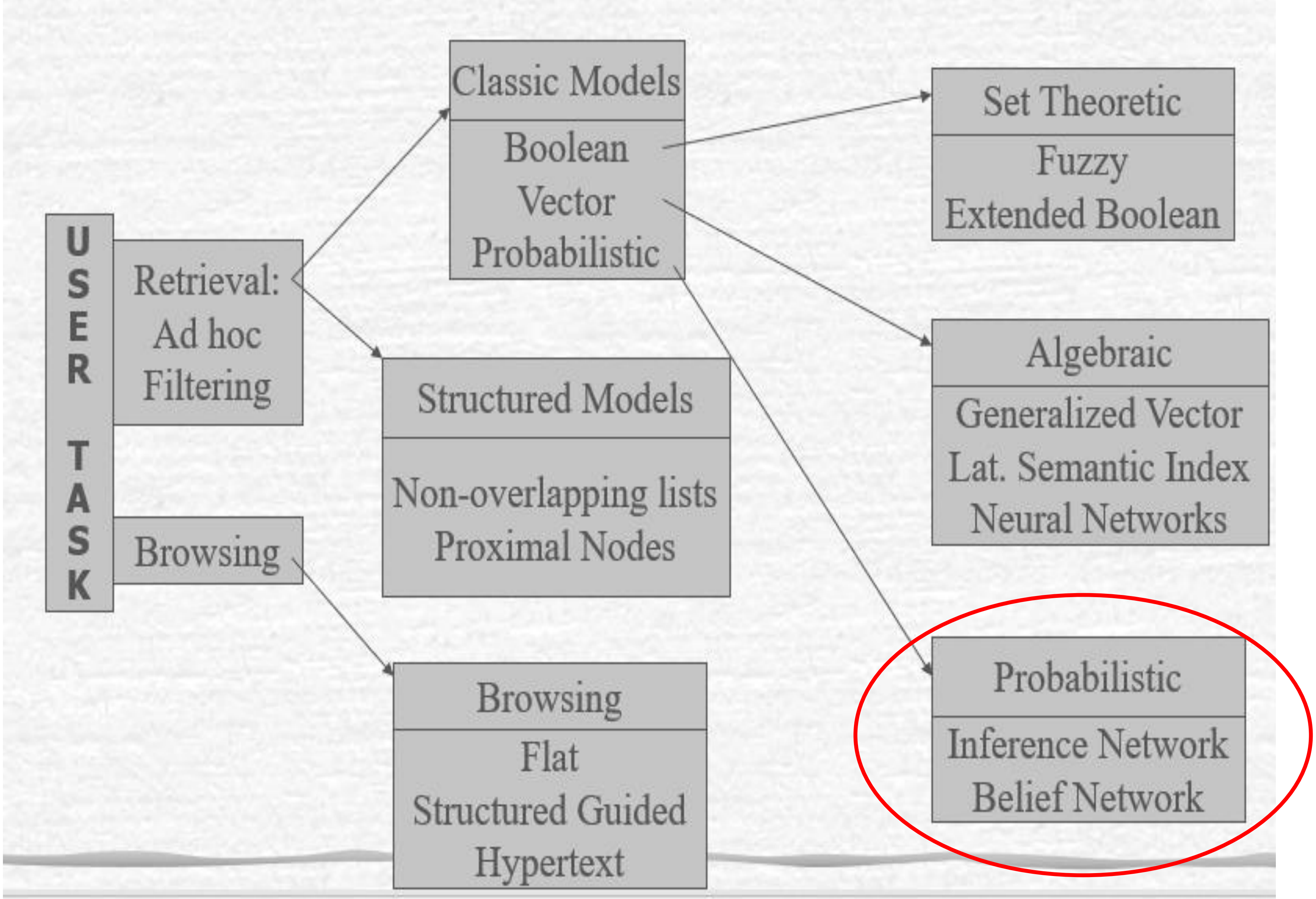
Information Retrieval

By

Dr Syed Khaldoon Khurshid



USER TASK

Retrieval:
Ad hoc
Filtering

Browsing

Classic Models
Boolean
Vector
Probabilistic

Structured Models
Non-overlapping lists
Proximal Nodes

Browsing
Flat
Structured Guided
Hypertext

Set Theoretic
Fuzzy
Extended Boolean

Algebraic
Generalized Vector
Lat. Semantic Index
Neural Networks

Probabilistic
Inference Network
Belief Network

Lecture of the week

- Probabilistic
  - Interference Network
  - Belief Network

Probabilistic Information Retrieval (IR) models

Probabilistic Information Retrieval (IR) models

- Probabilistic Information Retrieval (IR) models build on classical IR models by introducing the concept of uncertainty and probability. In these models, we consider **the likelihood of a document being relevant to a query as a probability**. Two key concepts in this extension are **“Inference"** and **"Belief Networks“.**
- **Inference:** In probabilistic IR, inference refers to the process of reasoning and making decisions based on probabilities. **It involves calculating the likelihood of a document being relevant given the information available.** Inference allows us to handle uncertainty when determining the relevance of documents.

Probabilistic Information Retrieval (IR) models

- **Belief Network:** A Belief Network, also known as a **Bayesian Network**, is a graphical representation of how different pieces of information are connected in a probabilistic model. It shows **the relationships between variables and how they influence each other**. In the context of IR, a Belief Network helps us model and understand the complex interplay of factors that contribute to document relevance.
- In simpler terms, **probabilistic IR models add a layer of probability to classical IR, enabling us to account for uncertainty and make informed decisions about document relevance.** Inference and Belief Networks are tools that help us navigate this probabilistic landscape, allowing us to find the most relevant documents effectively.

Simple Example of Inference

- Imagine you're on a treasure hunt! You have a map with different locations marked on it, and you want to find out which places might have treasure.
- **1. Treasure Clue: You heard a rumor that treasures are usually found near "old trees" and "rivers."**
- **2. Checking Locations: You visit three spots:**

  -Spot A: An old tree by a river.

