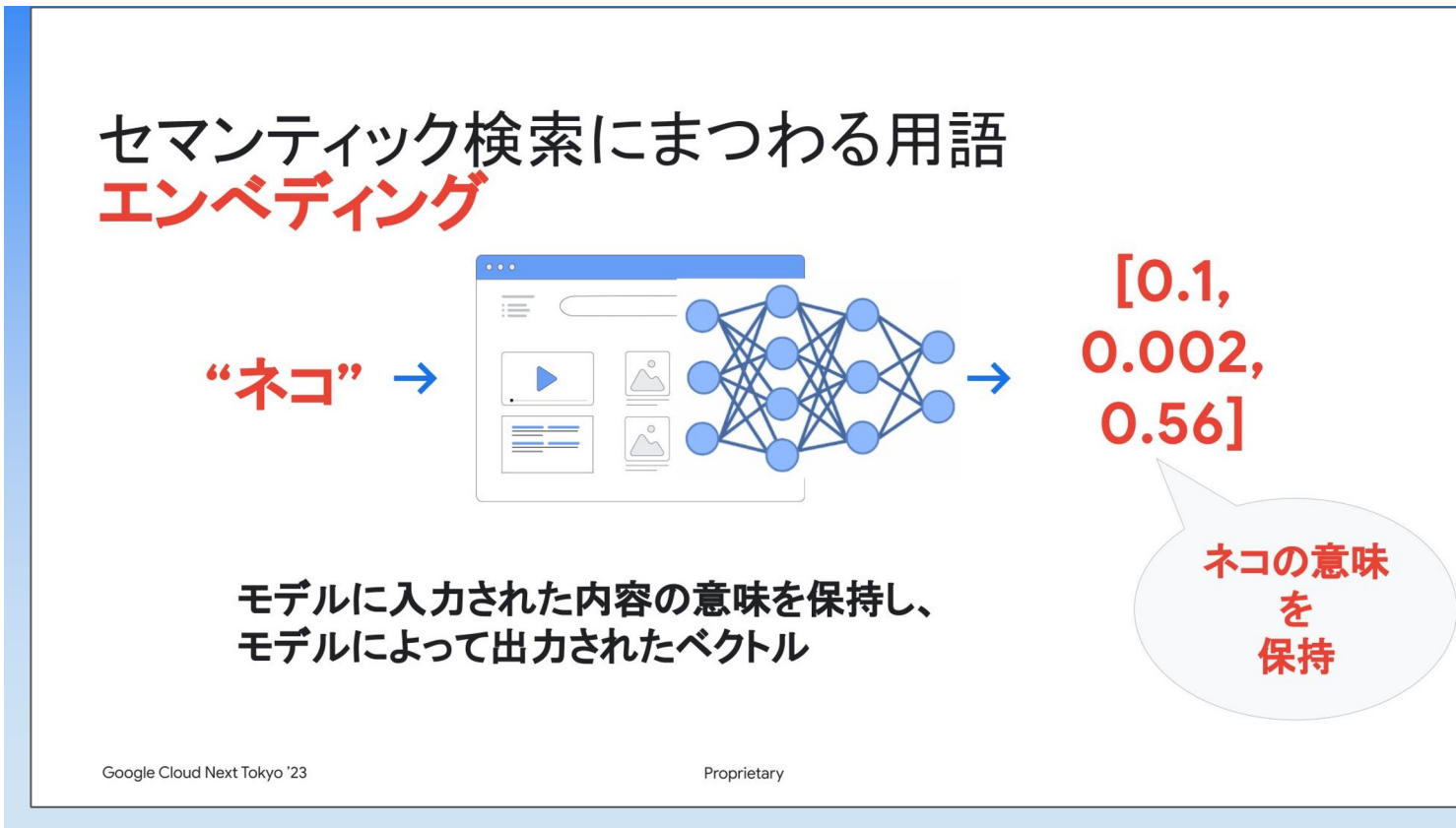
セマンティック検索の仕組みとベクトル

ここからは、セマンティック検索の裏側にある技術的な仕組みについて解説します。

まず、従来の検索の裏側にあるデータは、一般的には構造化データです。ユーザーが検索した時に、アプリケーション サーバーからデータベースが呼び出され、データベースから必要な情報を返します。しかし、生成 AI によって性能が飛躍的に向上したセマンティック検索では、裏側のデータはベクトルとして保持されます。ユーザーが検索を行うと、アプリケーション サーバーから AI が呼び出されて、検索クエリがベクトル化され、そして必要な情報が返される流れです。

先ほど紹介した Google 検索の SGE 機能も、裏側でベクトル化が行われています。検索クエリをベクトル化するこの AI はモデルと呼ばれます。モデルとは、例えばある入力に対して、その入力の内容に何らかの評価を行い、値を出力するものと捉えてください。プログラミングにおける関数と似たようなものですが、関数に統計学や機械学習の知識が入っている場合にモデルという言葉を用います。

