Chapter IR:V

V. Retrieval Models

- Overview of Retrieval Models
- Empirical Models
- Boolean Retrieval
- Vector Space Model
- Probabilistic Models
- Binary Independence Model
- Okapi BM25
- Hidden Variable Models
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Generative Models
- Language Models
- Combining Evidence
- Web Search
- 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.

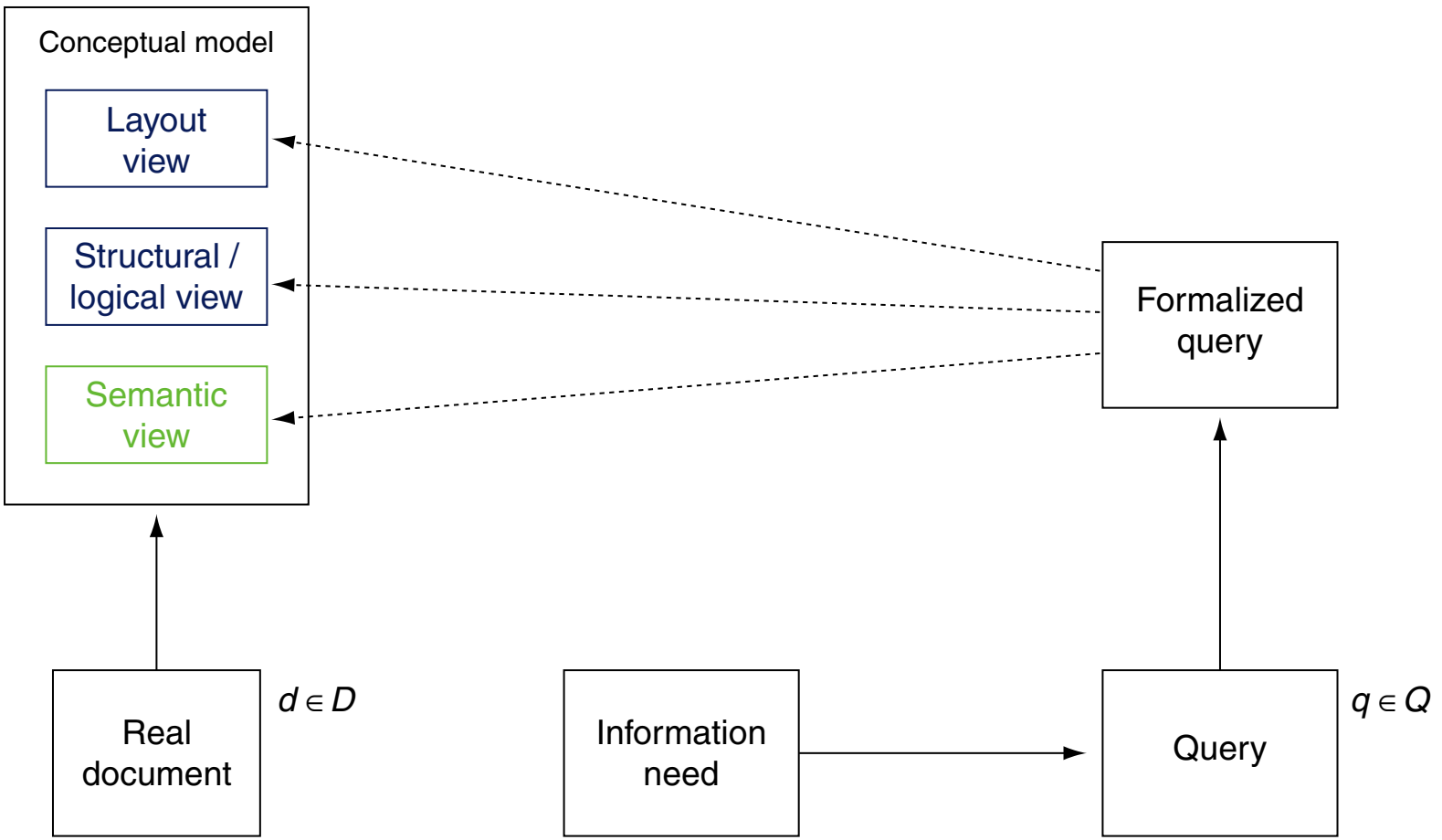
2. Structural / logical view
   Composition and logical structure of a document. Example:

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

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

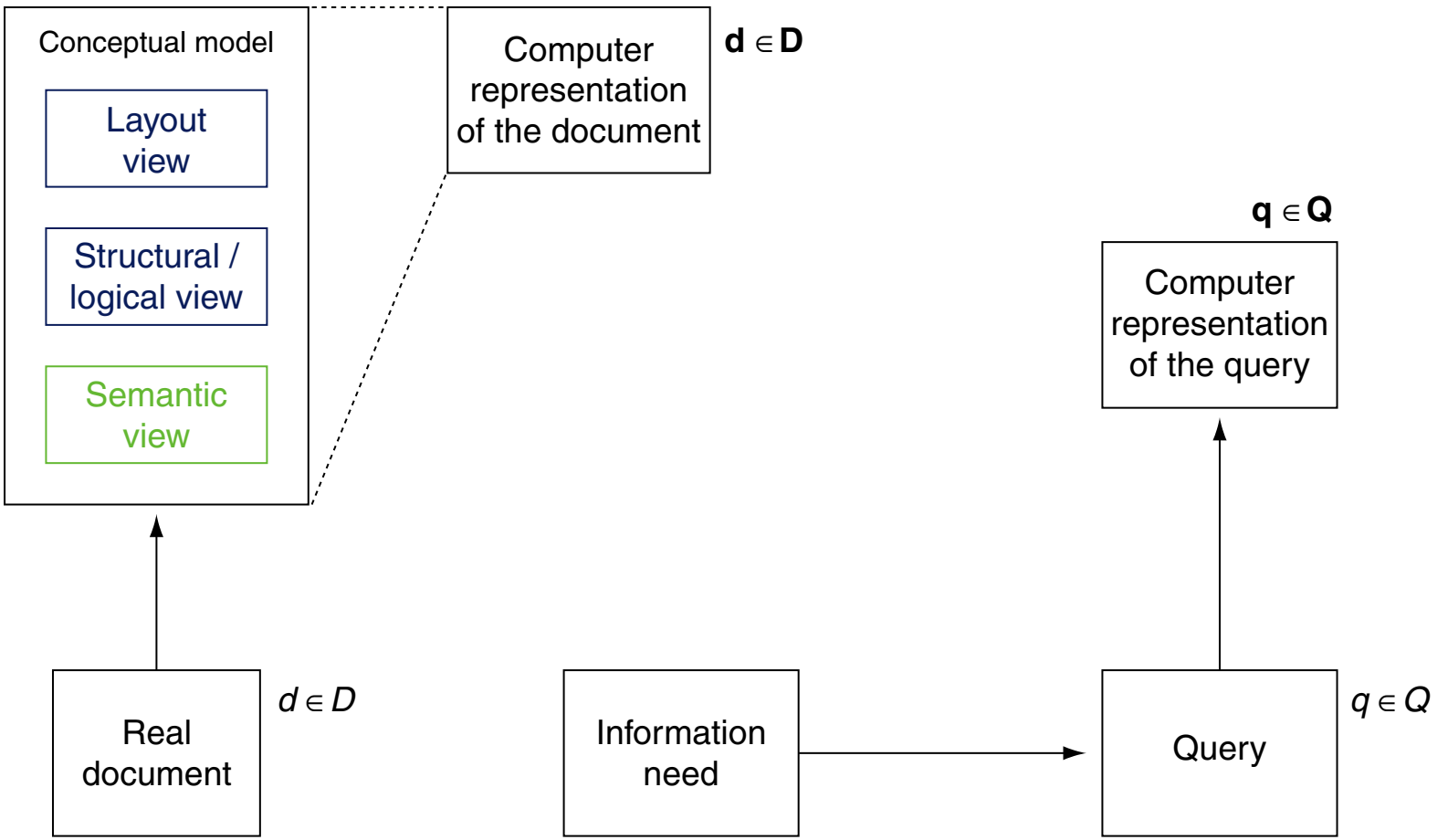
Overview of Retrieval Models

Retrieval Models



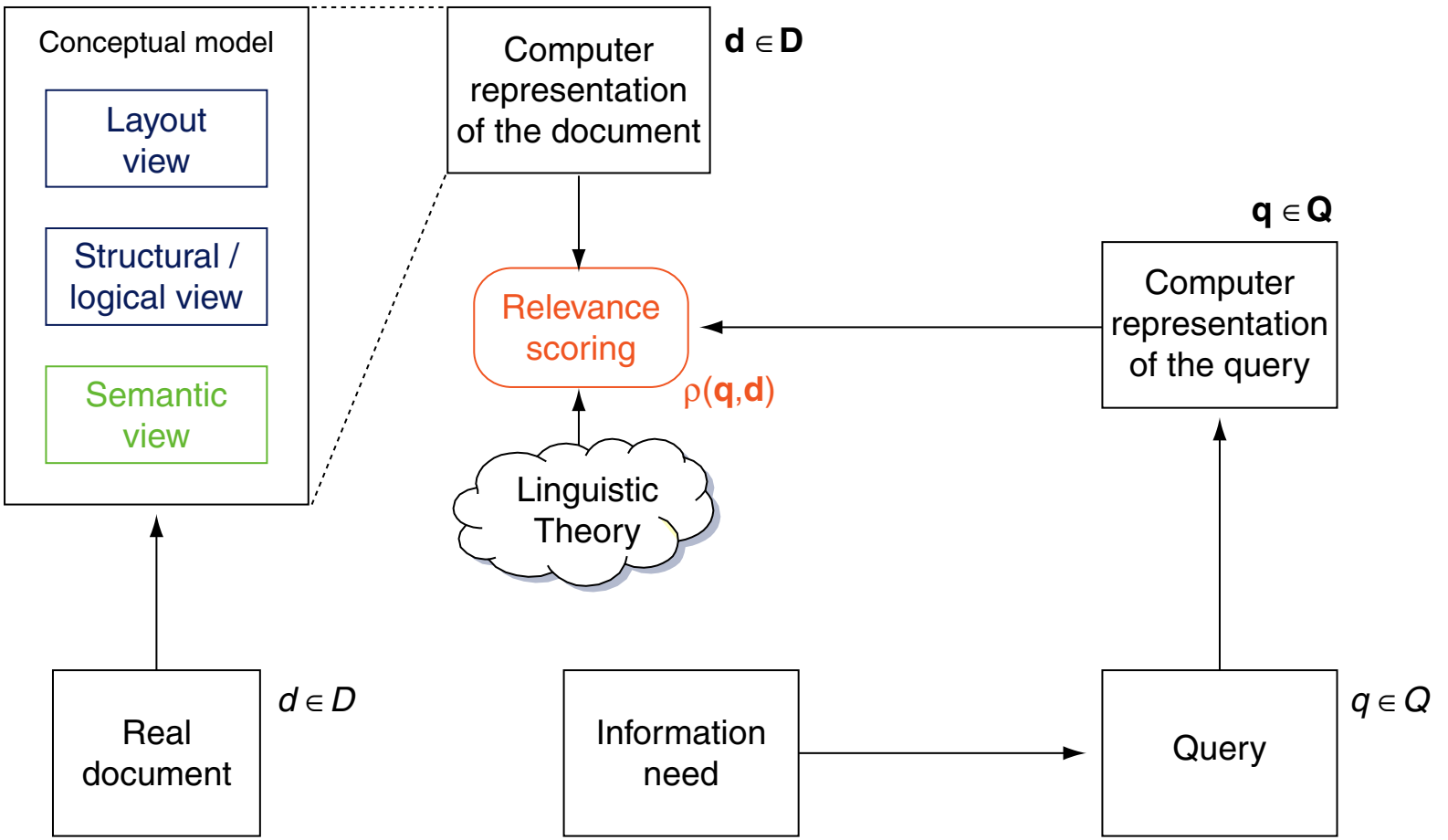
Overview of Retrieval Models