  -Spot B: A field with no trees.

  -Spot C: A new park with some trees but no river.
- **3. Making Decisions:**

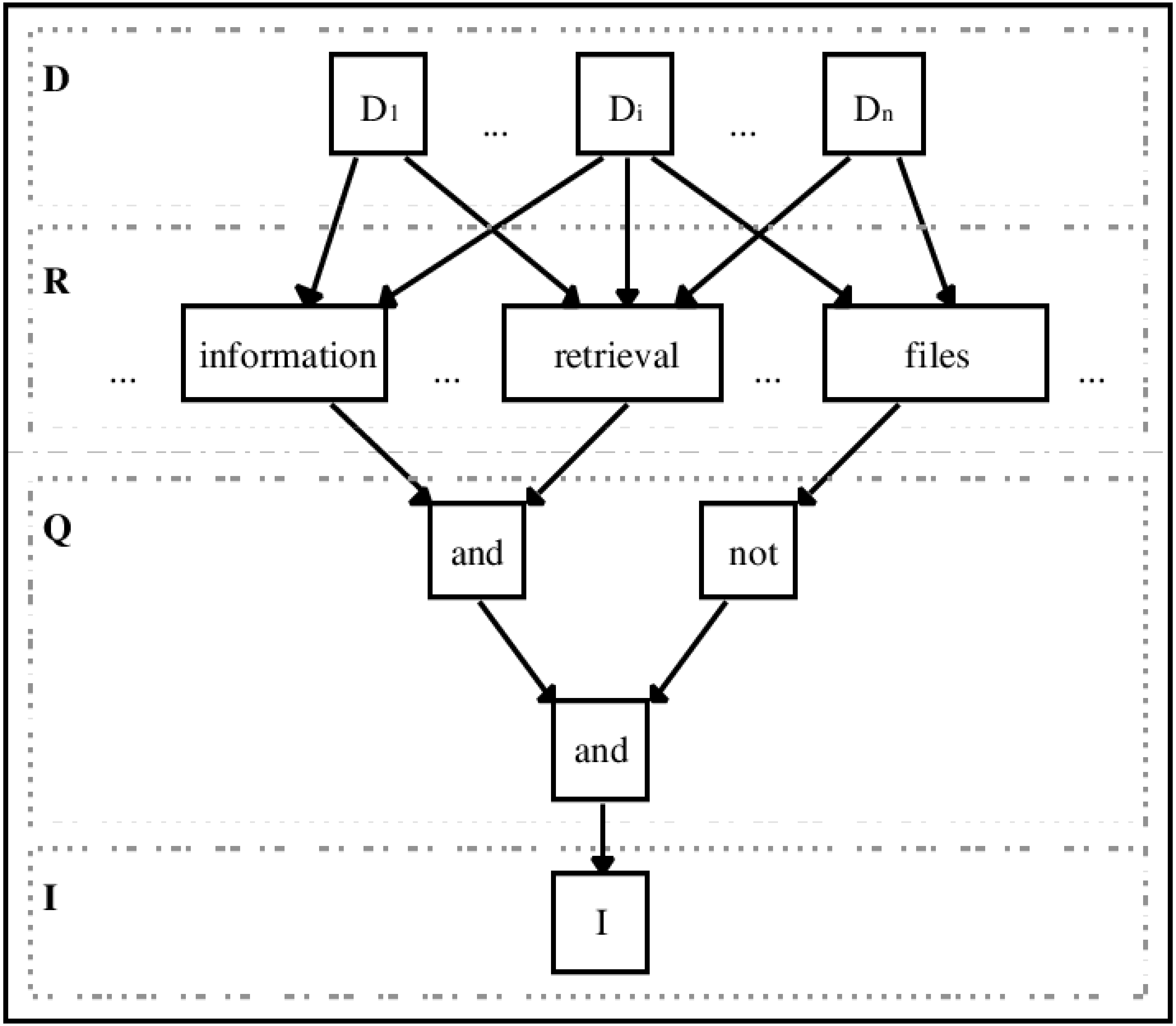
  - Spot A: Looks like a great place for treasure because it has both clues (an old tree and a river). So, you think it has a high chance of having treasure.

  - Spot B: This spot has no trees at all, so you think it probably doesn't have treasure.

  - Spot C: This spot has a tree but no river. It's not as likely to have treasure as Spot A, but it might still be worth checking.

Conclusion

- **Using inference is like using clues to decide where to search for treasure. You think about what you know (the clues) and then decide where to look based on how likely each spot is to have treasure. In the same way, when searching for information, we use clues (or probabilities) to figure out which documents are most likely to be useful!**



An example inference network created from the query (information and retrieval) and (not files).

Ref: Cunningham, Sally & Holmes, Geoffrey & Littin, Jamie & Beale, Russell & Witten, Ian. (1999). Applying Connectionist Models to Information Retrieval.

Simple real-life examples

- **Let's use a simple real-life example to illustrate the concepts of inference and Belief Networks in the context of Information Retrieval (IR).**
- **Scenario:** Imagine you are searching for a new Smartphone online, and you want to find the one with the best camera. You come across several reviews and user opinions. Here's how inference and a Belief Network could be applied:

1. **Inference:** In this case, **inference would involve assessing the probability of a Smartphone having a good camera based on available information**. You might consider factors like camera specifications (e.g., megapixels), user reviews, and expert opinions. **The probability of a Smartphone having a good camera is calculated based on these factors.**

Simple Example of a Belief Network

- Let’s say you’re a detective trying to solve a mystery about who might have taken the last cookie from the cookie jar.

1. Clues:

- Who was in the kitchen? (This is like the Query Node)

- Who likes cookies?(This is like the Document Node)

- Who actually took the cookie? (This is like the Relevance Node)

2. Gathering Information:

- People in the Kitchen: You know three friends were in the kitchen: Alice, Bob, and Charlie.

Who Likes Cookies:

- Alice likes cookies a lot.

