Chapter IR:III

III. Retrieval Models

- ❑ Overview of Retrieval Models
- ❑ Boolean Retrieval
- ❑ Vector Space Model
- ❑ Binary Independence Model
- ❑ Okapi BM25
- ❑ Divergence From Randomness
- ❑ Latent Semantic Indexing
- ❑ Explicit Semantic Analysis
- ❑ Language Models

- ❑ Combining Evidence
- ❑ Learning to Rank

Overview of Retrieval Models

Document Views

Information retrieval requires modeling and representing documents on a computer. We distinguish three orthogonal views on a document's content:

1. Layout view

   Presentation of a document on a (two-dimensional) medium.

2. Structural / logical view

   Composition and logical structure of a document. Example:

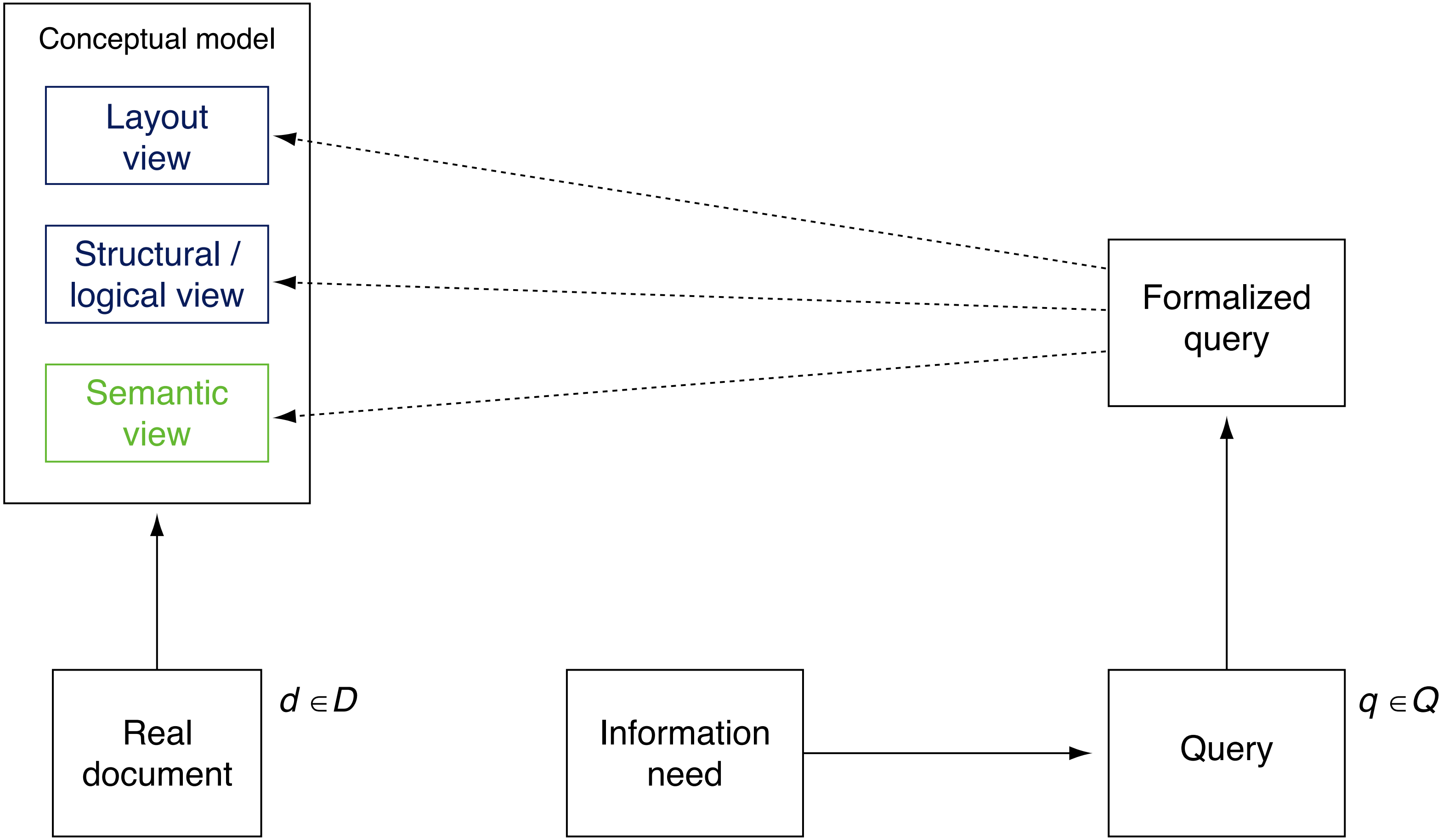
   ```
   \documentclass[twocolumn,english]{article}
   \title{...}
   \author{...}
   \section{...}
   ```

