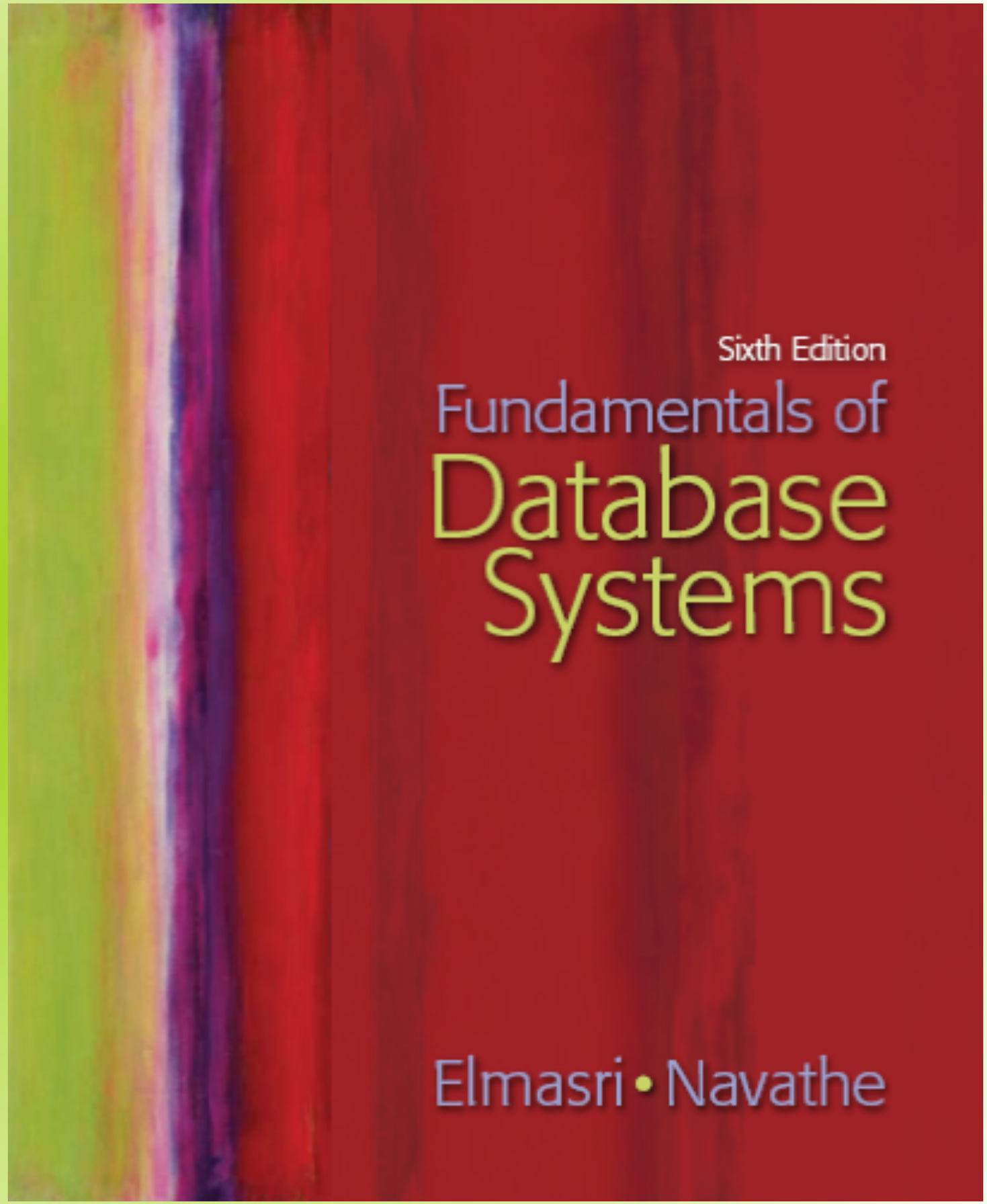
Information Retrieval

Luiz Celso Gomes Jr - André Santanchè
MC536 2013/2

Chapter 27

Introduction to Information Retrieval and Web Search



Addison-Wesley
is an imprint of
PEARSON

Outline

- Information Retrieval (IR) Concepts
- Retrieval Models
- Types of Queries in IR Systems
- Inverted Indexing
- Text Preprocessing
- Evaluation Measures of Search Relevance
- Web Search and Analysis
- Trends in Information Retrieval

Information Retrieval (IR) Concepts

- "Discipline that deals with the structure, analysis, organization, storage, searching, and retrieval of information"
- Process of retrieving documents from a collection in response to a query by a user
- Documents are **unstructured**
- Text, images, sound, etc

Information Retrieval (IR) Concepts (cont'd.)

- User's information need expressed as a free-form search request
  - Keyword search query
  - Also: images, speech, context… becoming more and more important
- High noise-to-signal ratio

Information Retrieval (IR) Concepts (cont'd.)

- IR systems characterized by:
  - Types of users
  - Types of data
  - Types of information needed
  - Levels of scale
- Examples:
  - Web search
  - Enterprise search systems
  - Desktop/mobile search engines
  - Image search
  - Library catalog search

Databases and IR Systems: A Comparison

Table 27.1 A Comparison of Databases and IR Systems

Databases	IR Systems
■ Structured data	■ Unstructured data
■ Schema driven	■ No fixed schema; various data models (e.g., vector space model)
■ Relational (or object, hierarchical, and network) model is predominant	■ Free-form query models
■ Structured query model	■ Rich data operations
■ Rich metadata operations	■ Search request returns list or pointers to documents
■ Query returns data	■ Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked)
■ Results are based on exact matching (always correct)	

Brief History of IR

- Inverted file organization
  - Based on keywords and their weights (SMART system in 1960s)
- Text Retrieval Conference (TREC)
- Search engine
  - Application of information retrieval to large-scale document collections
  - Crawler: Responsible for discovering, analyzing, and indexing new documents
- Google

IR Modern Trends

- Social Search
- Mobile
  - Context-aware
  - Conversational

Modes of Interaction in IR Systems

- Query
  - Set of terms
  - Used by searcher to specify information need

- Main modes of interaction with IR systems:
  - Retrieval: Extraction of information from a repository of documents through an IR query
  - Browsing: User visiting or navigating through similar or related documents

Modes of Interaction in IR Systems (cont'd.)