- Bob likes them a little.

- Charlie doesn’t like cookies at all.

Simple Example of a Belief Network

Probabilities:

- If Alice was in the kitchen, there’s a very high chance she took the cookie (let's say 90%).
- If Bob was in the kitchen, there’s a smaller chance (about 50%).
- If Charlie was in the kitchen, the chance is very low (around 10%).

3. Making Connections:

- The kitchen activity influences the likelihood of who took the cookie.
- Based on who was there and who likes cookies, you can use this information to figure out who is most likely the cookie thief.

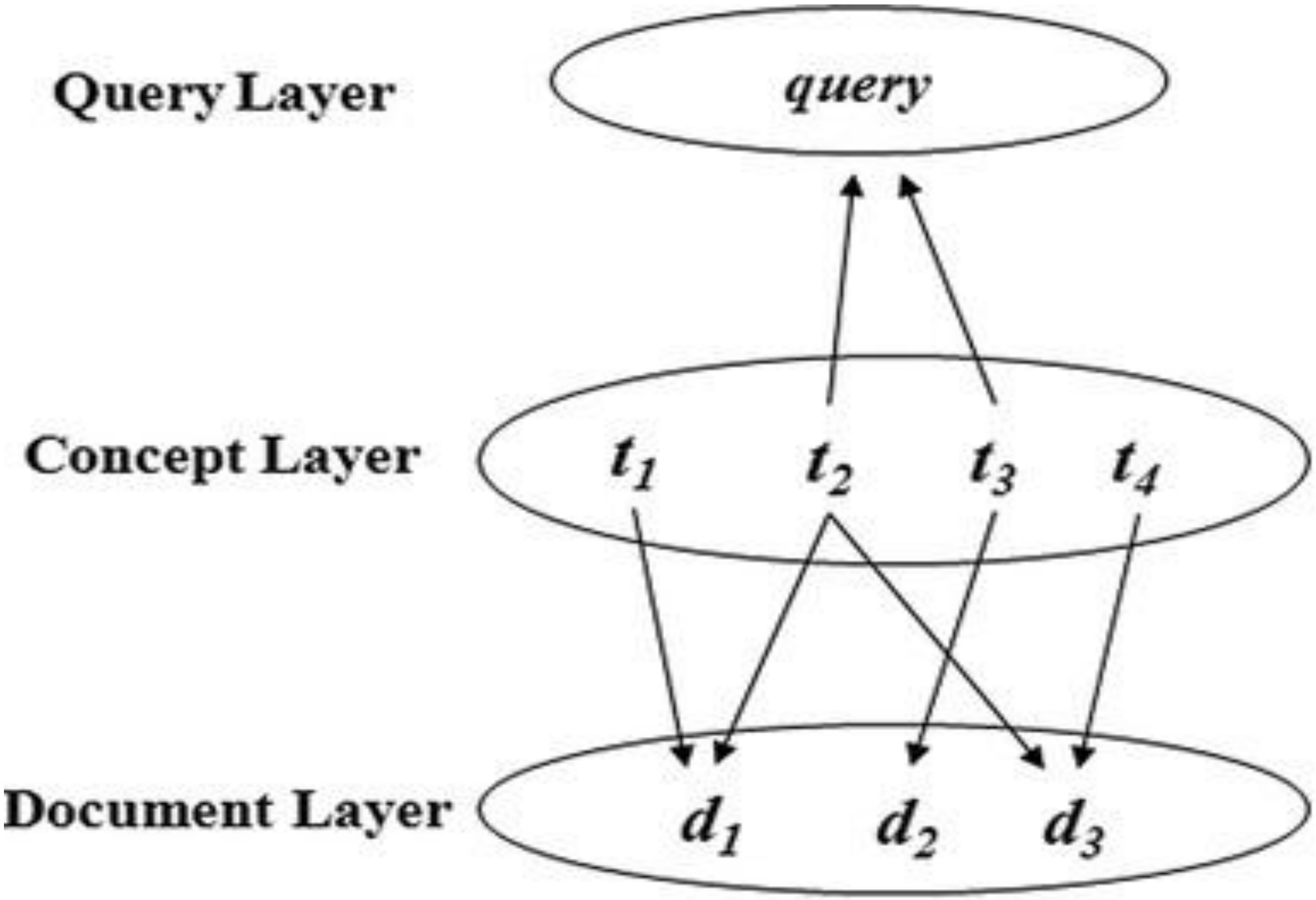
Conclusion

- A belief network helps you connect different pieces of information. In this case, you combine the fact that someone was in the kitchen with how much they like cookies to figure out who probably took the last cookie. **Just like in a belief network, each clue influences the final decision about who the cookie thief might be!**

Simple real-life examples

2. **Belief Network:** A Belief Network can be visualized as a graph. In this scenario, it would have nodes representing various factors:

- **Node 1: Camera Specifications (e.g., megapixels, aperture size)**
- **Node 2: User Reviews (positive or negative sentiment)**
- **Node 3: Expert Opinions (e.g., from tech blogs)**
- **Node 4: Overall Camera Quality (the variable we want to predict)**
- **The Belief Network shows how these factors are connected.** For example, camera specifications may influence overall camera quality, while user reviews and expert opinions might also contribute to the final assessment. **The connections indicate the influence or dependence of one variable on another.**



Bayesian belief network (BBN)-based model for information retrieval

**Ref: Lee, Jae-won & Lee, Sang-goo & Kim, Han-joon. (2011).
A probabilistic approach to semantic collaborative filtering using world
knowledge. J. Information Science. 37. 49-66. 10.1177/0165551510392318.**

Simple real-life examples

- Using this Belief Network, you can model how different factors affect your belief in the overall camera quality of a Smartphone. Inference techniques would then be used to calculate the probability of a Smartphone having a good camera based on the observed values of these factors.