3. Semantic view

   The meaning of a document or its message, allowing for its interpretation.

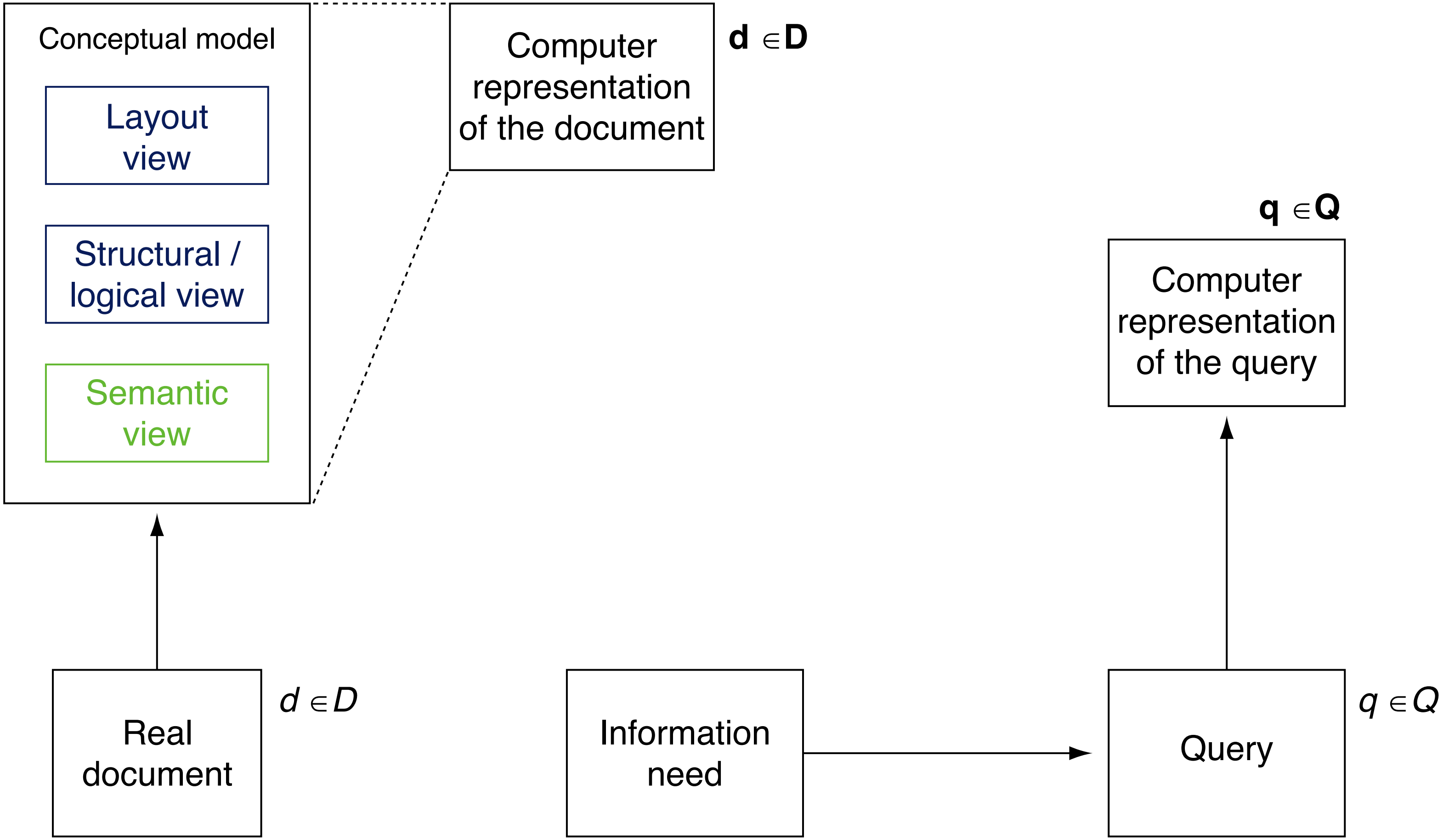
Overview of Retrieval Models

Retrieval Models



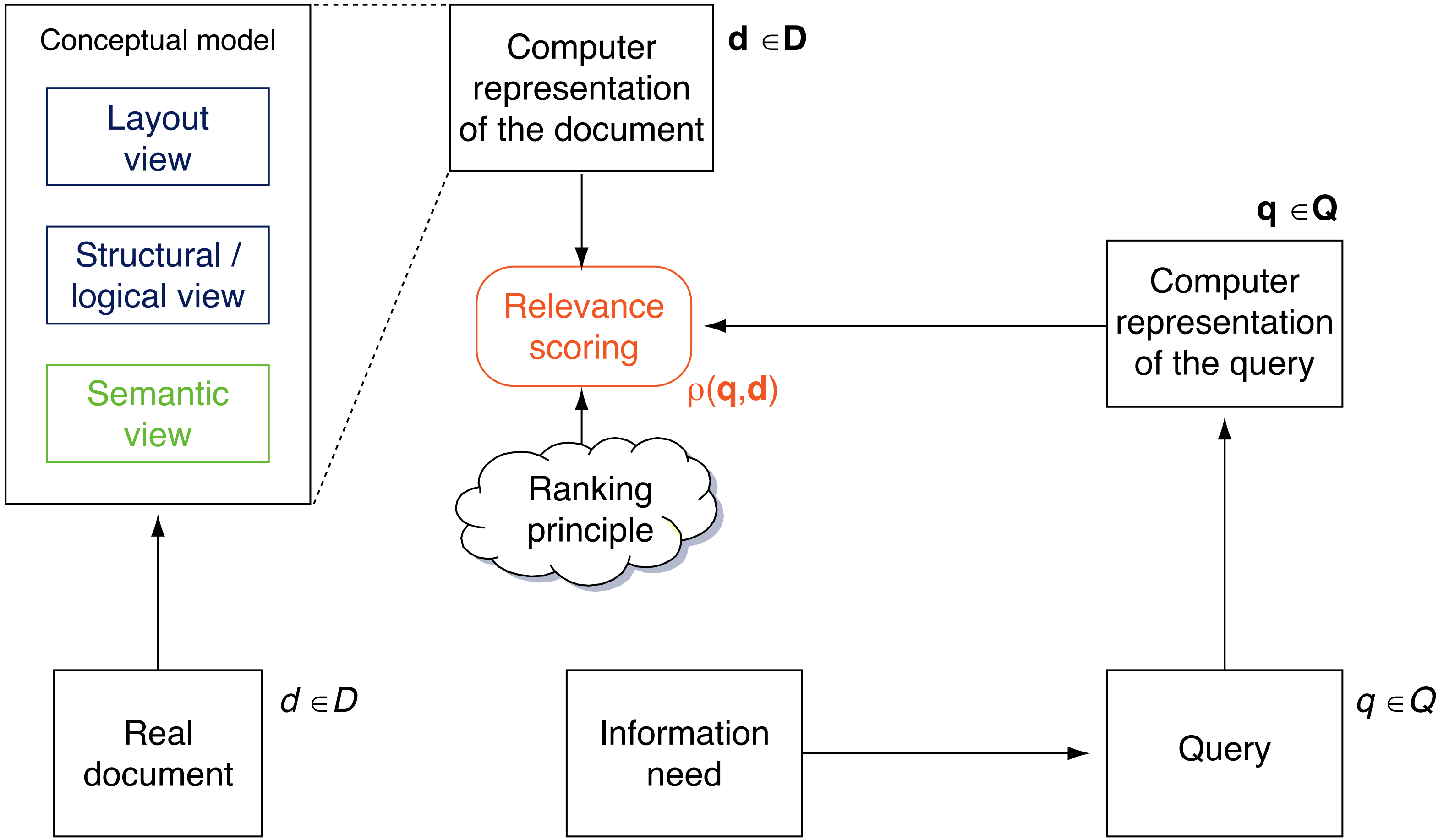
Overview of Retrieval Models

Retrieval Models


Conceptual model
Layout view
Structural / logical view
Semantic view
Computer representation of the document
$\mathbf{d} \in \mathbf{D}$
$\mathbf{q} \in \mathbf{Q}$
Computer representation of the query
Real document
$d \in D$
Information need
Query
$q \in Q$