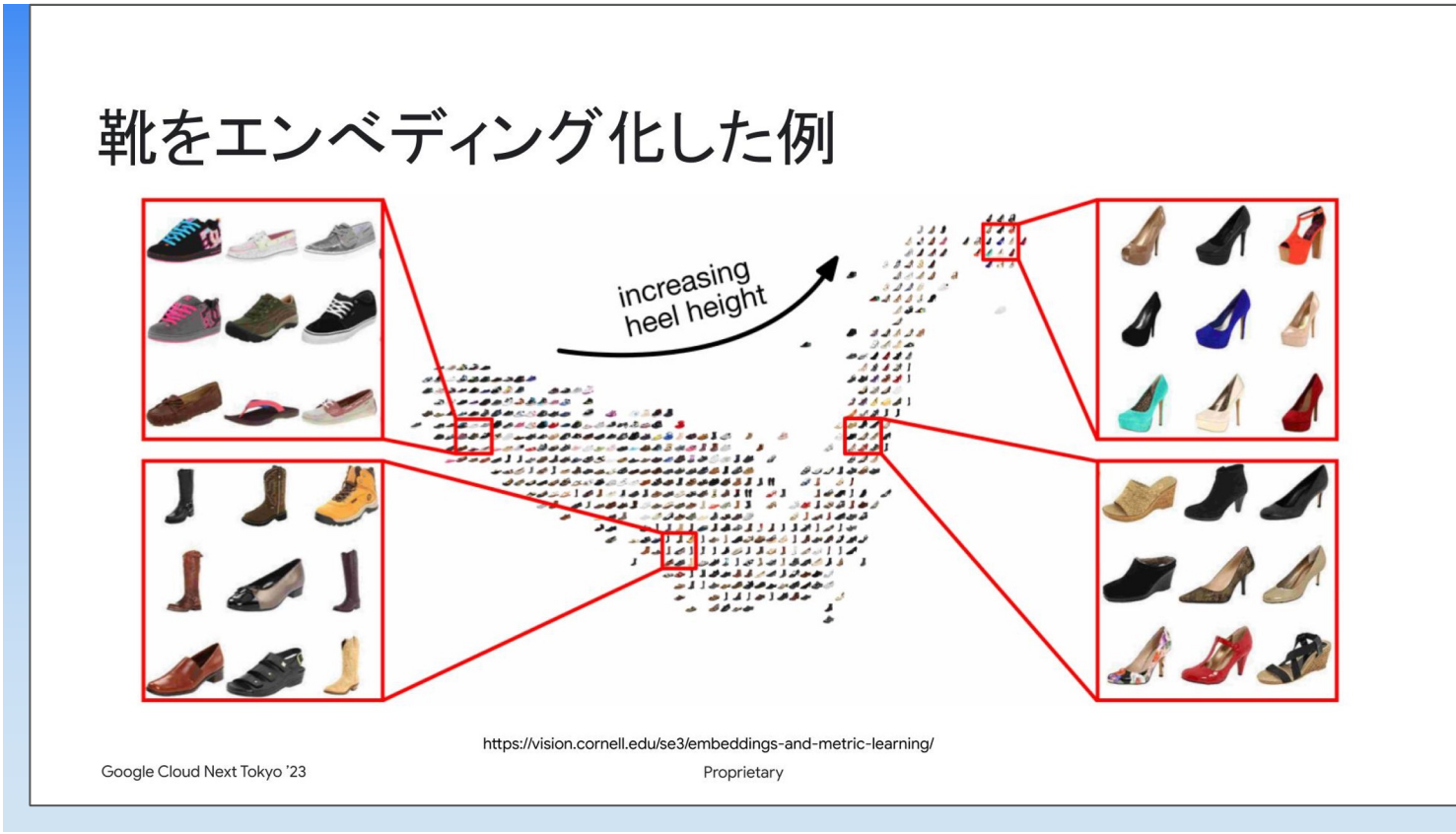
セマンティック検索の場合、ユーザーが検索した文字列は、ベクトルが出力として返されます。入力された文字列をモデルによって変換したベクトルは、エンベディングと呼ばれます。これは単なる数字の羅列ではなく、特別な意味を持つものです。



エンベディングは、モデルによって入力された内容の意味をベクトルの形に変換したもの

エンベディングはマルチモーダルであるため、画像やテキスト、音声といったあらゆるデータを、ベクトルという統一の形で扱えます。結果として視覚、聴覚、言語などのあらゆるインターフェースで検索ができるようになります。

以下は、エンベディング化された靴の関係性をわかりやすく2次元で落としたものです。さまざまな種類の靴の画像がモデルによってエンベディング化されていることがわかります。近くにあるエンベディングの画像を見ると、スニーカーなど高さが低めの画像が集まっていることが確認できます。このように、エンベディング化を行うと、人間に直感的な形でアイテムをグルーピング化でき、グルーピング化した中から、後述するベクトル検索を用いて似たデータを取得することが可能です。