- Web search
  - Combines browsing and retrieval
- Rank of a Webpage
  - Measure of relevance to query that generated result set

Generic IR Pipeline

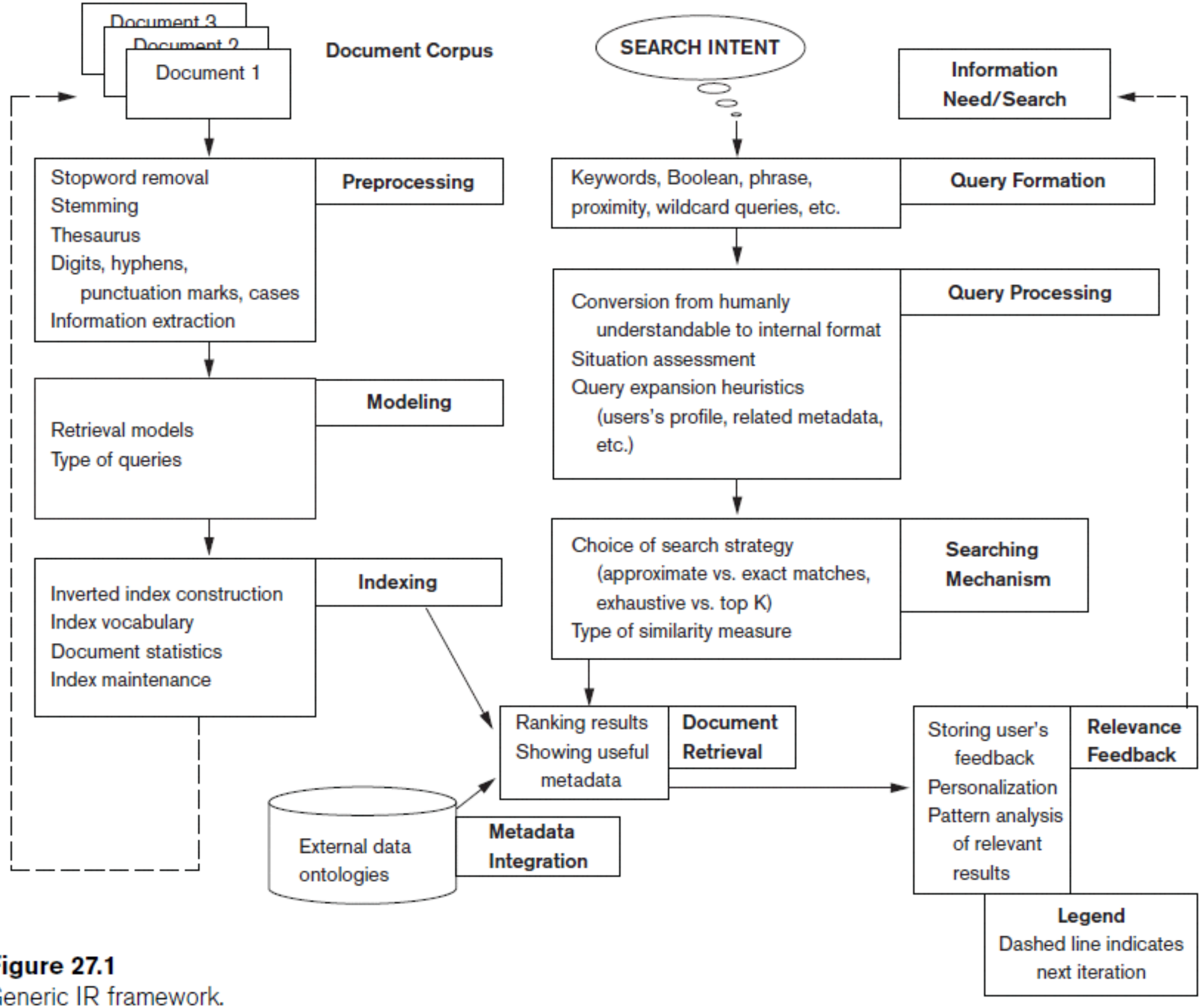


Figure 27.1
Generic IR framework.

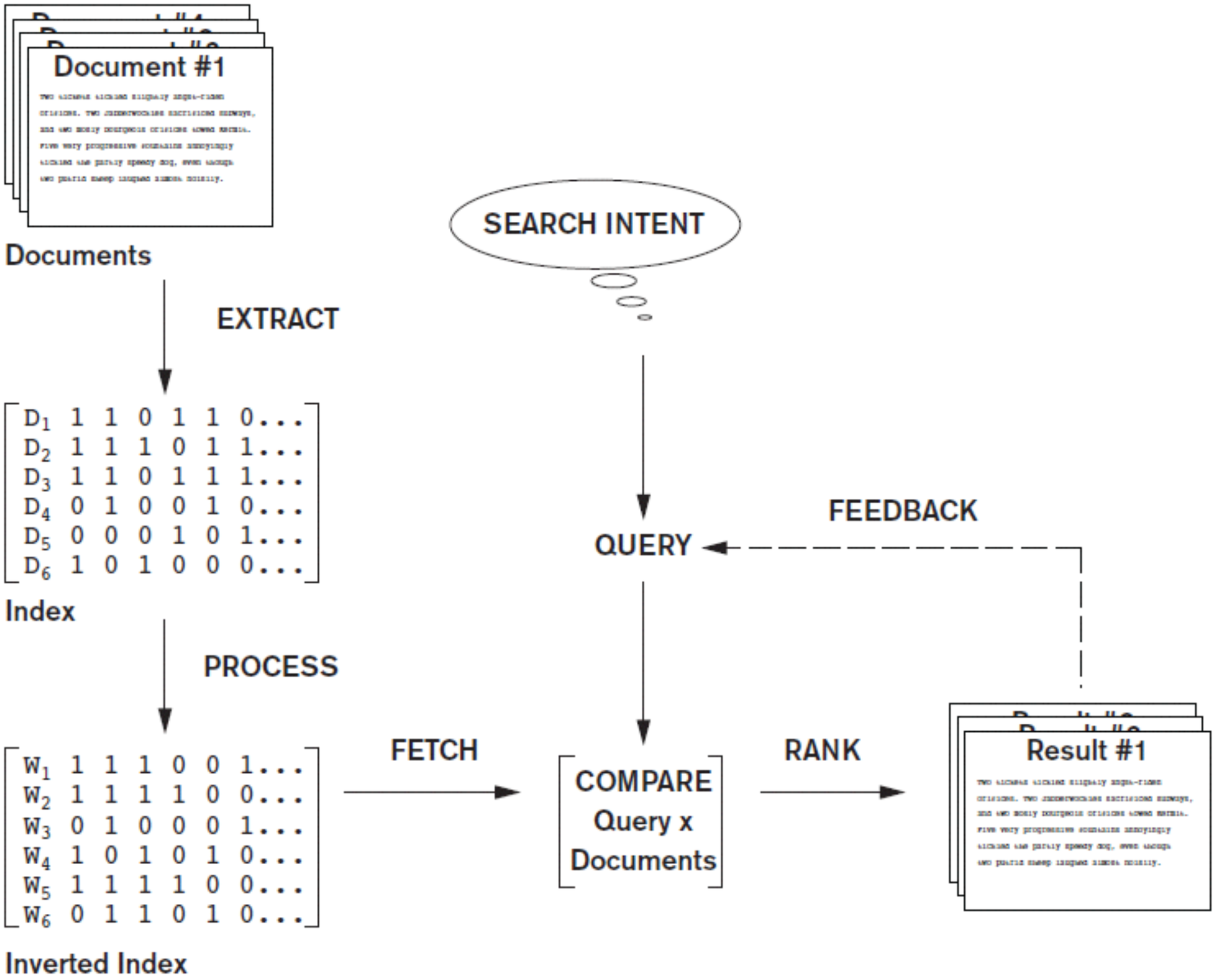
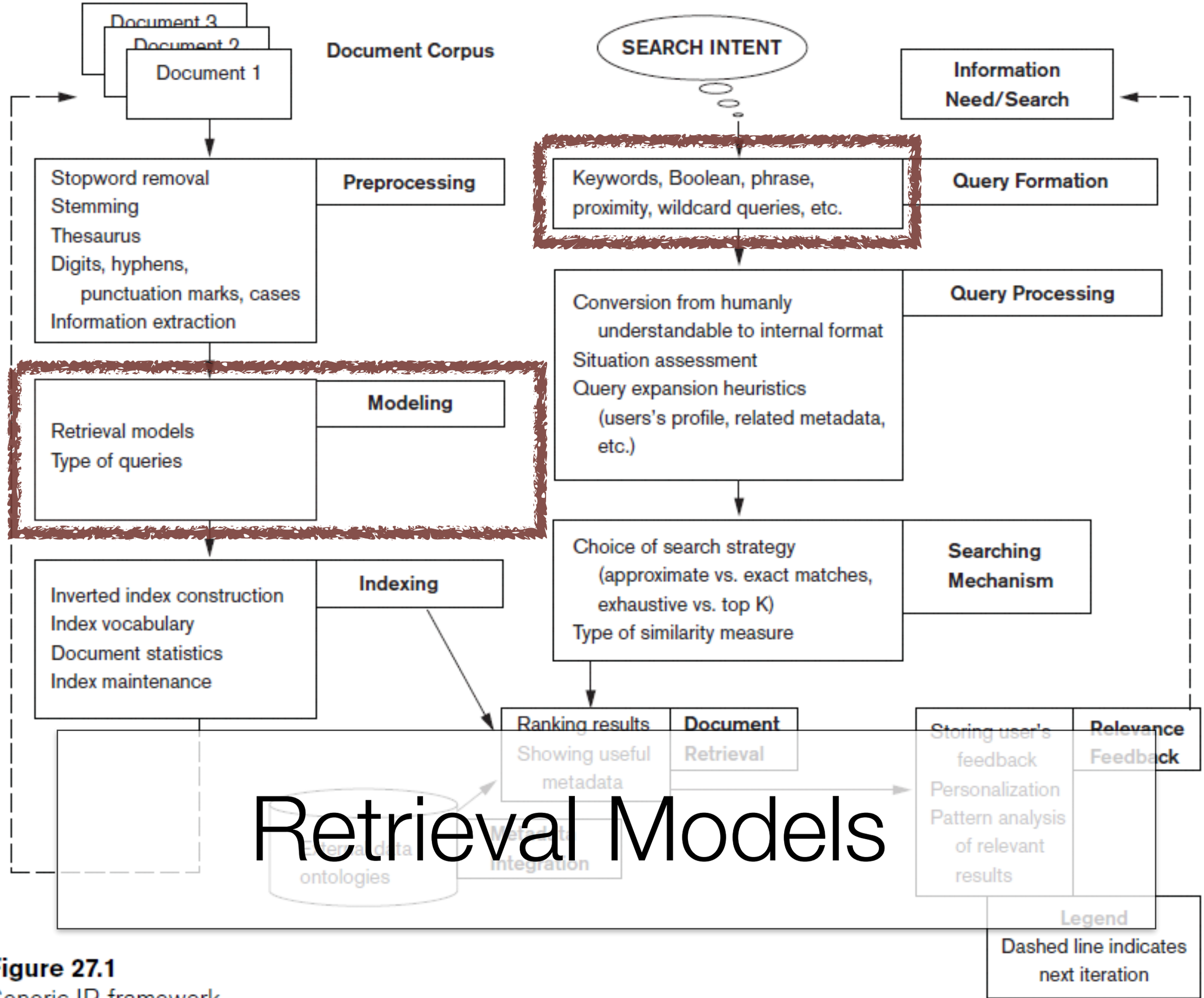