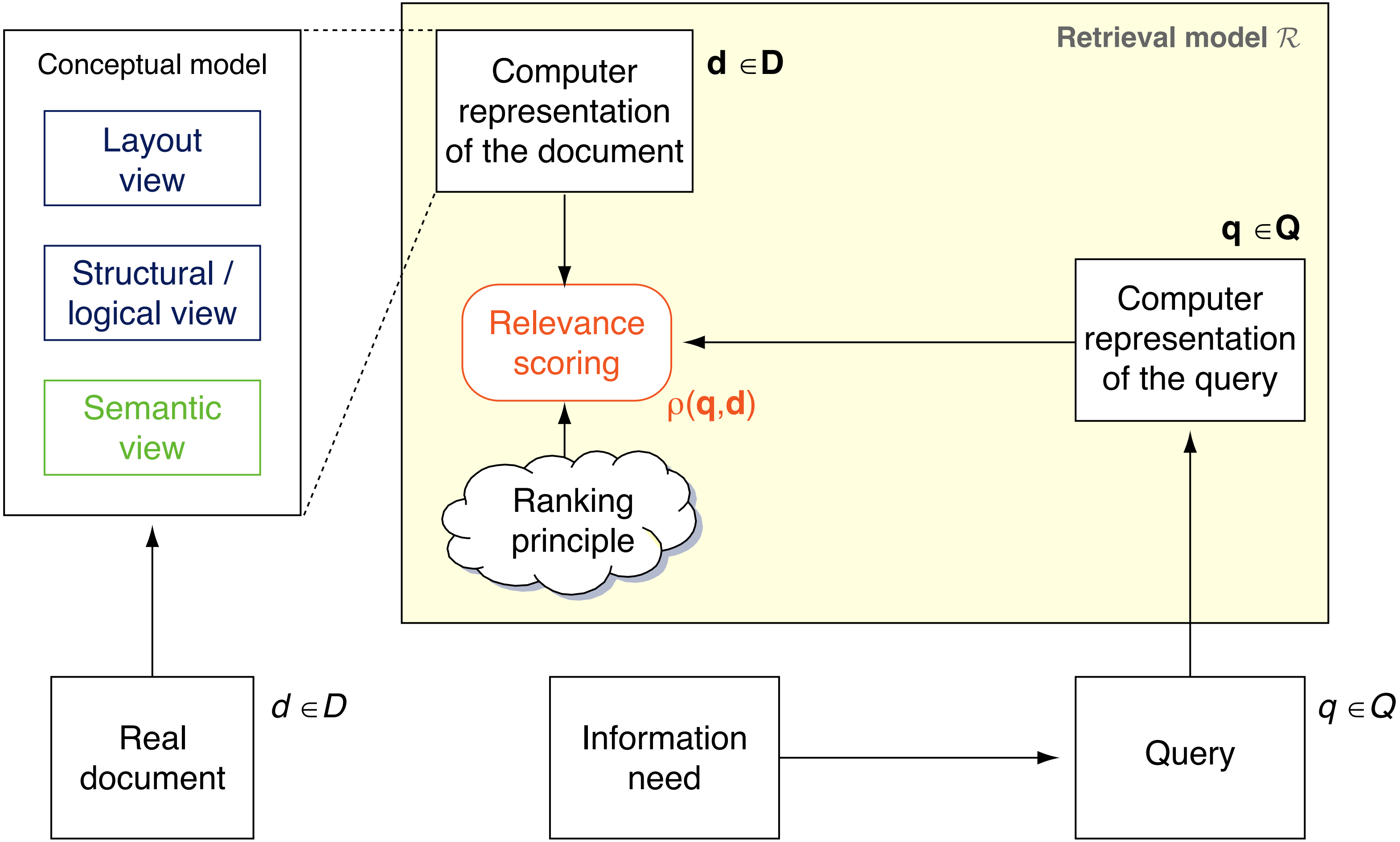
Overview of Retrieval Models

Retrieval Models


Conceptual model
Layout view
Structural / logical view
Semantic view
Computer representation of the document
$\mathbf{d} \in \mathbf{D}$
$\mathbf{q} \in \mathbf{Q}$
Relevance scoring
$\rho(\mathbf{q},\mathbf{d})$
Computer representation of the query
Ranking principle
Real document
$d \in D$
Information need
Query
$q \in Q$

Overview of Retrieval Models

Retrieval Models



Conceptual model
Layout view
Structural / logical view
Semantic view
Retrieval model $\mathcal{R}$
Computer representation of the document
d ∈**D**
q ∈**Q**
Relevance scoring
Computer representation of the query
$\rho(\mathbf{q},\mathbf{d})$
Ranking principle
Real document
d ∈*D*
Information need
Query
q ∈*Q*

Overview of Retrieval Models

Definition 1 (Retrieval Model, Relevance Function)

Let D denote the set of documents and Q the set of queries. A retrieval model $\mathcal{R}$ for D, Q is a tuple $\langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ defined as follows:

1. $\mathbf{D}$ is the set of document representations, where $\mathbf{d} \in \mathbf{D}$ represents $d \in D$.
   It may encode information from the layout view, the logical view, and the semantic view.

2. $\mathbf{Q}$ is the set of query representations.

Overview of Retrieval Models

Definition 1 (Retrieval Model, Relevance Function)

Let D denote the set of documents and Q the set of queries. A retrieval model $\mathcal{R}$ for D, Q is a tuple $\langle \mathbf{D}, \mathbf{Q}, \rho \rangle$ defined as follows:

1. $\mathbf{D}$ is the set of document representations, where $\mathbf{d} \in \mathbf{D}$ represents $d \in D$.
   It may encode information from the layout view, the logical view, and the semantic view.

2. $\mathbf{Q}$ is the set of query representations.