- In essence, Belief Networks and inference help you make more informed decisions about the camera quality of smart phones, similar to how **probabilistic models in IR help in ranking and retrieving documents based on their relevance to a query while considering various sources of uncertainty.**

Computational Example of Inference

- **Let's create a simple computational example of inference in the context of Information Retrieval (IR) using text data:**

  Scenario: We have a small document collection related to technology, and we want to infer the relevance of a document to a user query. In this example, we'll use a basic inference model.

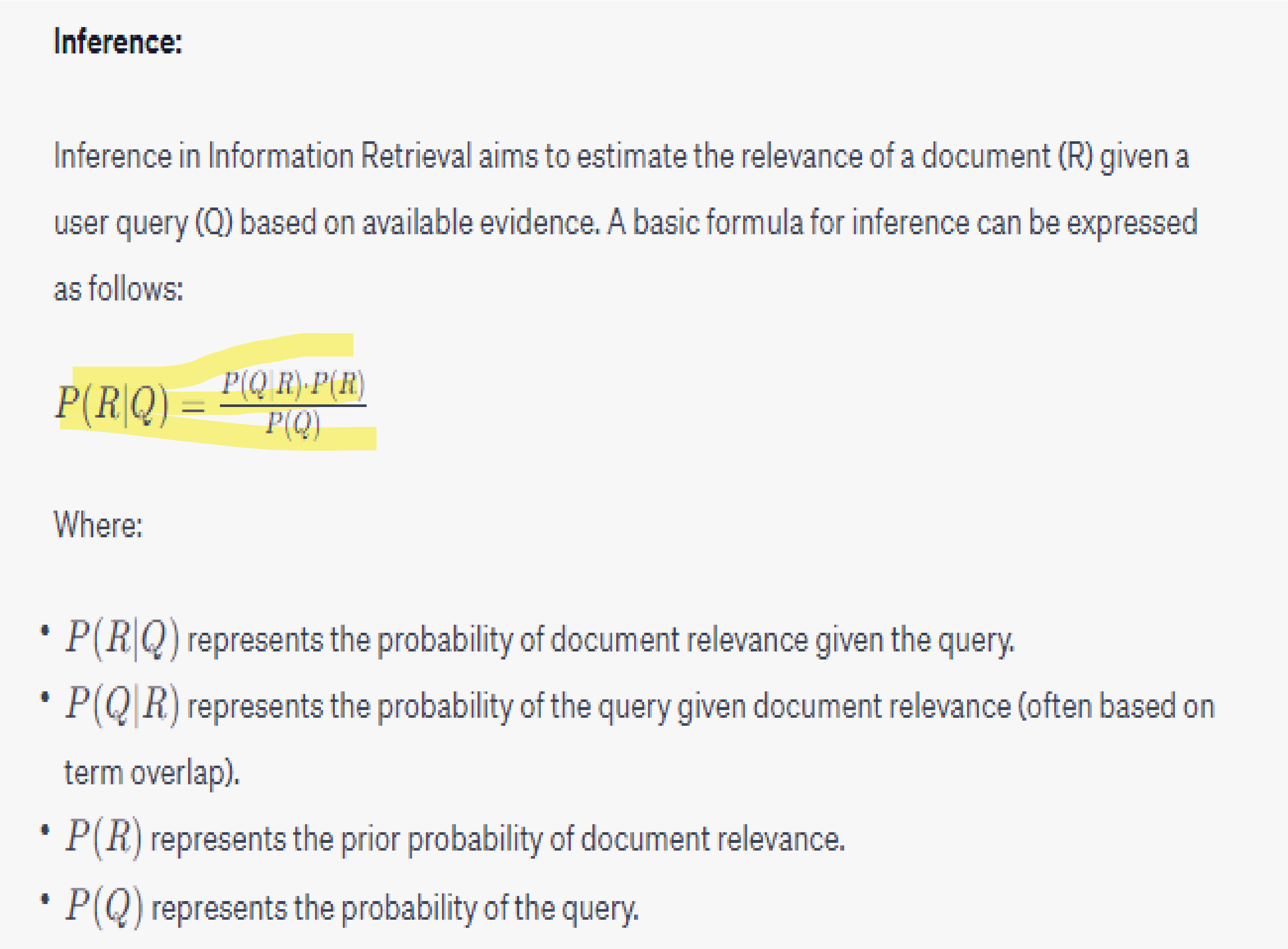
  Nodes:

1. **Query Node (Q):** Represents the user's query.
2. **Document Node (D):** Represents a document.
3. **Relevance Node (R):** Represents the relevance of the document to the query.

Computational Example of Inference

Probabilities:

1. **P(Q):** The probability distribution of the user's query. For simplicity, we'll assume equal probability for all queries.
2. **P(D|Q):** The probability distribution of a document given the query. We'll use a simple word **overlap approach**, **where the more words in the query that appear in the document, the higher the probability.**
3. **P(R|D, Q):** The probability distribution of document relevance given both the document and the query. We'll use a **threshold-based approach**, **where if the document contains a certain percentage of query words, it's considered relevant.**



Inference:

Inference in Information Retrieval aims to estimate the relevance of a document (R) given a user query (Q) based on available evidence. A basic formula for inference can be expressed as follows:

$$P(R|Q) = \frac{P(Q|R) \cdot P(R)}{P(Q)}$$

Where:

- $P(R|Q)$ represents the probability of document relevance given the query.
- $P(Q|R)$ represents the probability of the query given document relevance (often based on term overlap).
- $P(R)$ represents the prior probability of document relevance.
- $P(Q)$ represents the probability of the query.

Computed Example of Inference

Scenario:

- We have a small document collection related to travel, and we want to infer the relevance of a document to a user query about travel destinations.

Nodes:

- **Query Node (Q)**: Represents the user's query (e.g., "beach vacation").
- **Document Node (D)**: Represents a document in the collection (e.g., "document1").
- **Relevance Node (R)**: Represents the relevance of the document to the query.

Conditional Probabilities:

- **P(Q)**: The probability distribution of the user's query. For simplicity, we will assume equal probability for three queries:
  - P(Q = "beach vacation") = P(Q = "mountain hiking") = P(Q = "city tour") = 1/3.