Figure 27.2
Simplified IR process pipeline.



Retrieval Models

Figure 27.1
Generic IR framework.

Retrieval Models

- Three main statistical models
  - Boolean
  - Vector space
  - Probabilistic
- Semantic model

Boolean Model

- Documents represented as a set of terms
- Form queries using standard Boolean logic set-theoretic operators
  - AND, OR and NOT
- Retrieval and relevance
  - Binary concepts
- Lacks sophisticated ranking algorithms

Vector Space Model

- Documents
  - Represented as **terms** (features) and **weights** in an **n-dimensional vector space**
- Query
  - Specified as a terms vector
  - Compared to the document vectors for similarity/relevance assessment

Vector Space Model (cont'd.)

- Different similarity functions can be used
  - Cosine of the angle between the query and document vector commonly used
- TF-IDF
  - Statistical weight measure
  - Used to evaluate the importance of a document word in a collection of documents

TF-IDF

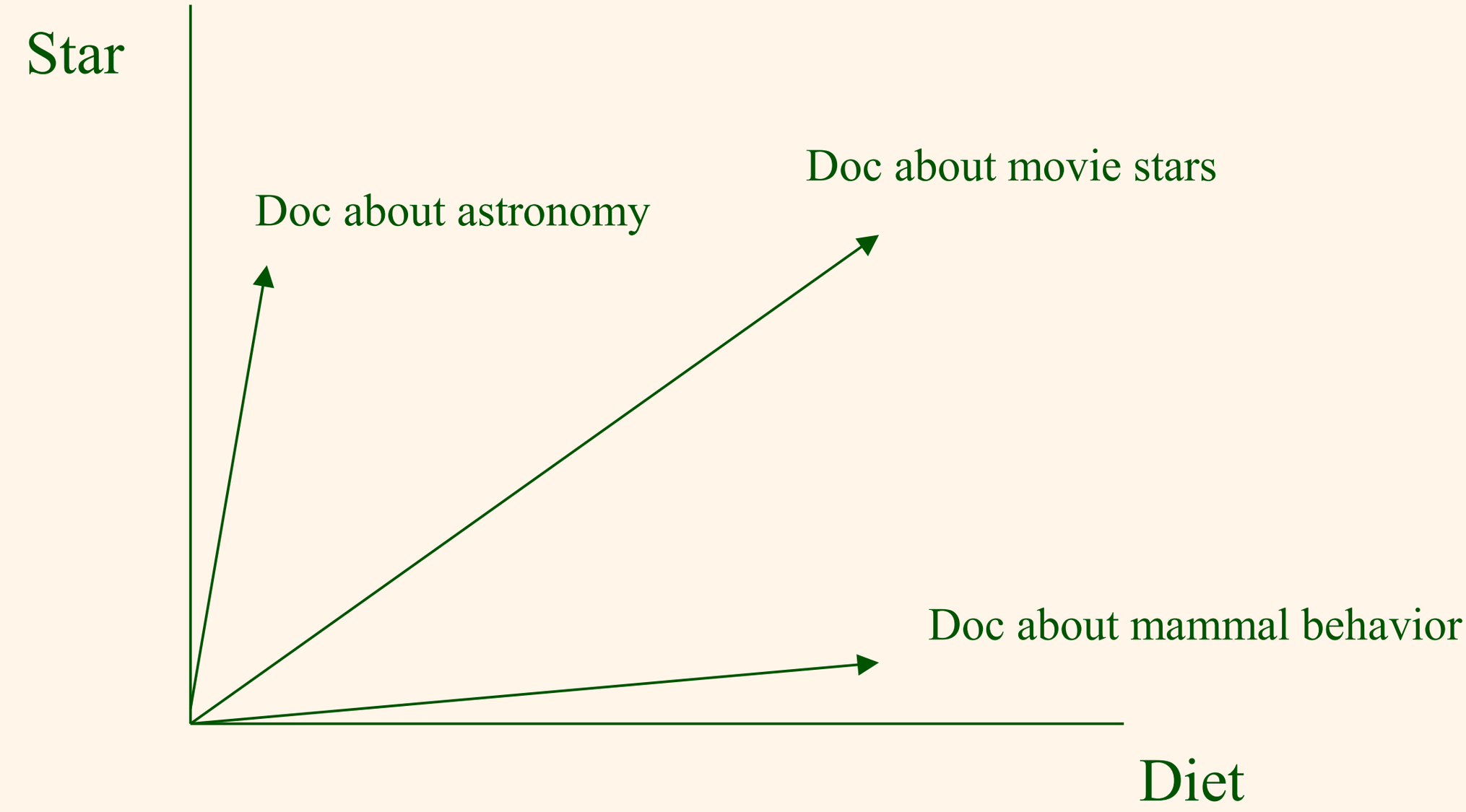
- Term Frequency–Inverse Document Frequency
- Numerical statistic which reflects how important a term (t) is to a document (d) in a collection or corpus (D)
- tf(t,d) = (number of t in d)/(size of d)
- df(t, D) = (number of times t appears in a document in D)/(number of documents in D)
- TF-IDF = tf/df
- Many more sophisticated ranking functions are variants of this simple model

