Supporting reading material for Week#2

Notion of Relevancy

DJ Hillman 1964

An extensional interpretation is suggested whereby a **concept** is associated with a **class of conceptually-similar documents.**

What is a model and what is a framework?

A framework is a set of guiding principles, concepts, and tools that provide a structure for organizing and understanding a subject or problem. Models are often derived from theoretical frameworks and can be used to test hypotheses or make predictions about real-world phenomena.

A framework is like a set of rules or ideas that helps us understand something better and figure out how to solve problems. A model is like a mini-version of that framework that we can use to test ideas or predict what might happen in the real world.

Framework: Think of **G**uiding **P**rinciples **O**rganize **S**olutions (**GPOS**).

Model: Think of **P**redicting **D**ata **T**ests **A**ctuals (**PDT-A**).

Framework: Organizing a birthday party (plan, gather supplies, execute the plan).

Model: Predicting how much time you need to finish a homework assignment based on past experiences.

Title: "The Great School Fair Project"

Once upon a time in a lively school, Ms. Harper's class was tasked with organizing the annual school fair. To tackle this big project, they used a **framework** and a **model**.

The Framework: The Fair Planning Guide

Ms. Harper introduced them to the **Fair Planning Guide**, which was their framework. It included a set of steps to help them organize everything:

Set the Goal: Decide what they wanted to achieve with the fair (fun activities and a successful event).

Plan the Steps: Break down the project into smaller tasks like setting up booths, arranging food, and organizing games.

Gather Resources: Collect the materials needed, such as decorations, supplies for booths, and a list of volunteers.

Execute the Plan: Start working on each task as scheduled.

Review and Adjust: Check how things are going and make changes if needed.

Complete the Project: Finish all preparations and ensure everything is ready for the fair.

The Model: Predicting Booth Popularity

To make sure the fair would be a hit, the class used a **model** to predict which booths would be the most popular. They named their model "Booth Buzz Predictor."

Define Variables: They identified factors affecting booth popularity, such as the type of activity (games or crafts), the location of the booth, and how much promotion it got.

Collect Data: They looked at previous fairs to see which booths had been most popular. They also surveyed students to find out what activities they were excited about.

Apply the Model: Using this information, they made a simple prediction formula. For example, if a booth was in a high-traffic area and had a fun game, it was likely to attract more kids.

Test the Prediction: On the day of the fair, they watched how many students visited each booth and compared it with their predictions.

Adjust: They noticed that booths with interactive activities were more popular than they had predicted, so they adjusted their future planning to include more interactive elements.

The Result

The fair was a huge success! By following their **Fair Planning Guide** (framework), they stayed organized and ensured every aspect of the fair was covered. Using their "Booth Buzz Predictor" (model), they were able to make smart decisions about which activities to emphasize. The students learned that a good **framework** helps them organize and manage tasks, while a well-designed **model** helps them make better predictions and adjustments.

And that's how Ms. Harper's class used a framework and a model to make their school fair an unforgettable event!

In the context of Information Retrieval (IR), the sentence means that the models used in IR are specifically designed for searching, retrieving, and organizing information from large sets of data. They have unique features and algorithms tailored for this purpose, which makes them different from general-purpose software that is used for a wide range of other tasks.

What is a Physical Analog Model?

A physical analog model is like a small version of something bigger or a simple object that helps us understand how something works in the real world. It's a model that physically looks like or behaves like the thing it represents.

Simple Example: The Water Cycle Model

Let's use the **water cycle** as an example:

What is the Water Cycle?

The water cycle is how water moves around the Earth. It goes from the ocean or lakes (where it's water), turns into clouds (as vapor), falls back to Earth as rain or snow, and then goes back into the water bodies

A ranking algorithm is **a procedure that ranks items in a dataset according to some criterion**. Ranking algorithms are used in many different applications, such as web search, recommender systems, and machine learning.

Retrieval models most frequently associated with distinct combinations of a document logical view and a user task

Ad hoc retrieval. standard retrieval task in which the user specifies his information need through a query which initiates a search (executed by the information system) for documents which are likely to be relevant to the user.

In ad-hoc system, we compare the coming queries with the documents collection available. On the other hand, filtering retrieval is based-on comparing the incoming documents with those queries specified in each user profile.

Filtering is **based on descriptions of individual or group information preferences, or profiles, that typically represent long-term interests**. Filtering also implies removal of data from an incoming stream rather than finding data in the stream; users see only the data that is extracted.

gi : A function returns the weight associated with ki in any t-dimensional vector ($g_i(d_j) = w_{i,j}$)

gi: This represents a function, denoted as "g sub i," which takes as input an index term ki and returns a weight associated with that specific index term. In other words, it's a function that calculates or retrieves the weight (wi,j) for a particular index term.

gi(dj): When you see "gi(dj)," it means that the function gi is applied to a document dj. In the context of Information Retrieval (IR), this function gi is used to determine the weight (wi,j) of a specific index term ki within a particular document dj.

wi,j: This represents the weight associated with the index term ki in the document dj. It quantifies the importance or relevance of the index term within that specific document. The weight wi,j is typically a numerical value.

t-dimensional vector: The expression refers to a vector space model where index terms and their weights are organized in a multidimensional space. The "t" denotes the number of dimensions in this vector space, usually representing the total number of unique index terms in the collection.

In summary, the expression describes a function gi that calculates the weight (wi,j) of a specific index term ki within a document dj, and this weight is used to represent the relevance of the index term within the document in a t-dimensional vector space. This is a fundamental concept in Information Retrieval for ranking and retrieving documents based on their relevance to user queries.

The IR models can be categorized as Classical Information Retrieval models, Non-Classical Information Retrieval models and Alternative models for Information Retrieval.

In Boolean model, the IR system retrieves the documents based **on the occurrence of query key words in the document**. It doesn't provide any ranking of documents based on the relevancy.

The Euclidean norm (magnitude) of the query vector (||Q||) and the Euclidean norm (magnitude) of the document vector (||D1||) are used in the cosine similarity formula to normalize the vectors. Normalization is a crucial step when calculating cosine similarity because it ensures that the similarity score remains in the range of -1 to 1, making it easier to interpret and compare.

Here's why we use these norms in the formula:

Normalization: By dividing the dot product of the vectors by the product of their magnitudes, we effectively normalize the similarity score. Normalization accounts for the differences in the overall length (magnitude) of vectors. Without normalization, longer vectors would tend to have higher dot products simply due to their length, which might not accurately reflect the similarity between the vectors.

Range: Cosine similarity measures the cosine of the angle between two vectors. When normalized, the cosine similarity score always falls within the

range of -1 (perfectly dissimilar) to 1 (perfectly similar). A score of 0 indicates orthogonality (no similarity). This standardized range allows for easy comparison and ranking of documents based on relevance to a query.

Independence from Vector Length: Cosine similarity is independent of the absolute length of the vectors. It only depends on the direction of the vectors in the vector space. This means that the similarity score remains consistent even if the length of the vectors (i.e., the number of terms in documents or queries) changes.

In essence, the Euclidean norms ensure that the similarity score is a meaningful and interpretable measure of how similar or dissimilar the query and document vectors are, irrespective of their lengths. This is especially important in Information Retrieval when ranking documents based on relevance to user queries, as it allows for fair and consistent comparisons.