靴の画像データをエンベディング化した例

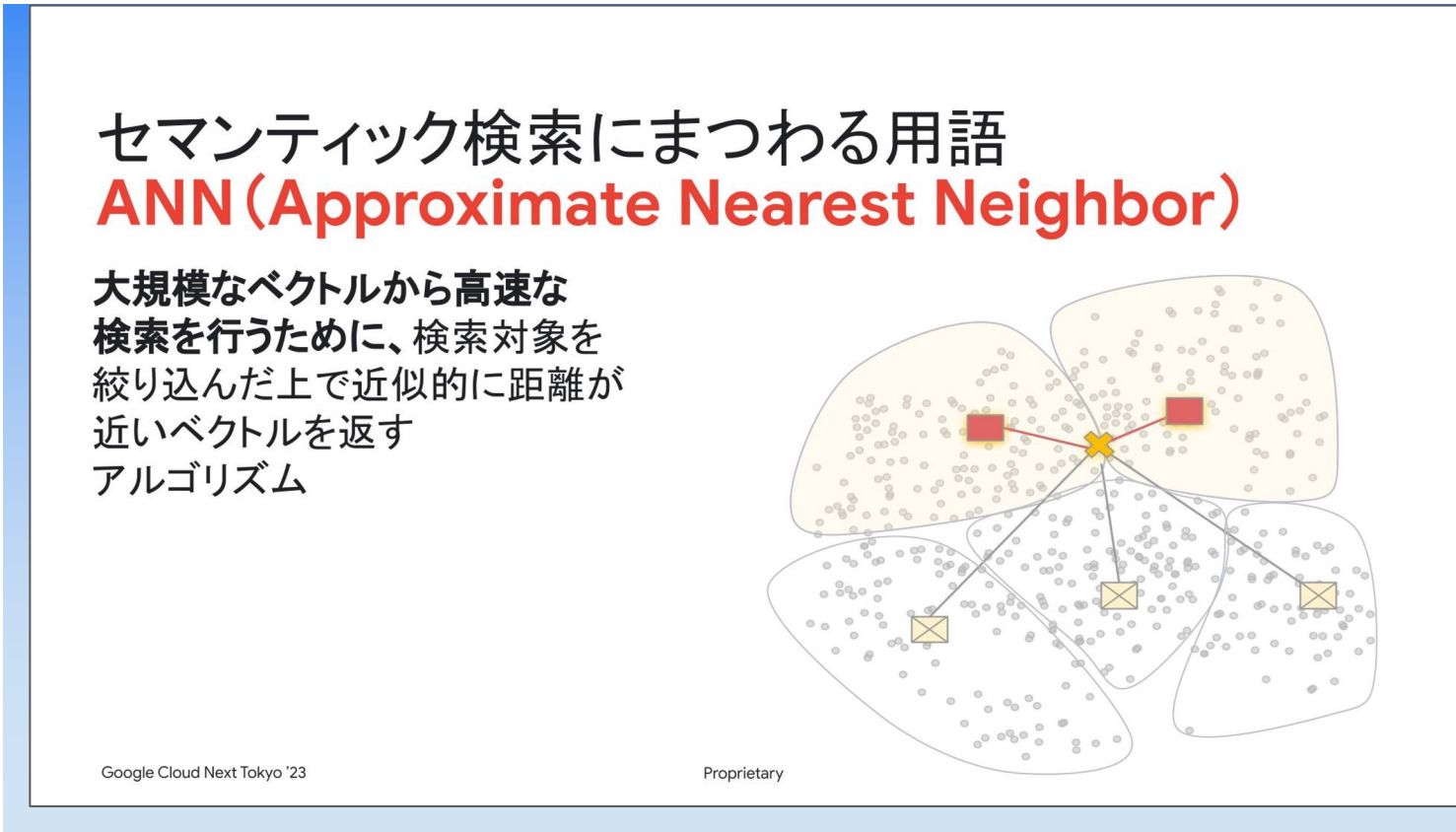
ベクトル検索では「ベクトル間の距離」を計測する

次に、エンベディングと同じくセマンティック検索の高度化に欠かせない「ベクトル検索」を紹介します。

ベクトル検索とは、ベクトル化された各データの距離を計算して類似度を判定するものです。その計算方法はさまざまな手法がありますが、注意したいのが、ベクトルという数字の配列を比較するため、単なる比較だけでも計算量が膨大になってしまうということです。

仮にベクトル検索を EC サイトに実際に導入する場合、数百万と膨大な商品数があるサイトの場合、計算はさらに重くなり、単純に実装しただけでは処理できなくなります。各商品のベクトルどうしの比較をユーザーごとに行えば、とてつもない計算量となります。

そこで必要になるのが、ベクトル検索に特化したアルゴリズムである「Approximate Nearest Neighbor（ANN）」です。全ベクトルから似ているベクトルを検索するのではなく、似ていそうなベクトルに範囲を絞り込み、検索を行う近似アルゴリズムです。