Document Vectors

Document ids

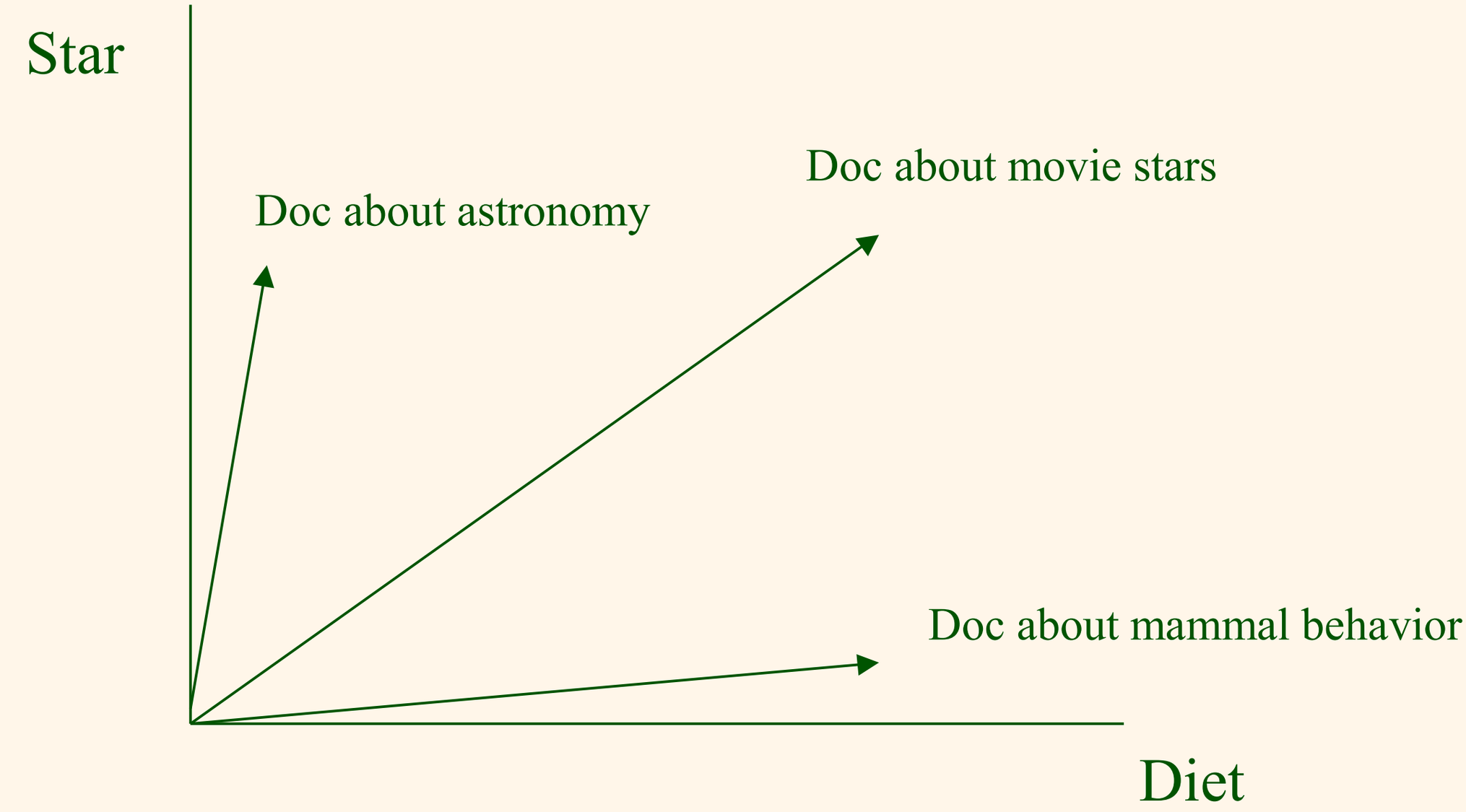
	nova	galaxy	heat	h'wood	film	role	diet	fur
A	10	5	3					
B	5	10						
C				10	8	7		
D				9	10	5		
E							10	10
F							9	10
G	5	7			9			
H		6	10	2	8			
I				7	5		1	3

We Can Plot the Vectors



Assumption: Documents that are "close" in space are similar.