3. $\rho(\mathbf{q}, \mathbf{d})$ denotes a relevance function, which quantifies the relevance between a query q and a document d via their representations $\mathbf{q} \in \mathbf{Q}$ and $\mathbf{d} \in \mathbf{D}$:

   $$\rho : \mathbf{Q} \times \mathbf{D} \to \mathbf{R}$$

   The values computed by ρ are called relevance scores.

$\mathcal{R}$ formalizes a certain ranking principle.

Remarks:

- A document representation encompasses certain elements and specific aspects of a real document. Examples for document representations include feature vectors and fingerprints.
- A retrieval model provides the theoretical foundations of how human information needs can be satisfied by drawing information from the three views. Examples for retrieval models include the vector space model, the binary independence model, and latent semantic indexing.
- An alternative name for a retrieval model is retrieval strategy.
- Most retrieval models are based on the semantic view of documents.
- An intensional definition of the sets $\mathbf{Q}$ and $\mathbf{D}$ can be given as functions $\alpha_Q : Q \to \mathbf{Q}$ and $\alpha_D : D \to \mathbf{D}$. [Fuhr 2004]

Overview of Retrieval Models

Probability Ranking Principle (PRP) [Robertson 1977, 2009]

If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best that can be obtained for the data.

Overview of Retrieval Models

Probability Ranking Principle (PRP) [Robertson 1977, 2009]

> If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best that can be obtained for the data.

Assumptions:

- Relevance *rel*(d, q) is a property of a document d given an information need's query q, assessable without reference to other documents.
- Relevance is binary: *rel*$(d, q) \in \{0, 1\}$.
- Relevance is that of an individual user submitting q.