Computed Example:

- **P(D|Q)**: The probability distribution of a document given the query. We will use word overlap:
- P(D = "document1" | Q = "beach vacation") = 0.9 (document1 mentions many beach-related terms).
- P(D = "document2" | Q = "beach vacation") = 0.5 (document2 has some relevant terms but is less focused).
- **P(R|D, Q)**: The probability distribution of document relevance given both the document and the query. We'll use a threshold-based approach:
- P(R = Relevant | D = "document1", Q = "beach vacation") = 1 (document1 is highly relevant).
- P(R = Relevant | D = "document2", Q = "beach vacation") = 0.5 (document2 is somewhat relevant).

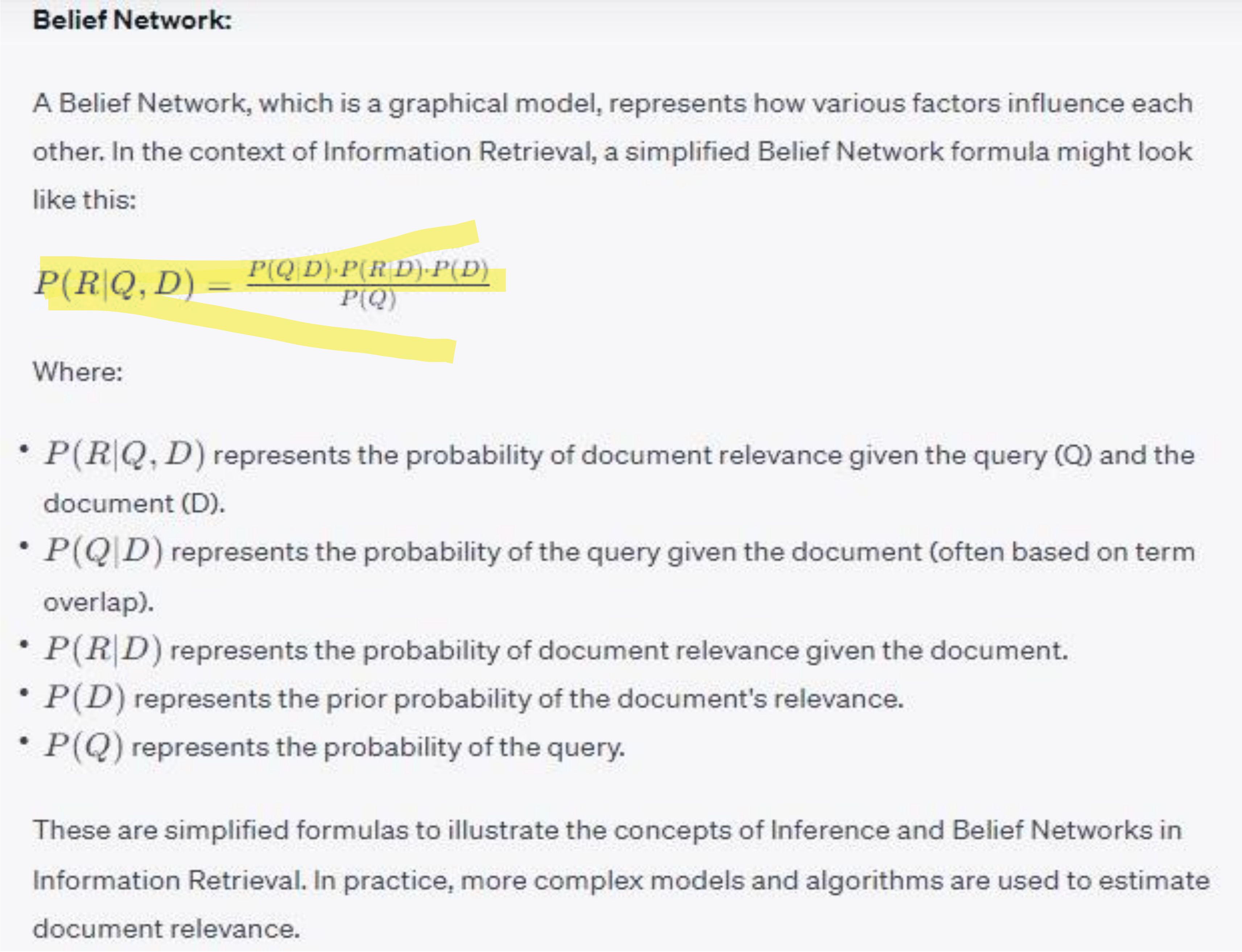
Conclusion:

- In the context of the provided example focused on the query "beach vacation," the inference model reveals the following key insights:
- **Relevance Assessment**:
  - The model evaluates the relevance of documents based on the probability of document content matching the user query. For instance, document1 shows a high relevance score of **0.9 (90%)**, indicating it is very likely to meet the user's needs.
- **Document Ranking**:
  - The inference model effectively ranks documents according to their relevance scores. Document1 is prioritized for retrieval due to its high score, while document2, with a score of **0.5 (50%)**, is deemed less relevant.

Conclusion:

- **Decision-Making Support**:
  - By utilizing calculated probabilities, the inference model assists in making informed decisions about which documents to present to the user, enhancing the relevance of search results and user satisfaction.
- **Optimization Opportunities**:
  - The model can be iteratively refined with additional data or user feedback, which can improve its accuracy and responsiveness to different queries over time.
- In summary, the inference model serves as a robust mechanism for evaluating and ranking document relevance, guiding effective information retrieval strategies based on user queries.

Computational Example of Belief Network



Belief Network:

A Belief Network, which is a graphical model, represents how various factors influence each other. In the context of Information Retrieval, a simplified Belief Network formula might look like this:

$$P(R|Q,D) = \frac{P(Q|D) \cdot P(R|D) \cdot P(D)}{P(Q)}$$

Where:

- $P(R|Q,D)$ represents the probability of document relevance given the query (Q) and the document (D).
- $P(Q|D)$ represents the probability of the query given the document (often based on term overlap).
- $P(R|D)$ represents the probability of document relevance given the document.
- $P(D)$ represents the prior probability of the document's relevance.
- $P(Q)$ represents the probability of the query.