Retrieval Models



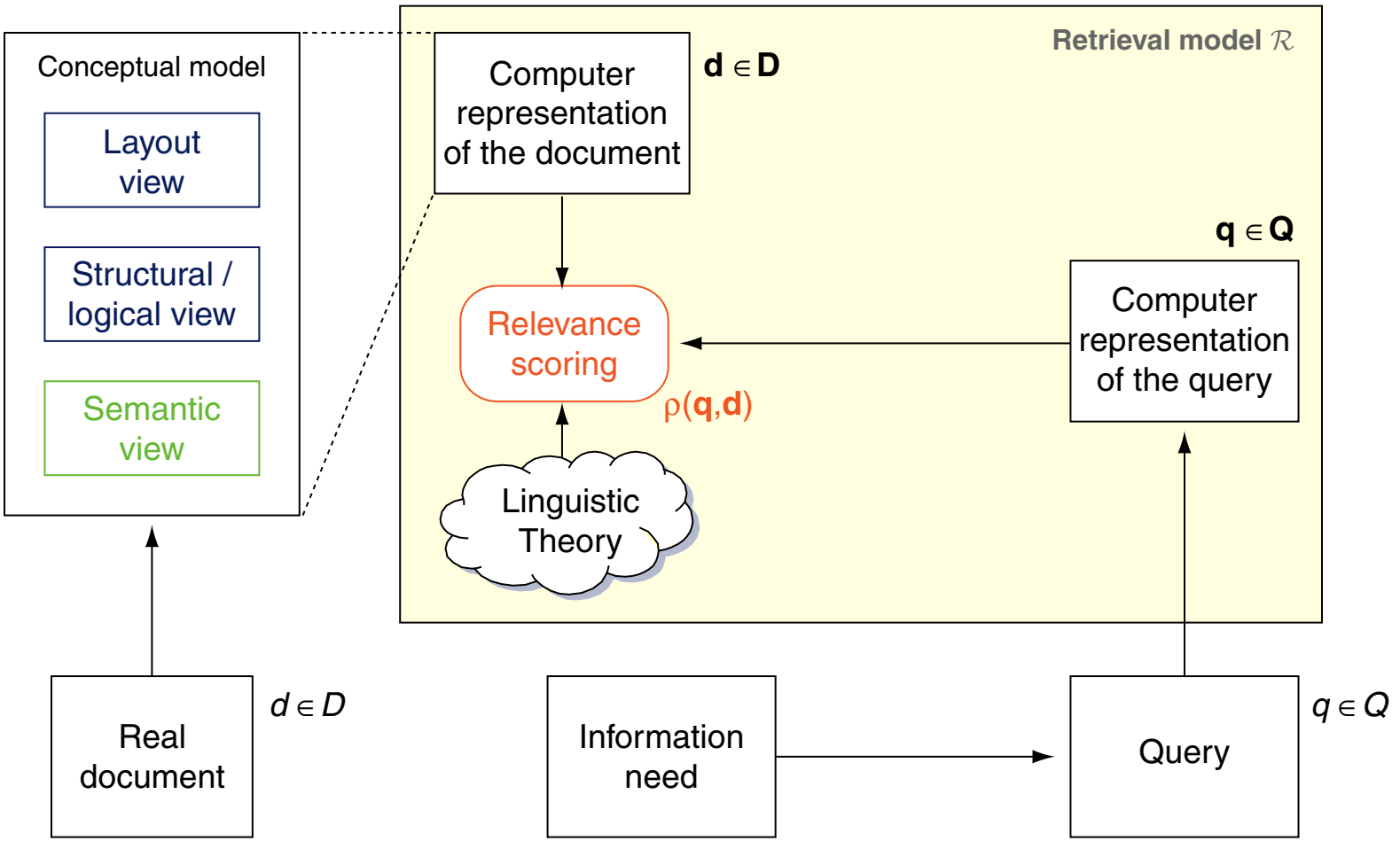
Overview of Retrieval Models

Retrieval Models



Overview of Retrieval Models

Retrieval Models



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 principle, paradigm, or linguistic theory of retrieval.

Remarks:

- A document representation encompasses certain elements and specific aspects of a real document. Examples for document representations include feature vectors, feature trees, 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

History of Retrieval Models [Stein 2013]