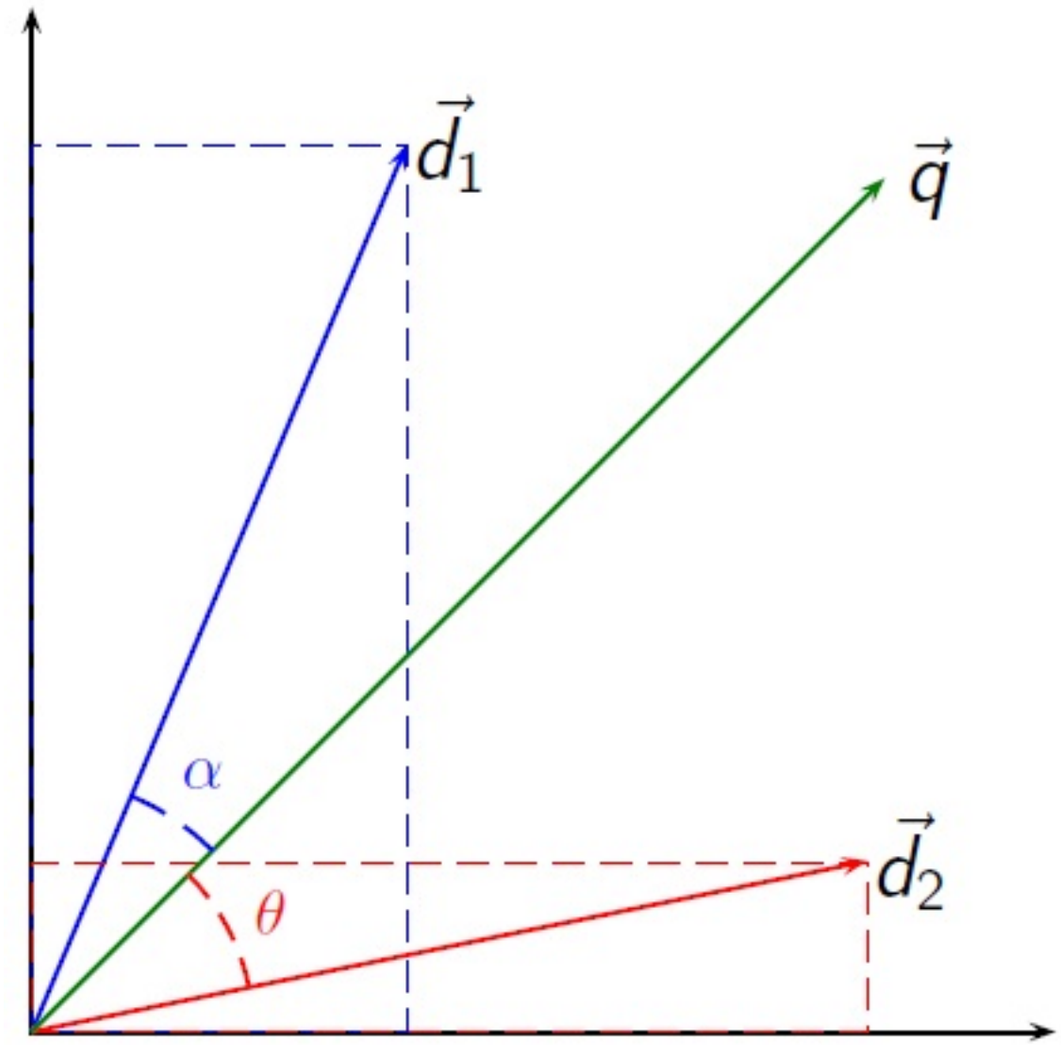
We Can Plot the Vectors



Assumption: Documents that are "close" in space are similar.

Cosine similarity

- + Simple model
- + Works well in average
- + Allows ranking
- - No term order information
- - Term similarity often not related to semantic similarity



$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|} \qquad \|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2}$$

Exercício 1

- Imagine que você precisa desenvolver um sistema de recuperação de fotos de rostos de pessoas. O sistema recebe como entrada uma foto de um rosto e retorna fotos ordenadas por similaridade com a foto da consulta.
- O modelo escolhido para a implementação foi o de Vector Space.
- Que tipo de informação poderia ser usada nos vetores?

Probabilistic Model

- Probability ranking principle
  - Decide whether the document belongs to the relevant set or the nonrelevant set for a query
- Conditional probabilities calculated using Bayes' Rule
- BM25 (Best Match 25)
  - Popular probabilistic ranking algorithm
- Okapi system

Semantic Model

- Knowledge-based IR systems
  - Based on semantic models
  - DBPedia
    - Cyc knowledge base
    - WordNet
- Latent Semantic Indexing/Analysis
- Case: Wikipedia-based cross-language IR

Types of Queries

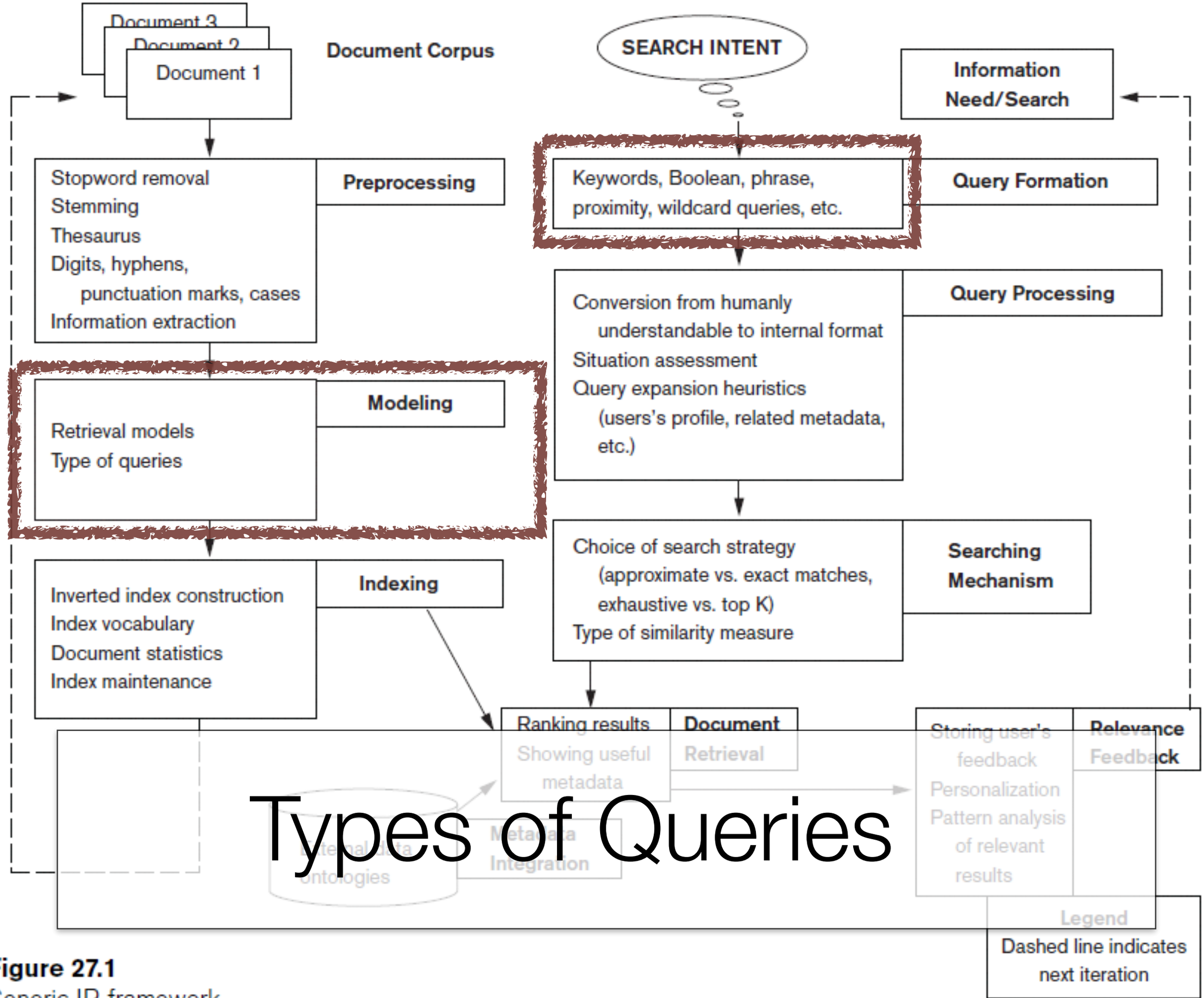


Figure 27.1
Generic IR framework.