Ranking by probability of relevance provably maximizes several objective functions:

- Expected recall
- Expected precision
- Expected utility

Remarks:

- The probability ranking principle has not been shown to hold in general as of yet, but under the above assumptions.

- In a counterexample, William S. Cooper considers different users with different information needs who formulate the same query.

- Including more knowledge about the user accounts takes user relevance into account: $P(\textit{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q}, \mathbf{u})$, where $\mathbf{u}$ is a user model.

Overview of Retrieval Models

Ranking Principles in IR

A ranking principle states a criterion and shows that ranking documents by this criterion achieves an objective, usually the maximization of an objective function.

- Binary user relevance *rel*$(d, q) \in \{0, 1\}$ of a document d to a query q.
  Objective: Return relevant documents.

- Semantic similarity $\varphi(\mathbf{d}, \mathbf{q})$ of a document d to a query q.
  Objective: Return documents that are similar to the topic of the query.

- Probability of user relevance $P($*rel*$(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$ of d to q.
  Objective: Return documents that satisfy the user's information need.

- Probability $P(\mathbf{d} \mid \mathbf{q})$ of d having been generated for the topic of query q.
  Objective: Return documents that fit the topic of the query.

- Amount of information $-\log_2 P(\mathbf{d} \mid \mathit{tf}(t_1), \ldots, \mathit{tf}(t_{|\mathbf{q}|}))$ carried by $\mathbf{q}$'s terms in $\mathbf{d}$.
  Objective: Return documents that carry much information about $\mathbf{q}$'s terms.

Overview of Retrieval Models

Types of Retrieval Models [Amati 2018]

Logical models: Evaluation for a given document d and a query q of the truth of

$$\mathcal{I}(\mathbf{d} \rightarrow \mathbf{q})$$

Algebraic models: Computation of the similarity between d and q in a vector space:

$$\varphi(\mathbf{d}, \mathbf{q})$$

Probabilistic models: Estimation of the probability of a user's relevance *rel* of d for q:

$$P(\textit{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$$

Bayesian models: Estimation of the probability of generating d from q:

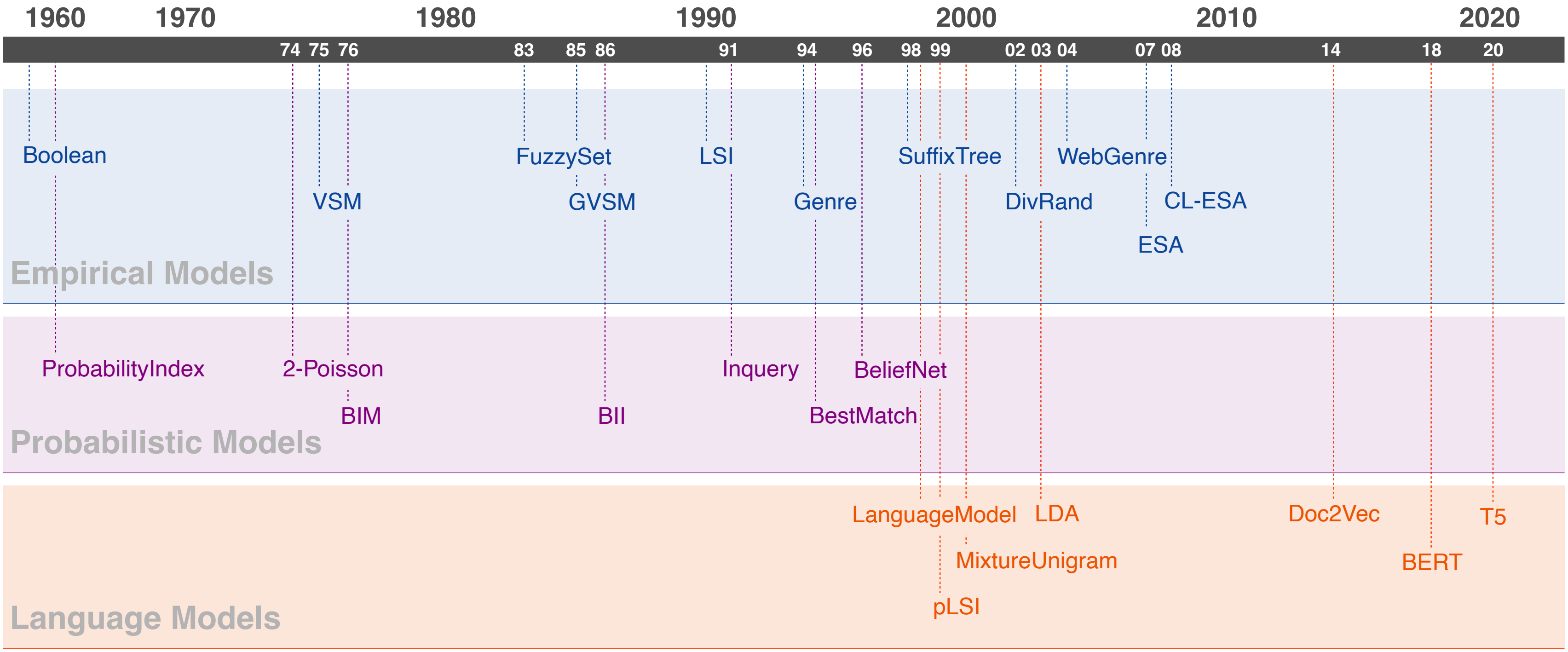
$$P(\mathbf{d} \mid \mathbf{q})$$

Information theoretic models: Computation of the number of bits necessary to code *tf*(t_i) many of $\mathbf{q}$'s i-th term t_i in $\mathbf{d}$ for $i \in \{1, \ldots, |\mathbf{q}|\}$:

$$-\log_2 P(\mathbf{d} \mid \textit{tf}(t_1), \ldots, \textit{tf}(t_{|\mathbf{q}|}))$$

Overview of Retrieval Models

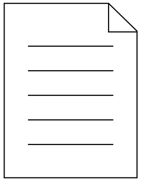
History of Retrieval Models [Stein 2013]



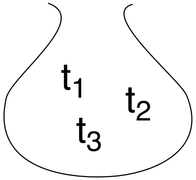
Overview of Retrieval Models

Document Modeling

Analytic models



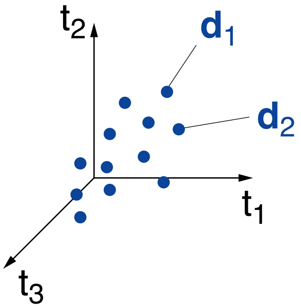
d



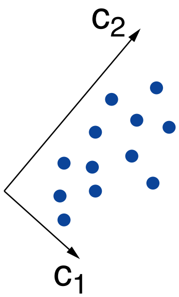
“bag of words” model

	$\mathbf{d}_1$	$\mathbf{d}_2$	⋯
t_1	5	3	
t_2	1	2	⋮
t_3	7	6	

term-document matrix

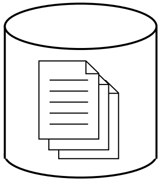


term space

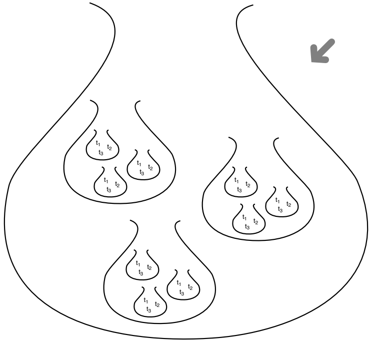


concept space

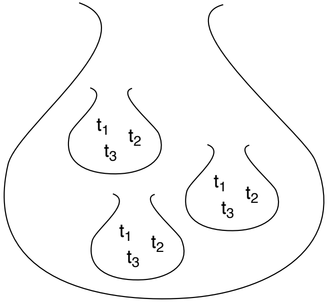
embedding



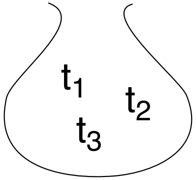
Synthetic models /
Generative models



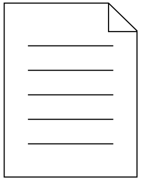
distribution of topic distributions



topic distribution



urn model /
topic model



d