These are simplified formulas to illustrate the concepts of Inference and Belief Networks in Information Retrieval. In practice, more complex models and algorithms are used to estimate document relevance.

Understanding the Mathematical Expression

- The expression which is being provided is a part of a probabilistic framework for Information Retrieval (IR) that utilizes belief networks or graphical models. Let's break down the components step by step.
- **1. (P(R | Q, D) :**

  - This represents the conditional probability of retrieving a relevant document (R) given a query (Q) and a document (D).

  - It answers the question: "What is the probability that document (D) is relevant to the query (Q)?"
- **2. (P(Q, D)):**

  -This is the joint probability of the query (Q) and document (D) occurring together.

  - It encompasses the likelihood of observing both (Q) and (D) simultaneously.

Understanding the Mathematical Expression

- **3. (P(R, D)):**

  - This term denotes the joint probability of document (D) being relevant (R) and the document itself.

  - It helps to quantify how likely it is to retrieve (D) when considering its relevance.

- **4. (P(D)):**

  - This is the marginal probability of the document (D).

  -It indicates how likely document (D) is to be selected or encountered, regardless of the query.

- **5. (P(Q)):**

  - This is the marginal probability of the query (Q).

  - It reflects the overall likelihood of the query occurring in the system.

Understanding the Mathematical Expression

Putting It All Together:

-The right side of the equation combines these probabilities to derive the conditional probability on the left.

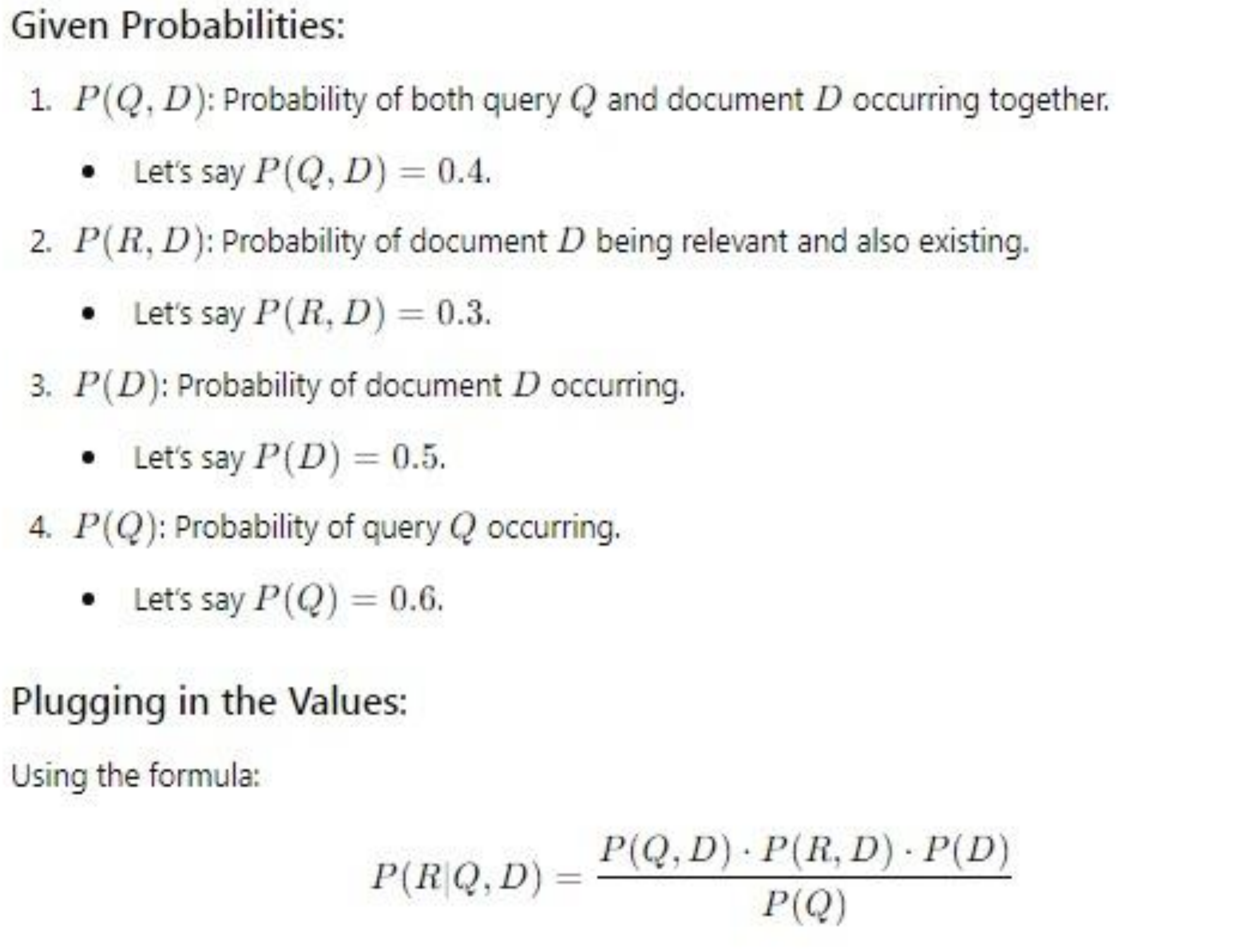
-By multiplying (P(Q, D)), (P(R, D)), and (P(D)), you're essentially incorporating the dependencies and interactions between the query, document, and relevance.

-Dividing by (P(Q)) normalizes this value, ensuring that the result is a valid probability (summing to 1 over all relevant documents).

Conclusion:

- This expression highlights how various probabilities interact within the belief network framework, **allowing for an understanding of document relevance in response to a query.** It can be particularly useful in developing models for information retrieval systems.