Types of Queries in IR Systems

- Keywords
  - Consist of words, phrases, and other characterizations of documents
  - Used by IR system to build inverted index
- Queries compared to set of index keywords
- Most IR systems
  - Allow use of Boolean and other operators to build a complex query

Keyword Queries

- Simplest and most commonly used forms of IR queries
- Keywords implicitly connected by a logical ~~AND~~ AND or OR operator
  may or may not…
- Remove ~~stop~~words - Most commonly occurring words (a, the, of)
- IR systems ~~do not~~ may pay attention to the ordering of these words in the query

Boolean Queries

- AND: both terms must be found
- OR: either term found
- NOT: record containing keyword omitted
- ( ): used for nesting
- +: equivalent to AND
- – Boolean operators: equivalent to AND NOT
- Document retrieved if query logically true as exact match in document

Phrase Queries

- Phrases encoded in inverted index or implemented differently
- Phrase generally enclosed within double quotes
- More restricted and specific version of proximity searching

Proximity Queries

- Accounts for how close within a record multiple terms should be to each other
- Common option requires terms to be in the exact order
- Various operator names
  - NEAR, ADJ(adjacent), or AFTER
- Computationally expensive

Wildcard Queries

- Support regular expressions and pattern matching-based searching
  - ‘Data*’ would retrieve data, database, datapoint, dataset
- Involves preprocessing overhead
- Not considered worth the cost by many Web search engines today
- Retrieval models do not directly provide support for this query type

Natural Language Queries

- Few natural language search engines
- Active area of research

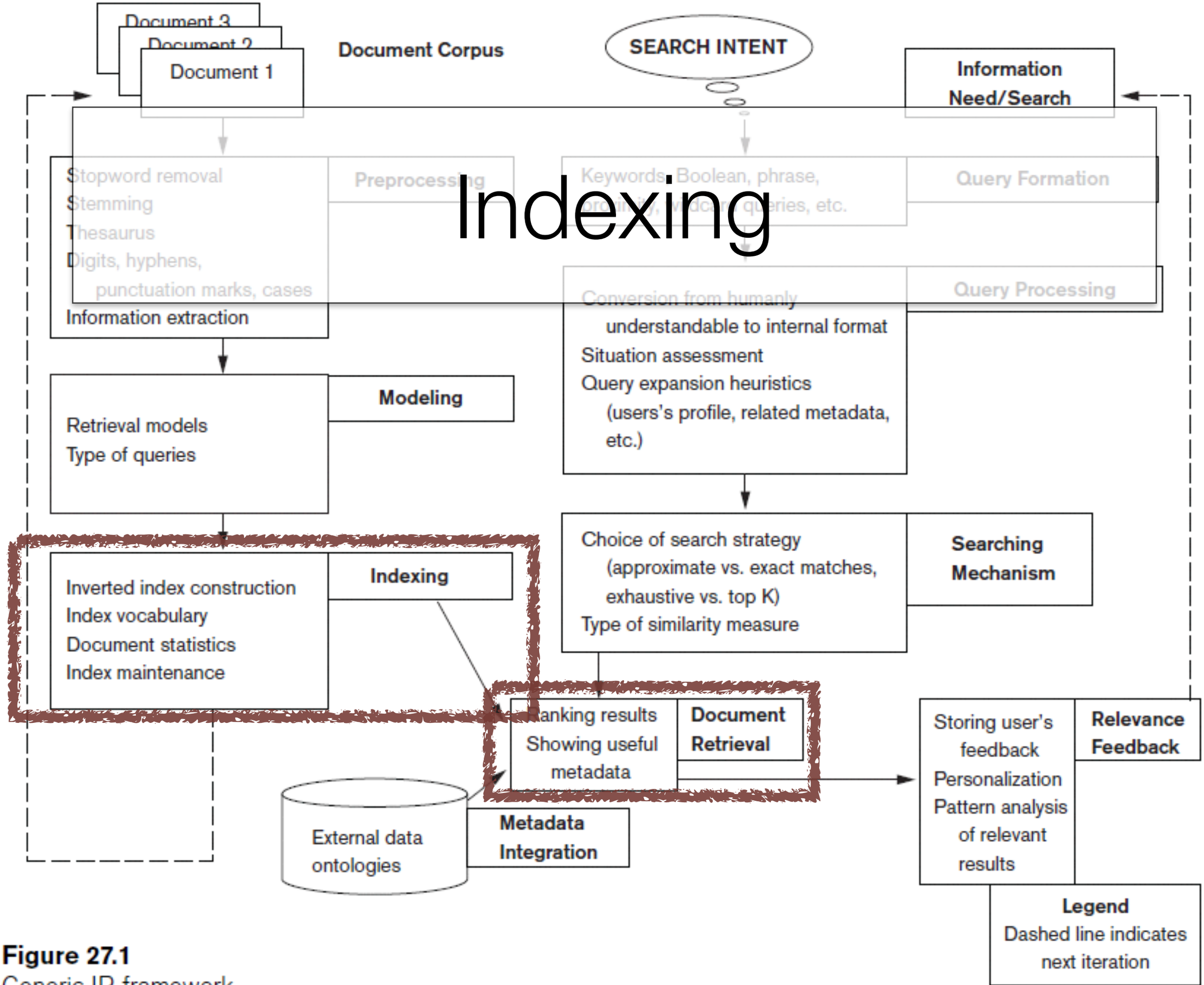


Figure 27.1
Generic IR framework.

Inverted Indexing

- Vocabulary
  - Set of distinct query terms in the document set
- Inverted index
  - Data structure that attaches distinct terms with a list of all documents that contains term

Document 1

This example shows an example of an inverted index.

Document 2

Inverted index is a data structure for associating terms to documents.

Document 2

Stock market index is used for capturing the sentiments of the financial market.

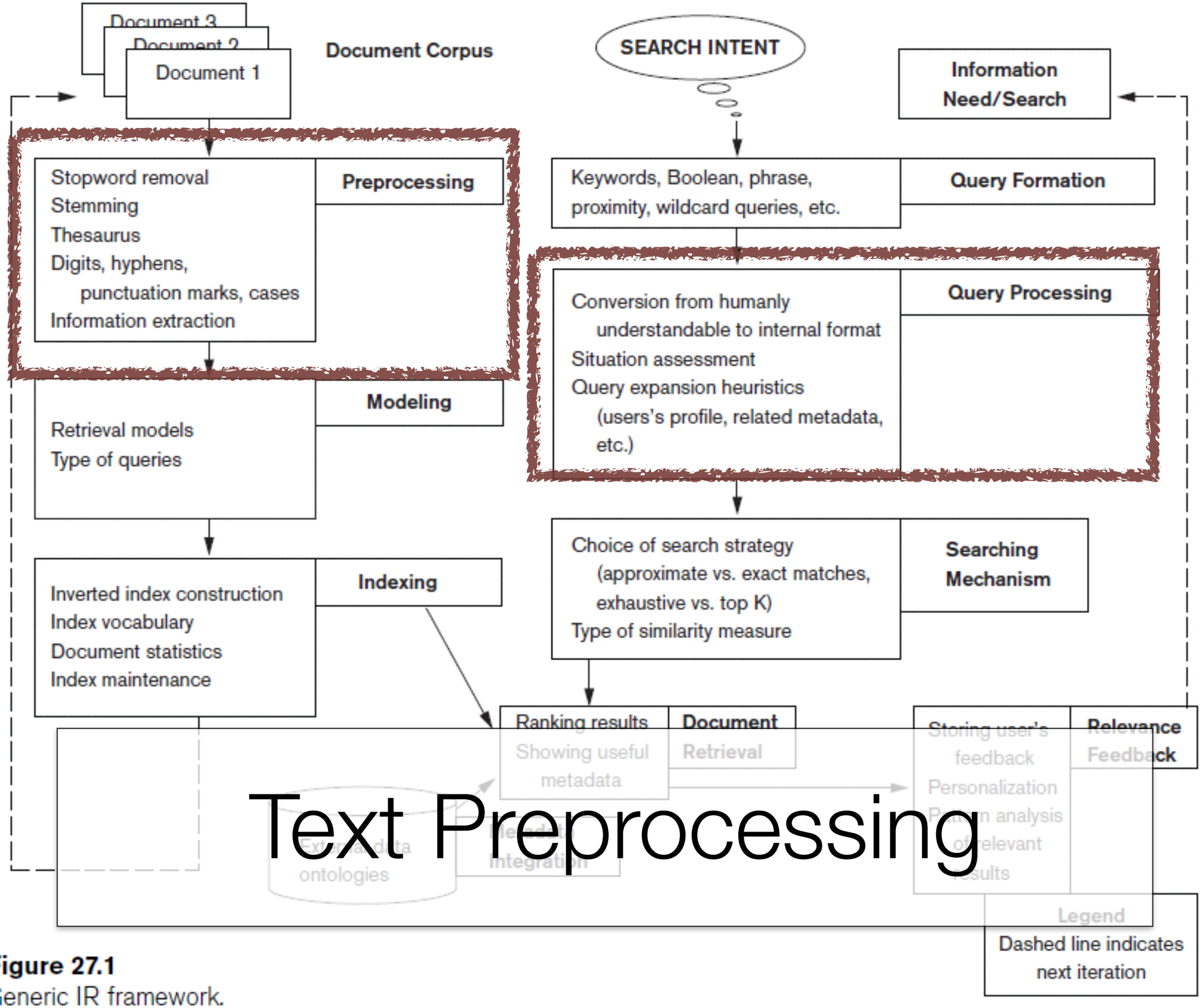
ID	Term	Document: position
1.	example	1:2, 1:5
2.	inverted	1:8, 2:1
3.	index	1:9, 2:2, 3:3
4.	market	3:2, 3:13