Approximate Nearest Neighbor（ANN）アルゴリズム

靴の例では、スニーカーの画像を調べる際に、すべての靴を検索対象として調べるのではなく、スニーカーのクラスターに範囲を絞り込み検索を行います。Google では、この ANN に特化したマネージドサービス「Vertex AI Vector Search」を提供しています。水平スケーリングやシャーディングといった運用を自動化する機能や、Google 規模の自動スケーリング機能を有するため、数億単位のベクトルや高負荷な QPS（クエリ毎秒）にも対応できます。

Vertex AI Vector Search の裏側では、Google が開発した ScaNN と呼ばれる ANN アルゴリズムが用いられています。これは、Google 画像検索をはじめ、YouTube、Google Playといった Google のあらゆるサービスで使われています。

問い合わせ対応業務に有効なベクトル検索

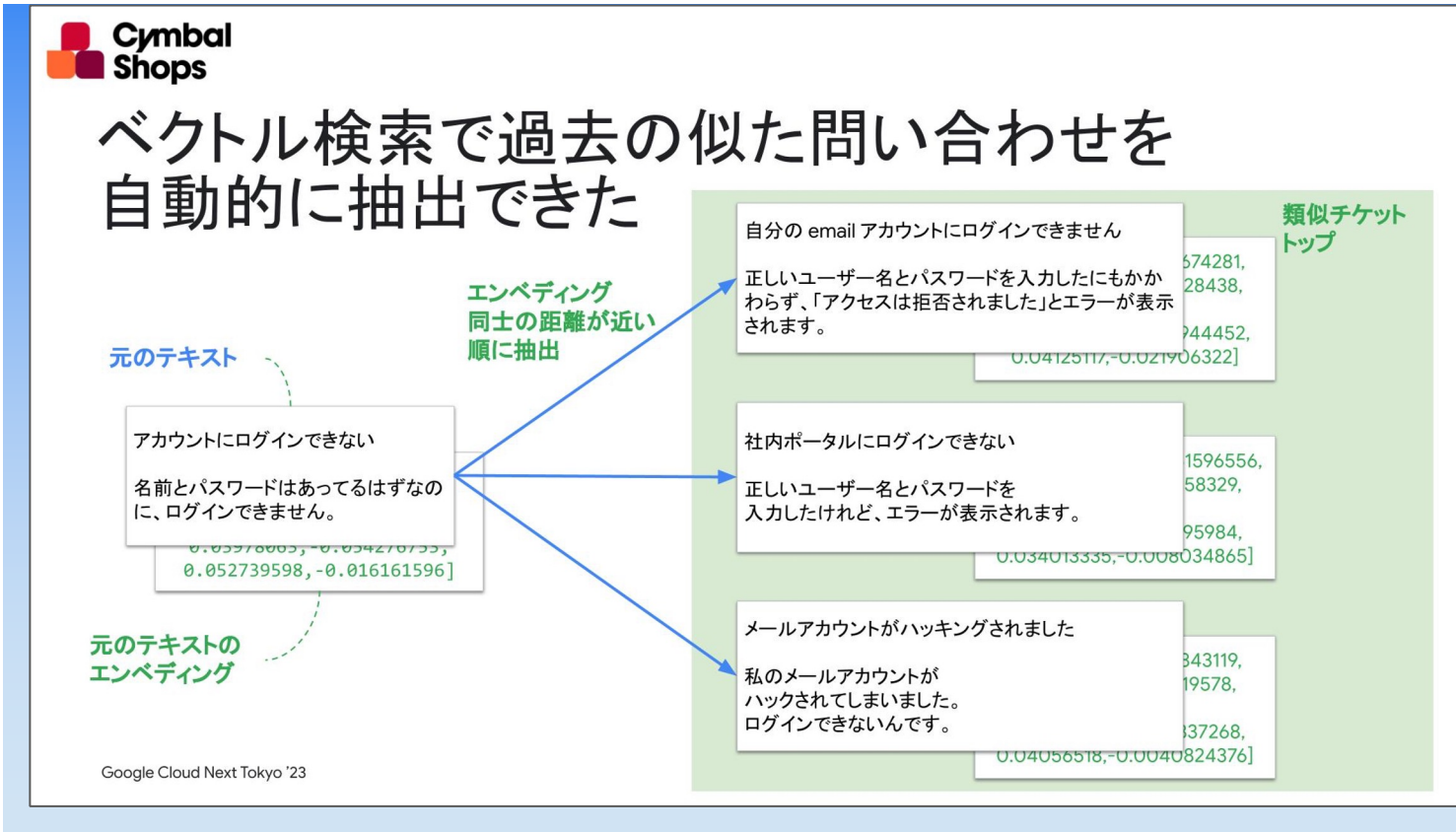
ここまでベクトル検索の概要に触れてきましたが、ここからは通常のリレーショナル データベースにて、このベクトル検索ができたらどのような世界が広がるか、社内 IT ヘルプデスクの問い合わせシステムを例に紹介していきます。