Computed Example of Belief Network



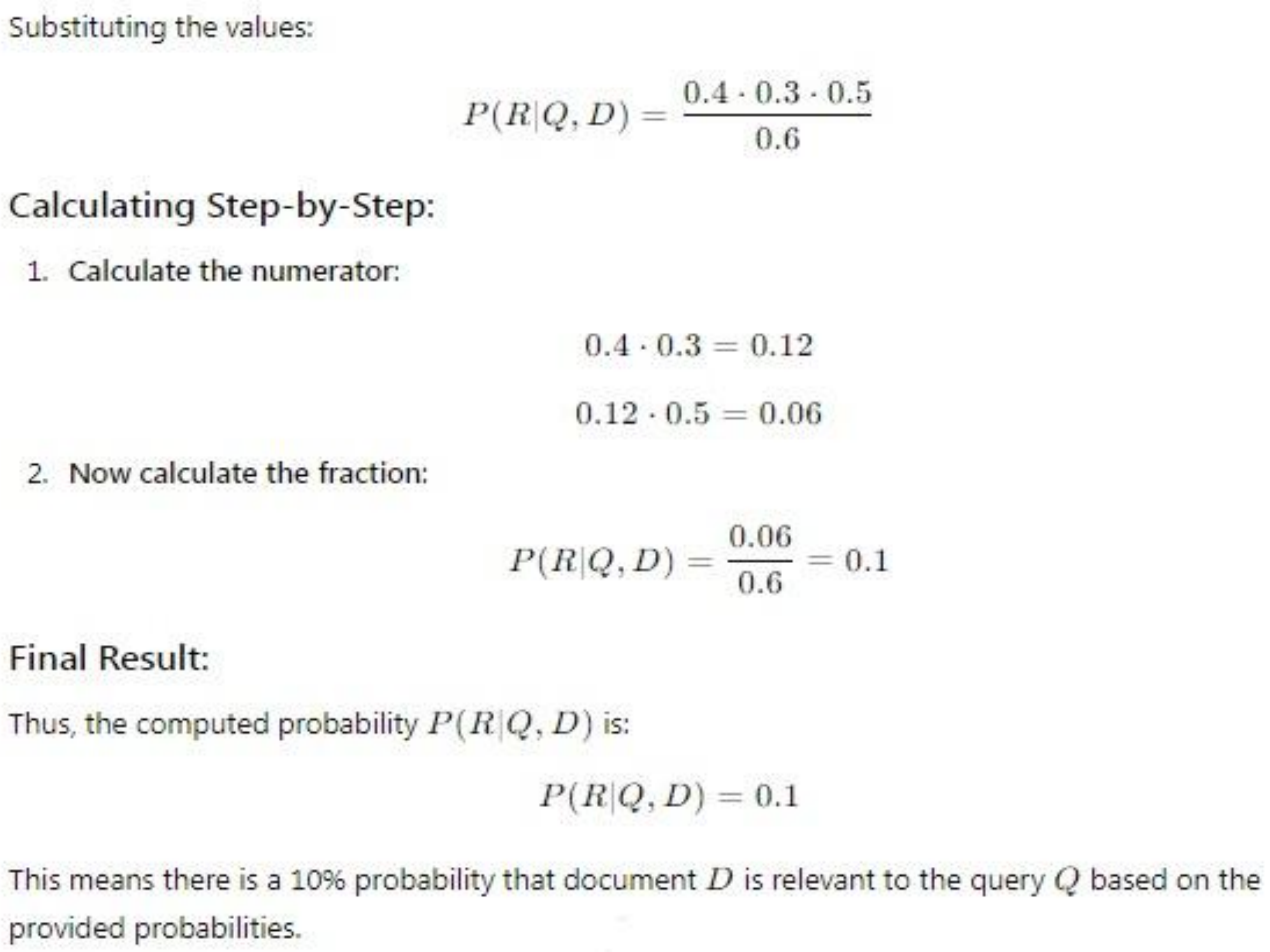
Given Probabilities:

1. $P(Q, D)$: Probability of both query Q and document D occurring together.
   - Let's say $P(Q, D) = 0.4$.
2. $P(R, D)$: Probability of document D being relevant and also existing.
   - Let's say $P(R, D) = 0.3$.
3. $P(D)$: Probability of document D occurring.
   - Let's say $P(D) = 0.5$.
4. $P(Q)$: Probability of query Q occurring.
   - Let's say $P(Q) = 0.6$.

Plugging in the Values:

Using the formula:

$$P(R|Q, D) = \frac{P(Q, D) \cdot P(R, D) \cdot P(D)}{P(Q)}$$



Substituting the values:

$$P(R|Q,D) = \frac{0.4 \cdot 0.3 \cdot 0.5}{0.6}$$

Calculating Step-by-Step:

1. Calculate the numerator:

$$0.4 \cdot 0.3 = 0.12$$

$$0.12 \cdot 0.5 = 0.06$$

2. Now calculate the fraction:

$$P(R|Q,D) = \frac{0.06}{0.6} = 0.1$$

Final Result:

Thus, the computed probability $P(R|Q,D)$ is:

$$P(R|Q,D) = 0.1$$

This means there is a 10% probability that document D is relevant to the query Q based on the provided probabilities.

Conclusions:

- Based on the explanations and the computed example regarding the conditional probability in the belief network for information retrieval, we can draw several conclusions:

1. Understanding Relevance:

- The expression (P(R | Q, D)) quantifies the likelihood of a document being relevant to a specific query. In our example, this probability was calculated as 0.1 (or 10%), indicating a relatively low chance of relevance based on the given data.

2. Importance of Probabilities:

- The computation illustrates how various probabilities interact:

- Joint probabilities (P(Q, D)) and (P(R, D)) provide insights into the relationships between queries and documents.

- Marginal probabilities (P(D)) and (P(Q)) help to contextualize these relationships within the overall system, affecting the final calculation.