Figure 27.4
Example of an inverted index.

Exercício 2

ID	Term	Document: position
1.	example	1:2, 1:5
2.	inverted	1:8, 2:1
3.	index	1:9, 2:2, 3:3
4.	market	3:2, 3:13

Faça o pseudo-código para um algoritmo que, baseado num índice invertido como o do exemplo ao lado, processe consultas booleanas conjuntivas (a AND b AND c AND…).



Text Preprocessing

Figure 27.1
Generic IR framework.

Text Preprocessing

- Commonly used text preprocessing techniques
- Part of text processing task

Stopword Removal

- Stopwords
  - Very commonly used words in a language
  - Expected to occur in 80 percent or more of the documents
    - the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by, and it
- Removal must be performed before indexing
- Queries can be preprocessed for stopword removal

Stemming

- Stem
  - Word obtained after trimming the suffix and prefix of an original word
- Reduces different forms of the word formed by inflection
- Most famous stemming algorithm:
  - Martin Porter’s stemming algorithm

Utilizing a Thesaurus

- Thesaurus: Precompiled list of important concepts and the main word that describes each
  - Synonym converted to its matching concept during preprocessing
- Examples:
- UMLS: Large biomedical thesaurus of concepts/meta concepts/relationships
- WordNet: Manually constructed thesaurus that groups words into strict synonym sets

Other Preprocessing Steps: Digits, Hyphens, Punctuation Marks, Cases

- Digits, dates, phone numbers, e-mail addresses, and URLs may or may not be removed during preprocessing
- Hyphens and punctuation marks
  - May be handled in different ways
- Most information retrieval systems perform case-insensitive search
- Text preprocessing steps language specific

Information Extraction

- Generic term
- Extracting structured content from text
- Mostly used to identify contextually relevant features that involve text analysis, matching, and categorization

Exercício 3

- Considerando o sistema de recuperação de rostos do Exercício 1, que tipo de pré-processamento deve ser feito nas imagens da base e nas imagens das consultas?

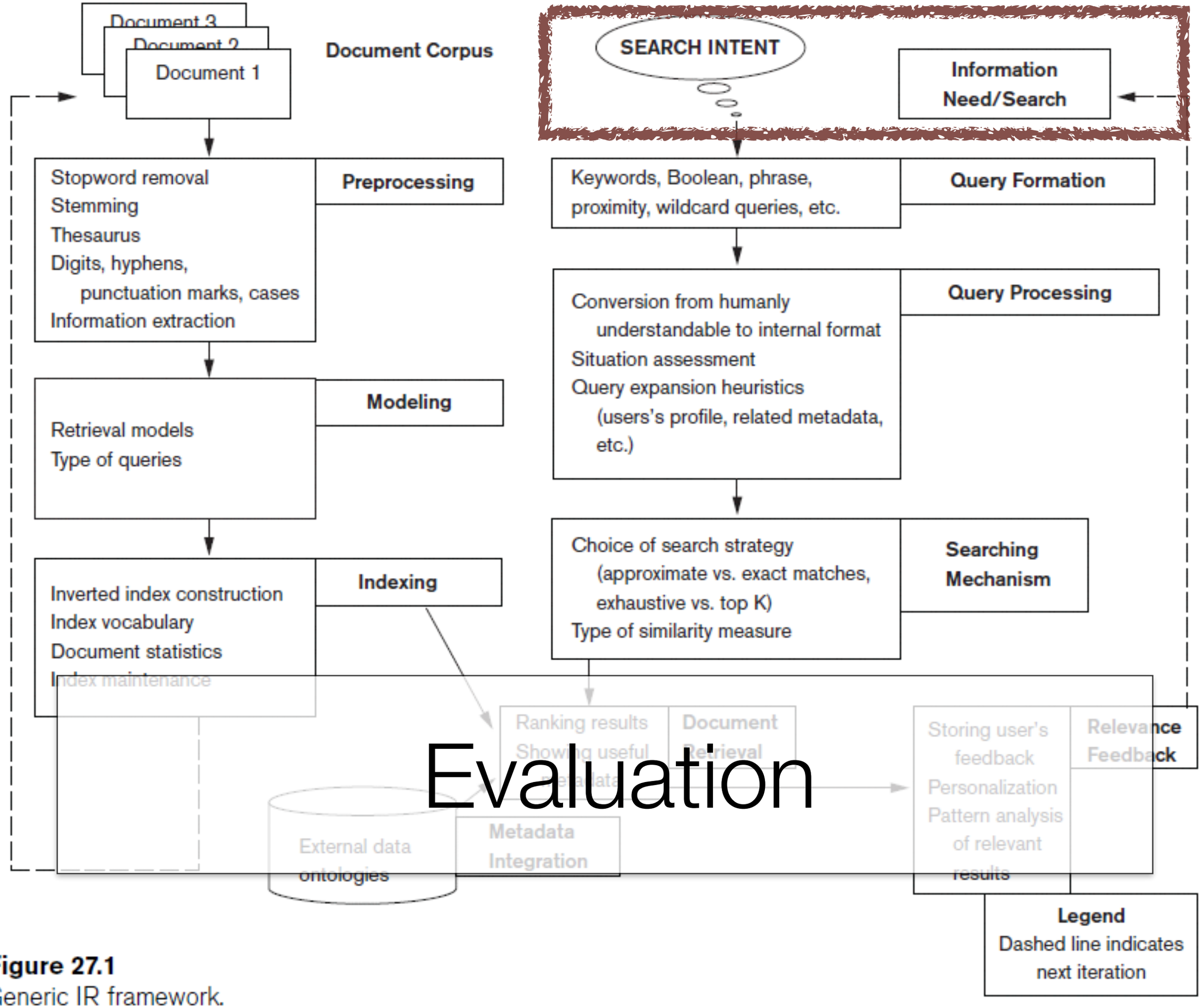


Figure 27.1
Generic IR framework.