IT ヘルプデスクは、社内のさまざまなシステムに関する問い合わせを受け付ける窓口であり、似たような問い合わせが多く寄せられます。当然、同じような回答があるため、正しい回答を提示するまでに時間がかかってしまうという課題もあります。

それを改善する 1 つのアプローチには、例えば、問い合わせ内容を分析して、それに応じた適切な過去の類似チケットを自動的に提示するという方法が考えられます。この場合、一般的にはキーワード検索でヒットさせる、つまりキーワードに紐づく内容をデータベースの中からフィルタリングして抽出する方法が用いられます。

確かにこの手法でも過去の類似チケットを提示できますが、そもそもそのキーワードの選定が難しい場合があります。さらに、「たまたまそのキーワードが入っていただけで関連性の低いチケット」が検索されてしまうという課題も考えられます。

ここで有用なのがベクトル検索です。ベクトルとは数字の配列のことですが、ここではただの配列ではありません。元々の変換前の文字、変換前のデータ情報の意味が内包された数字表現となります。これを用いて、例えば問い合わせ内容のテキストをベクトル（エンベディング）に変換します。これを過去のチケットのエンベディングと比較（ベクトル同士の距離を計測）することで、より過去の問い合わせに近いチケットを探して抽出することができるようになります。



ベクトル検索を利用してデータベースから適切な情報を抽出するイメージ

データベースにベクトル検索を組み合わせる具体的なイメージ

では、実際にこのようなベクトル検索を行うためには何が必要でしょうか。1つは、入力テキストをエンベディング、つまりベクトル表現に変換するための AI モデルです。次に、ベクトルどうしの距離を計算する機能です。その計算方法はユークリッド距離やコサイン距離、内積などさまざまな手法があります。最後に、素早く検索するためのアルゴリズム（先程の例では ANN）も必要です。

これを実現するのが、Google Cloud が提供するベクトル検索特化型のエンジン「Vertex AI Vector Search」です。後述しますが、具体的な流れとしては、まずはベクトル検索によって類似したベクトルの ID を取り出し、次にその ID を使って、リレーショナル データベースの中を検索し、該当のチケットの情報を抽出するという流れになります。

ここでは、通常のデータベースが持つフィルタリング機能、すなわち SQL における WHERE 句や全文検索機能なども、ベクトル検索と組み合わせて利用できます。普段利用しているリレーショナル データベースのテーブルのある列にベクトル情報を格納するイメージです。他の列に保持されている情報をフィルタリング条件に組み合わせることで、よりユーザーの意図に沿った検索結果を返していきます。

既存のリレーショナル データベースにベクトル検索を組み合わせるイメージ

Google Cloud のデータベースでベクトル検索を行う方法

上述のアプローチを具体的に Google Cloud を用いてどのように実践するかを紹介しましょう。

「Vertex AI Embeddings for Text」には、入力したテキストをエンベディングに変換するための API が搭載されています。これは「BigQuery」など Google Cloud の各種マネージド サービスから直接呼び出すことができます。つまり SQL で呼び出せるということです。

