15 The Semantic Index

Consider this chapter optional, because the Semantic Index is behind the scenes from administrators, security practitioners, and especially, users.

Laypeople paying attention to Microsoft marketing are more commonly hearing about WorkIQ. That's a branded amalgamation of the data, memory, and skills available to M365 Copilot. The Semantic Index is a key component of WorkIQ.

How the Semantic Index Works

The semantic index converts information from the Microsoft Graph into vectors. Vectors are numerical values (complex combinations of 1s and 0s) that represent data (about your emails, chats, meeting transcripts, and files). Vectors make it much easier for a computer to process numbers than words. This is very important when dealing with a lot of data, like years of email!

Think of the semantic index as a highly skilled librarian in a vast digital library. This librarian doesn't just organize books by their physical location; *they understand the meaning behind each book*. When you ask a question, they guide you to the most relevant content from *many* books, based on context and concepts. They recognize hidden connections and tailor their recommendations to individual preferences.

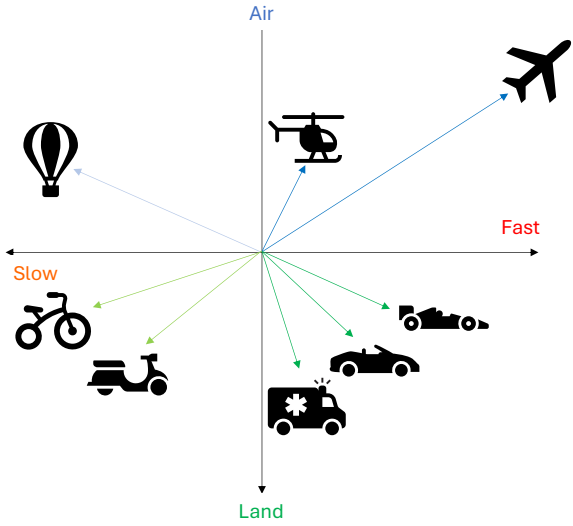
Data Sources

The semantic index doesn't house any data by itself. It crawls through data in the Graph, creating vectorized indices, which are like a map of the concepts within the data. **These vectors can represent words, images, or other types of data.**

To Vectors, Relationships Matter

In the oversimplified example below, each line is a vector, showing the relationship of a type of travel and its speed. Air vehicles appear above the x-axis, and ground vehicles below the x-axis. The tip of each vector would have a mathematical representation for the software to quickly process. For instance, the airplane might have coordinates (5,4) while the race car is located at (3, -1). The car is not nearly as fast relative to the plane, but it's much faster relative to the other land vehicles. Let that sink in – *it's the context or relationships amongst words in the index that helps an LLM be so good at predicting what to create next.*

The relationship or "closeness" of each entity is important. The slower land vehicles on the bottom left can also make up another grouping, "2-wheelers" compared to those on the right with four. LLMs use groupings to decide what letter / word is most logically next.



This visual is oversimplified since it's only two-dimensional. *LLMs use thousands of vectors*. It gets difficult to conceptualize, but imagine a *third axis coming in and out of the page*, categorizing "water" vehicles. Out to the right and coming towards you would be speedboats, and over to the left and away from you would be canoes and paddleboards, with the latter two being grouped closely since they're slow and manually powered.

That'd be three dimensions, and unless you took collegiate courses in Matrix Algebra, any more are hard to imagine. Yet a **semantic index can use thousands of vectors** to represent all the words, sentences, documents, presentations, and other content it knows about. Unlike traditional keyword search engines, **semantic search indexes learn implicitly**, adjusting their internal parameters (weights) and connectedness between words and concepts.

Vectors, their semantic relationships, and their contextual connectedness **send a much richer prompt to the LLM than what was originally typed**.

What Microsoft 365 data gets indexed?

Two kinds of data are indexed.
1. **User-level indexing** creates a personal index for each user about the set of data that they work with. This includes any text-based content that they create or engage with, such as their emails, Teams meeting transcripts, their documents in OneDrive, or documents that they comment on or *that are shared with them*.
2. **Tenant-level indexing** adds SharePoint Online files that are available to two or more people in the organization. However, it only shows the results to a user if the user already has access to the content. Access is controlled by SharePoint permissions, and the SharePoint Online site must be searchable. **There may be valid reasons to exclude a site from indexing, which is covered in Chapter 17, "Securing Copilot."**

Also note: Delegated Mailboxes and Shared Mailboxes are indexed, as are natively Archived Mailboxes, but Archived SharePoint Data and other data using M365 Backup are not indexed.

Where is the index stored?

Again, there is a difference, and each is stored in a different place:
1. The User-level index is stored in the user's Exchange Online mailbox.
2. The Tenant-level index is stored in a separate and secure container within the customer's tenant. It's located in the region where the SharePoint site is located.

When and how long does indexing occur?

The semantic index starts crawling a person's Graph data once they are licensed with Microsoft 365 Copilot.[1] It can take a while (some people have seen it take >48 hours), especially if the person has a long tenure and / or stores a lot of content.

After the semantic index finishes indexing for the first time, documents created by users are indexed almost immediately in the user's mailbox. New documents that are added to SharePoint Online sites that are *shared with two or more users* are indexed every day. When a document that has been indexed at the user or tenant level is changed, the updates are indexed right away.

You've made it through the most complex topic in the book! Read on to gain a better understanding of why LLMs may be wrong sometimes.

[1] https://learn.microsoft.com/en-us/microsoftsearch/semantic-index-for-copilot#index-updates