3. Influence of Data:

- The values assigned to the probabilities significantly impact the outcome. Small changes in (P(Q, D)) or (P(R, D)) could lead to different conclusions about the relevance of documents. This highlights the importance of accurate and representative data in information retrieval systems.

4. Use in IR Systems:

- This framework can guide the development of more sophisticated information retrieval systems. By modeling relationships probabilistically, systems can better understand user queries and the relevance of documents, potentially improving search results.

5. Probabilistic Reasoning:

- The use of probability allows for uncertainty in information retrieval, reflecting real-world scenarios where relevance is not always clear-cut. This reasoning is essential for designing systems that are robust and user-friendly.

Conclusion of probabilistic models :

- Overall, the expression and its calculation demonstrate the power of probabilistic models in evaluating document relevance in response to user queries. These models provide a structured approach to handle uncertainty, which is crucial for effective information retrieval.

Pros and Cons of Extended Classical IR Models

Set-Theoretic IR Models (e.g., Fuzzy and Extended Boolean Models):

Pros:

1. **Expressive:** These models allow for more flexible representation of documents and queries through the use of sets and logic operations.
2. **Precision Control:** Fuzzy models can handle partial matching, allowing for a degree of relevance in retrieval, which classical models lack.
3. **Interconnectedness:** Extended Boolean models can capture relationships between terms, enhancing retrieval precision.

Pros and Cons of Extended Classical IR Models

Cons:

1. **Complexity:** Fuzzy and extended Boolean models can be more complex to implement and understand for users compared to classical models.

2. **Scalability:** Managing the workings of these models for large document collections can be challenging.

3. **User Learning Curve:** Users may need to learn the degrees of these models, which can be a barrier to adoption.

Pros and Cons of Extended Classical IR Models

Algebraic IR Models (e.g., Neural Networks):

Pros:

1. **Learning:** Neural networks can adapt and learn from data, potentially improving retrieval relevance over time.
2. **Semantic Understanding:** They can capture semantic relationships between terms, enhancing retrieval accuracy.
3. **Applicability:** They are versatile and applicable to various types of information, including text, images, and multimedia.

Pros and Cons of Extended Classical IR Models

Cons:

1. **Complexity:** Building and fine-tuning probabilistic models can be complex and require expertise in probability and statistics.
2. **Data Dependency:** They rely on a substantial amount of data for accurate estimation, which may not be available in some scenarios.
3. **Initial Learning Curve:** Users may require time to understand the probabilistic model's concepts and parameters.

- These are high-level pros and cons, and the specific advantages and disadvantages can vary based on the context and implementation of each model.

Pros of Extended Classical IR Models (Probabilistic Models)

- When discussing Extended Classical Information Retrieval (IR) Models, particularly focusing on probabilistic models such as Interference Models and Belief Networks, there are several pros and cons to consider.

1. **Handling Uncertainty:**

   Probabilistic models effectively manage uncertainty and variability in data, allowing for more interpretations of relevance.

2. **Incorporating Prior Knowledge:**

   These models can incorporate prior information about documents and queries, improving retrieval performance when data is sparse.

Pros of Extended Classical IR Models (Probabilistic Models)

3. Flexible Representations:

Models like belief networks provide flexible frameworks for representing complex dependencies among variables, enabling richer representations of information.

4. Improved Ranking:

Probabilistic models can yield better document ranking by estimating the probability of relevance based on various factors, leading to more relevant search results.

5. Bayesian Framework:

The Bayesian approach allows for continuous learning and updating of beliefs based on new evidence, enhancing adaptability.

Cons of Extended Classical IR Models (Probabilistic Models)

1. **Complexity:**

   The implementation of probabilistic models, especially belief networks, can be complex and computationally intensive, making them harder to deploy in real-time systems.

2. **Data Requirements:**

   These models often require substantial amounts of training data to estimate probabilities accurately, which may not always be available.

3. **Assumptions of Independence:**

   Some probabilistic models rely on strong independence assumptions, which may not hold in practice, leading to suboptimal performance.

Cons of Extended Classical IR Models (Probabilistic Models)

4. Interpretability:

The probabilistic nature of these models can make them less interpretable compared to simpler models, making it challenging to understand why certain documents are retrieved.

5. Parameter Sensitivity:

The effectiveness of probabilistic models can be sensitive to the choice of parameters, and tuning these can be difficult without extensive experimentation.

Cons of Extended Classical IR Models (Probabilistic Models)

Conclusion

- Extended Classical IR Models, specifically probabilistic models like Interference Models and Belief Networks, offer significant advantages in managing uncertainty and improving retrieval performance. However, their complexity, data requirements, and sensitivity to assumptions pose challenges that need to be addressed in practical applications. Balancing these pros and cons is crucial for developing effective IR systems.

Mid-term

- All the contents delivered in the class will be **the syllabus of the Mid-term exam**.
- **Mid-term** carries **30 Marks** and its weightage is **10 percent.**

REFERENCES:

[1] “INFORMATION RETRIEVAL MODELS,” *Ebrary*, 2013. https://ebrary.net/201793/psychology/retrieval_models (accessed Feb. 22, 2023).

[2] “Ch_2 Information Retrieval Models,” *Rutgers.edu*, 2023. https://aspoerri.comminfo.rutgers.edu/InfoCrystal/Ch_2.html (accessed Feb. 22, 2023).

[3] “Fig. 2. Horizontal taxonomy.,” *ResearchGate*, 2021. https://www.researchgate.net/figure/Horizontal-taxonomy_fig2_47397195 (accessed Feb. 22, 2023).

[4] “4.1. IR MODELS – BASIC CONCEPTS – Wachemo University e-Learning Platform,” *Wachemo-elearning.net*, 2023.https://wachemo/elearning.net/courses/information-storage-and-retirievalitec3081/lessons/chapter-four-ir-model/topic/4-1-ir-models-basic-concepts/ (accessed Feb. 22, 2023).

Any
Questions
?