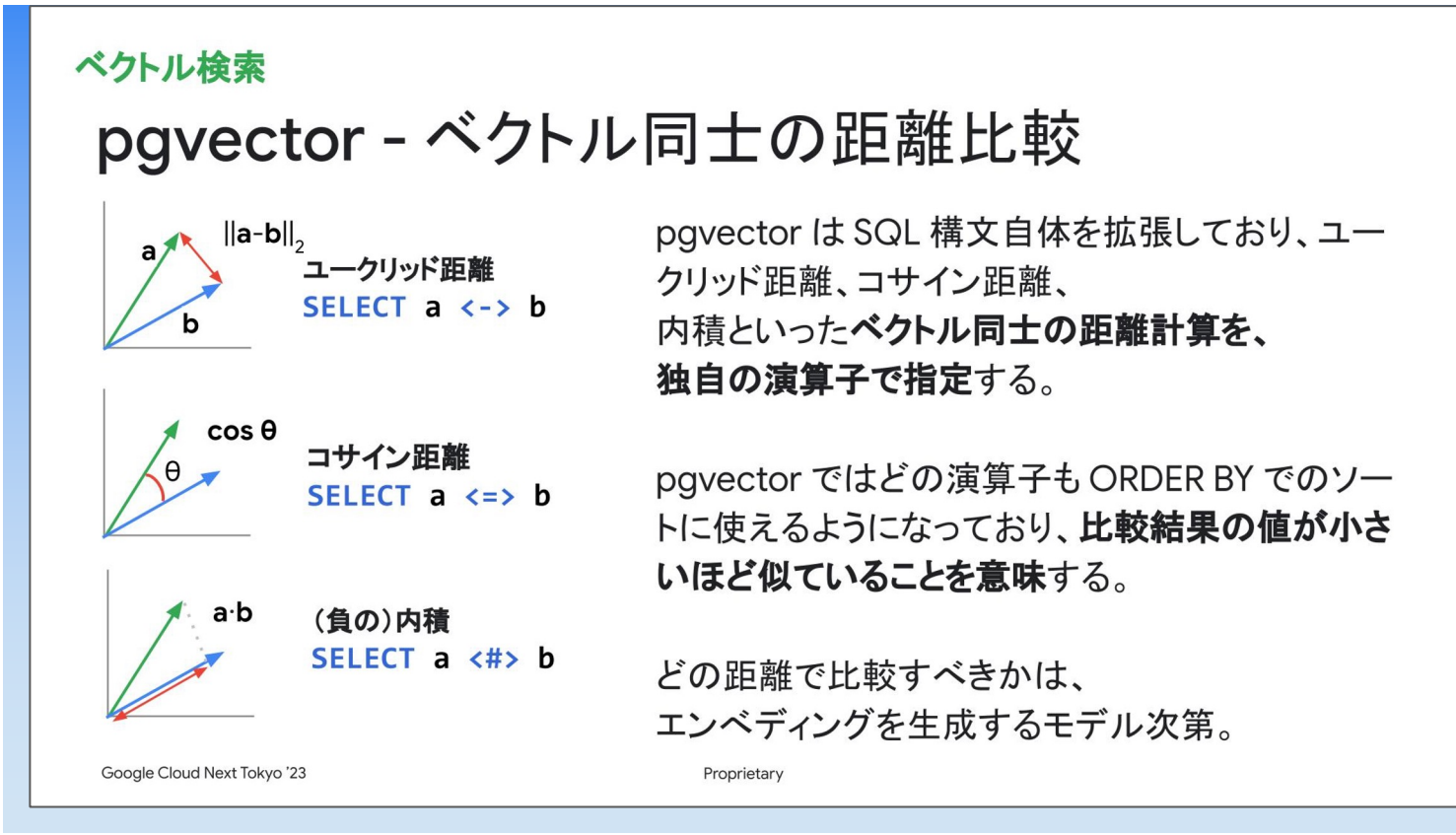
Google Cloud のマネージドリレーショナル データベースには、PostgreSQL をベースに高速化した「AlloyDB for PostgreSQL」や、Google 自身が使用してきた分散データベース「Cloud Spanner」など、いくつかの種類があります。これらのサービスは Vertex AI とネイティブに連携しており、SQL の SELECT 文で直接 AI のモデルを呼び出すことができます。

データベースと Vertex AI 連携機能の拡充

今回は AlloyDB for PostgreSQL を用いた例を紹介しましょう。まず AlloyDB for PostgreSQL では、EMBEDDING という関数が実装されています。モデル名と変換したい元のテキストを引数として入力すると、ベクトルデータが返されます。返ってきたベクトルを列として保存したいので、一般的なリレーショナル データベースが持っている生成列（Generated Column）機能を用いて、自動的に値を生成して埋め込んでいきます。こうすることで、簡単にデータベースにエンベディングの列を追加することが可能です。

次に行うのが、実際のベクトル検索です。OSS 版 PostgreSQL の拡張機能「pgvector」を用いてベクトル検索を行います。pgvector には、データベースに格納するデータ型として「ベクター型」をサポートしており、さらにベクター型どうしの距離を計算するための演算子や ANN 近似値検索をするためのインデックスを生成する機能も搭載されています。

ベクトルどうしの距離を計算する演算子は複数あり、そのうちどれを使えばよいかは AI モデルにより異なります。ただし、どの演算子を用いたとしても、比較結果の値が小さいほど似ているということに変わりはありません。SELECT 文の ORDER BY でソートすれば、近い順に並びます。