Evaluation Measures of Search Relevance

- Topical relevance
  - Measures extent to which topic of a result matches topic of query
- User relevance
  - Describes “goodness” of a retrieved result with regard to user’s information need
- Web information retrieval
  - Must evaluate document ranking order

Recall and Precision

- Recall
  - Number of relevant documents retrieved by a search / Total number of existing relevant documents
- Precision
  - Number of relevant documents retrieved by a search / Total number of documents retrieved by that search
- Other combinations…

Searching the Web

- Hyperlink components
  - Destination page
  - Anchor text
- Hub
  - Web page or a Website that links to a collection of prominent sites (authorities) on a common topic

Analyzing the Link Structure of Web Pages

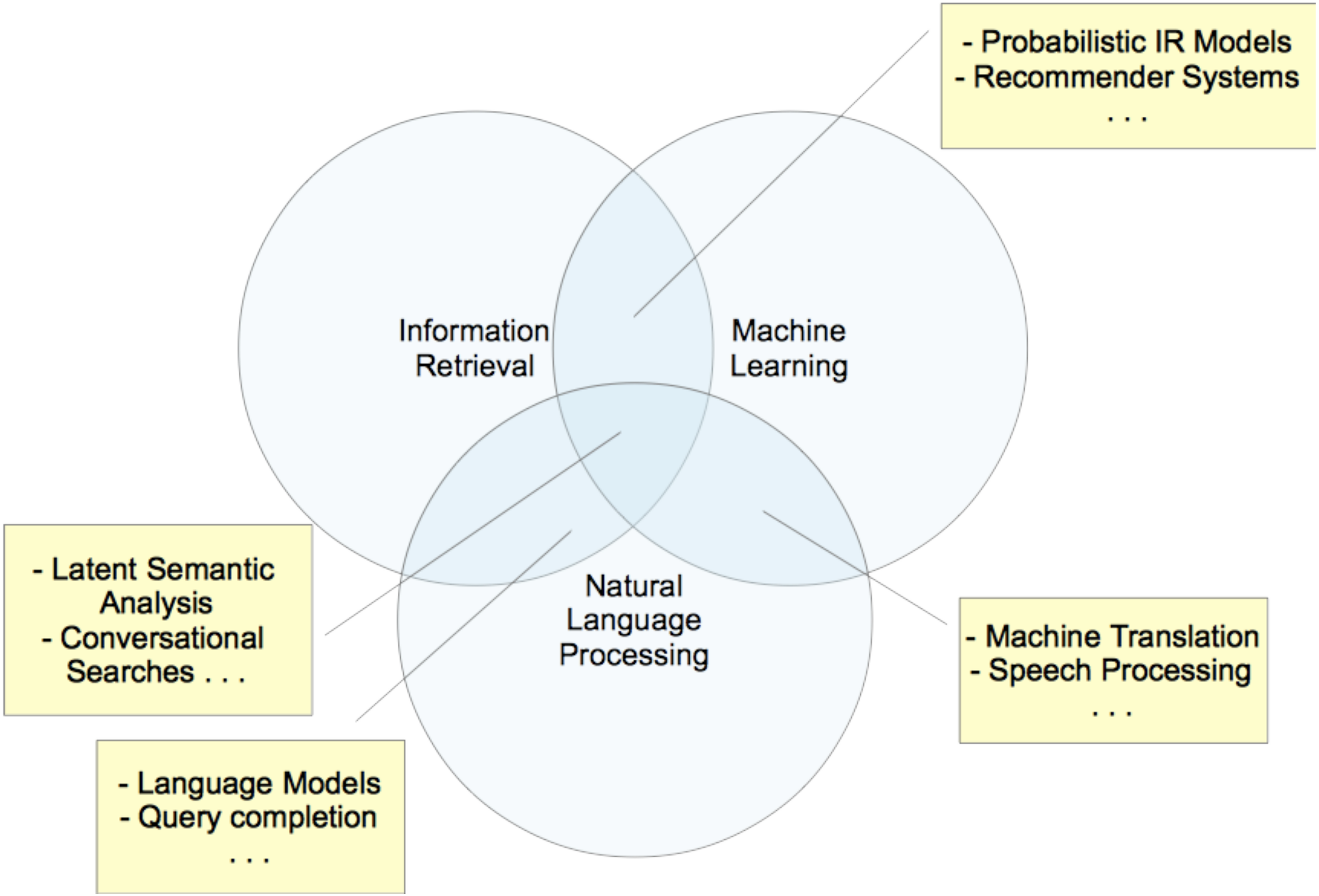
- The PageRank ranking algorithm
  - Used by Google
  - Highly linked pages are more important (have greater authority) than pages with fewer links
  - Measure of query-independent importance of a page/node
- HITS Ranking Algorithm
  - Contains two main steps: a sampling component and a weight-propagation component

Trends in Information Retrieval

- Faceted search
  - Allows users to explore by filtering available information
  - Facet: Defines properties or characteristics of a class of objects
- Social search
  - New phenomenon facilitated by recent Web technologies: collaborative social search, guided participation
- Conversational search (CS)
  - Interactive and collaborative information finding interaction
  - Aided by intelligent agents

Summary

- IR introduction
- Basic terminology, query and browsing modes, semantics, retrieval modes
- Web search analysis
- Content, structure, usage
- Algorithms
- Current trends


- Probabilistic IR Models
- Recommender Systems
. . .
Information Retrieval
Machine Learning
Natural Language Processing
- Latent Semantic Analysis
- Conversational Searches . . .
- Language Models
- Query completion
. . .
- Machine Translation
- Speech Processing
. . .

Exercício 1

- Imagine que você precisa desenvolver um sistema de recuperação de fotos de rostos de pessoas. O sistema recebe como entrada uma foto de um rosto e retorna fotos ordenadas por similaridade com a foto da consulta. O modelo escolhido para a implementação foi o de Vector Space.
- Que tipo de informação poderia ser usada nos vetores?
- Resposta: distância entre os olhos, distância entre os olhos e boca, porcentagem de prevalência de cores (histograma). . .

Exercício 2

- Faça o pseudo-código para um algoritmo que, baseado num índice invertido, processe consultas booleanas conjuntivas (a AND b AND c AND…).
- D = {} //documentos de resposta
- Para cada termo t na consulta
  - D = interseção de D e documentos contendo t obtidos no índice.
- Retorna D

Exercício 3

- Considerando o sistema de recuperação de rostos do Exercício 1, que tipo de pré-processamento deve ser feito nas imagens da base e nas imagens das consultas?

- Resposta: redimensionamento de imagens (para padronizar), recortar a imagem para que o rosto ocupe espaços parecidos em todas as fotos, rotacionar as fotos para alinhar, ajustar cores, iluminação, contraste, converter para preto e branco para otimizar . . .