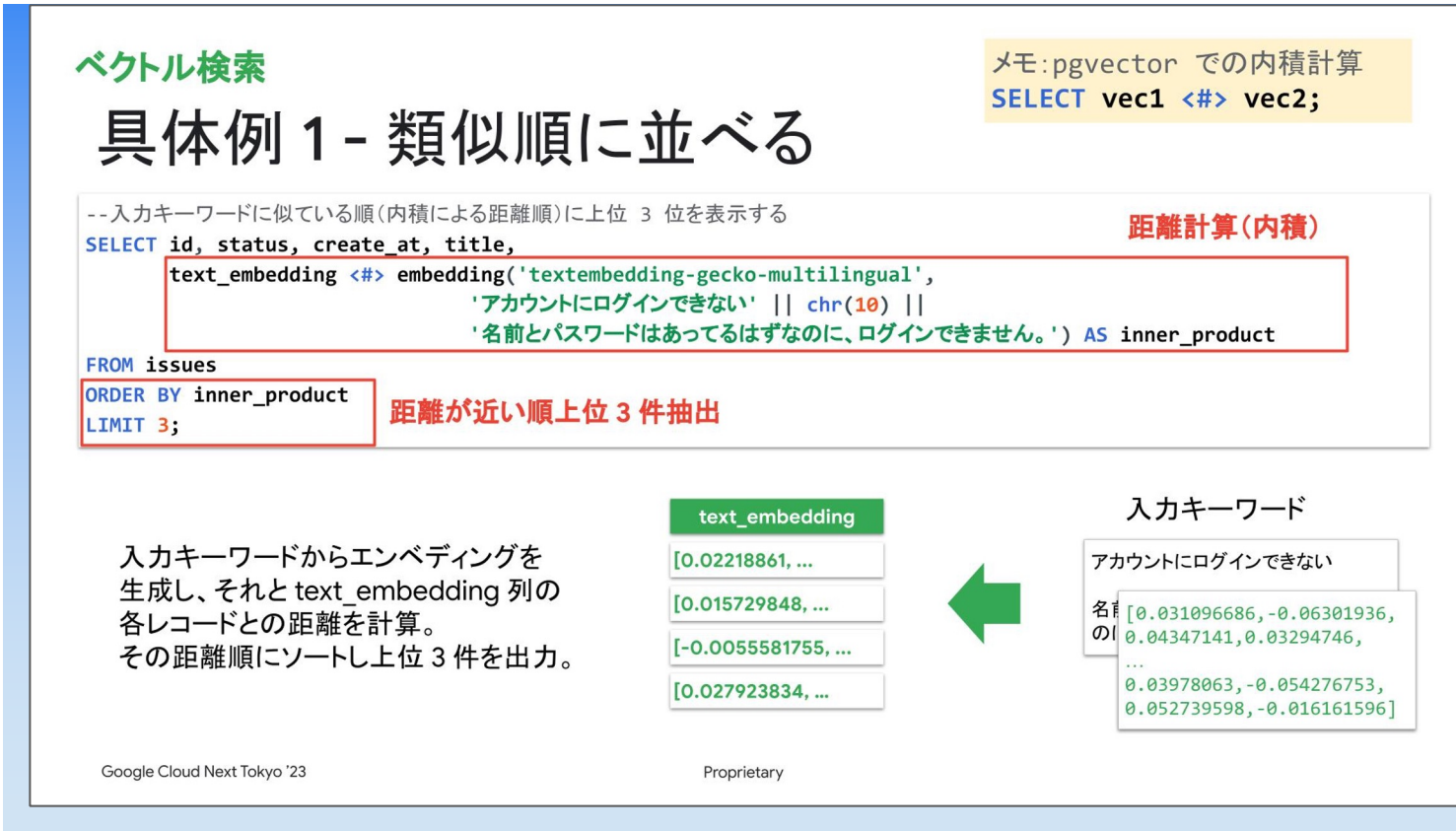


ベクトルどうしの距離を比較する演算子は複数存在する

今回は、pgvector を例に取り説明していますが、ベクトル検索はデータベースによって実装方法が異なる点に注意してください。例えば、BigQuery の場合は、BigQuery から機械学習を扱う BigQuery ML の中の ML.DISTANCE 関数を用いて距離を測ります。

ベクトル検索と構造化データによるフィルタリングを組み合わせる

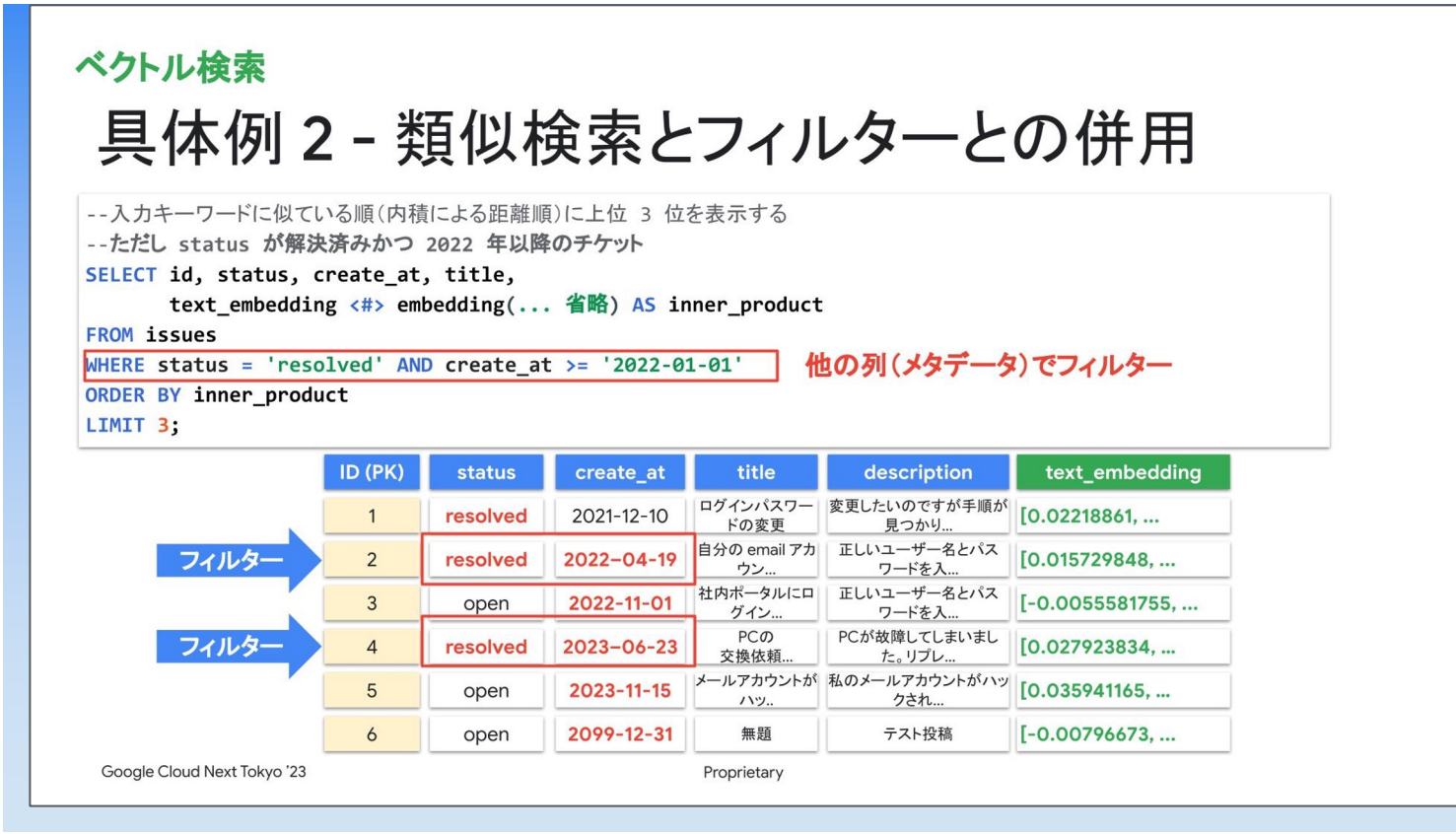
この機能を踏まえ、先ほどのヘルプデスクの例にて実際の記述を見ていきましょう。まず問い合わせのチケットが入っているテーブルがある前提とし、下図の SELECT 文では、ID、チケットが解決済みか否かどうかのステータス、チケットの作成日時、タイトル・説明（問い合わせの内容）のほかに、今回のテーマであるエンベディングの比較をしています。



ベクトル間の内積を計算して類似度を調べるサンプル

具体的には「embedding」の引数に問い合わせのテキストを埋め込んで API をコールし、返されたエンベディングと既存のテーブル内に格納されたエンベディングを比較します。最後に ORDER BY で上位3件を抽出しています。

上記の例に、フィルタリングを加えた例も紹介します。下図では、WHERE 句を追加して他の列を用いたフィルタリング、つまりベクトルのメタデータとデータベース内の構造化データを併用した検索を行っている例です。この例では、作成日の列（create_at）と解決済みか否か（status）の列をフィルタリング条件に含むことで、直近 2 年かつ解決済みのチケットが抽出されます。



データベース内の既存の情報を用いてベクトル検索と他の条件と組み合わせるイメージ

同様に、キーワード検索と併用することももちろん可能です。上記の WHERE 句の条件に件名（title 列）を追加して絞り込むことで、より類似度の高いチケット情報を抽出することができるでしょう。

既存のシステムにベクトル検索を組み合わせる発想を

今回は、AlloyDB for PostgreSQL によるリアルタイム処理を例に取りましたが、選択肢はさまざまです。アプリケーションの裏のシステムとして使うのであれば AlloyDB for PostgreSQL の場合もありますし、既に BigQuery をお使いの方で、バッチ処理としてベクトル検索をすることも考えられます。

そのため、「ベクトル検索をしたいけれども、どのデータベースを選べば良いのか、どのサービスを選べば良いのか」という声もよく聞かれますが、ベクトル検索をするためにサービスを選ぶのではなく、普段、既に使っているサービスにベクトル検索が加わることで、どのようなことができるかと考えてみるのも有効です。

もちろん、データベースの技術ですべてを解決する必要はないので、Vertex AI Vector Search なども活用していただきたいと思います。いずれにしても、現在では、従来と比べてより手軽にベクトル検索ができるようになっているので、ぜひ自社の業務課題の解決にお役立ていただければと思います。

参照リンク

1. Cloud FinOps の概要
2. AI と ML のソリューションの概要
3. AI と機械学習のプロダクトの概要
4. 今さら聞けない！ベクトル検索超入門 ～データベース ユーザーは何をおさえておけばいいのか？～ アーカイブ動画視聴ページ

製品、サービスに関するお問い合わせ

goo.gl/CCZL78

Google Cloud の詳細については、上記 URL もしくは QR コードからアクセスしていただくか、同ページ「お問い合わせ」よりお問い合わせください。

Google Cloud