



Knowledge Retrieval

Franz J. Kurfess

***Computer Science Department
California Polytechnic State University
San Luis Obispo, CA, U.S.A.***

Knowledge Retrieval

Franz J. Kurfess

***Computer Science Department
California Polytechnic State University
San Luis Obispo, CA, U.S.A.***

Acknowledgements

*Some of the material in these slides was developed for a lecture series sponsored by the **European Community** under the **BPD** program with **Vilnius University** as host institution*

Use and Distribution of these Slides

These slides are primarily intended for the students in classes I teach. In some cases, I only make PDF versions publicly available. If you would like to get a copy of the originals (Apple KeyNote or Microsoft PowerPoint), please contact me via email at fkurfess@calpoly.edu. I hereby grant permission to use them in educational settings. If you do so, it would be nice to send me an email about it. If you're considering using them in a commercial environment, please contact me first.

Overview Knowledge Retrieval

- ❖ Finding Out About

- ❖ Keywords and Queries; Documents; Indexing

- ❖ Data Retrieval

- ❖ Access via Address, Field, Name

- ❖ Information Retrieval

- ❖ Access via Content (Values); Parsing; Matching Against Indices; Retrieval Assessment

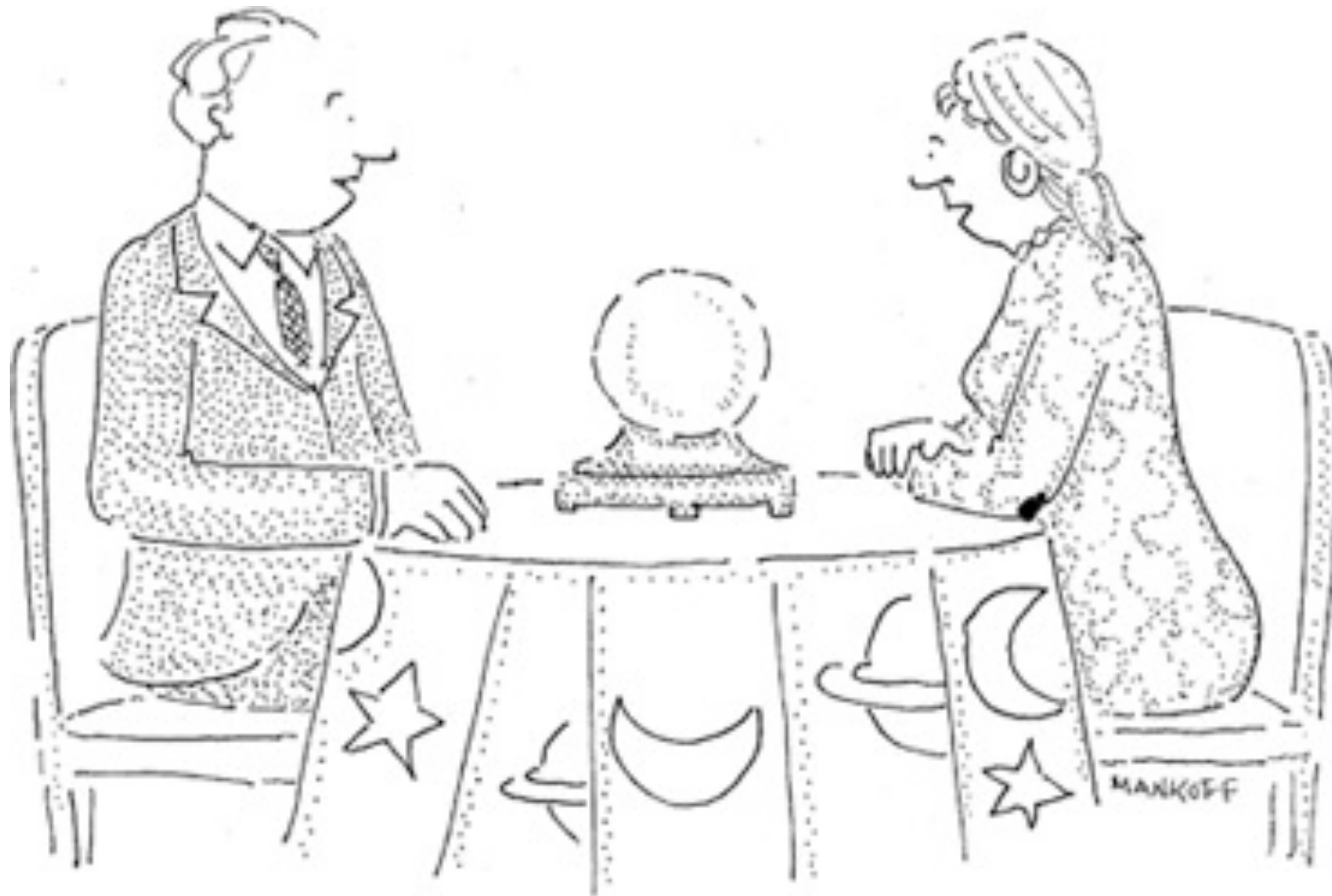
- ❖ Knowledge Retrieval

- ❖ Access via Structure; Meaning; Context; Usage

- ❖ Knowledge Discovery

- ❖ Data Mining; Rule Extraction

Finding Out About



*"What's the final episode
of 'Seinfeld' about?"*

"It's about nothing."

Finding Out About

- ❖ Keywords
- ❖ Queries
- ❖ Documents
- ❖ Indexing

Keywords

- ❖ linguistic atoms used to characterize the subject or content of a document
 - ❖ words
 - ❖ pieces of words (stems)
 - ❖ phrases
- ❖ provide the basis for a match between
 - ❖ the user's characterization of information need
 - ❖ the contents of the document
- ❖ problems
 - ❖ ambiguity

Queries

- ❖ formulated in a query language
 - ❖ natural language
 - ❖ interaction with human information providers
 - ❖ artificial language
 - ❖ interaction with computers
 - ❖ especially search engines
- ❖ vocabulary
 - ❖ controlled
 - ❖ limited set of keywords may be used
 - ❖ uncontrolled
 - ❖ any keywords may be used
- ❖ syntax

Documents

- ❖ general interpretation

- ❖ any document that can be represented digitally
 - ❖ text, image, music, video, program, etc.

- ❖ practical interpretation

- ❖ passage of text
 - ❖ strings of characters in an alphabet
 - ❖ written natural language
 - ❖ length may vary
- ❖ longer documents may be composed of shorter ones

Aboutness of Documents

- ❖ describes the suitability of a document as answer to a query
- ❖ assumptions
 - ❖ all documents have equal aboutness
 - ❖ the probability of any document in a corpus to be considered relevant is equal for all documents
 - ❖ simplistic; not valid in reality
- ❖ a paragraph is the smallest unit of text with appreciable aboutness

Structural Aspects of Documents

- ❖ documents may be composed of documents
 - ❖ paragraphs, subsections, sections, chapters, parts
 - ❖ footnotes, references
- ❖ documents may contain meta-data
 - ❖ information about the document
 - ❖ not part of the content of the document itself
 - ❖ may be used for organization and retrieval purposes
 - ❖ can be abused by creators
 - ❖ usually to increase the perceived relevance

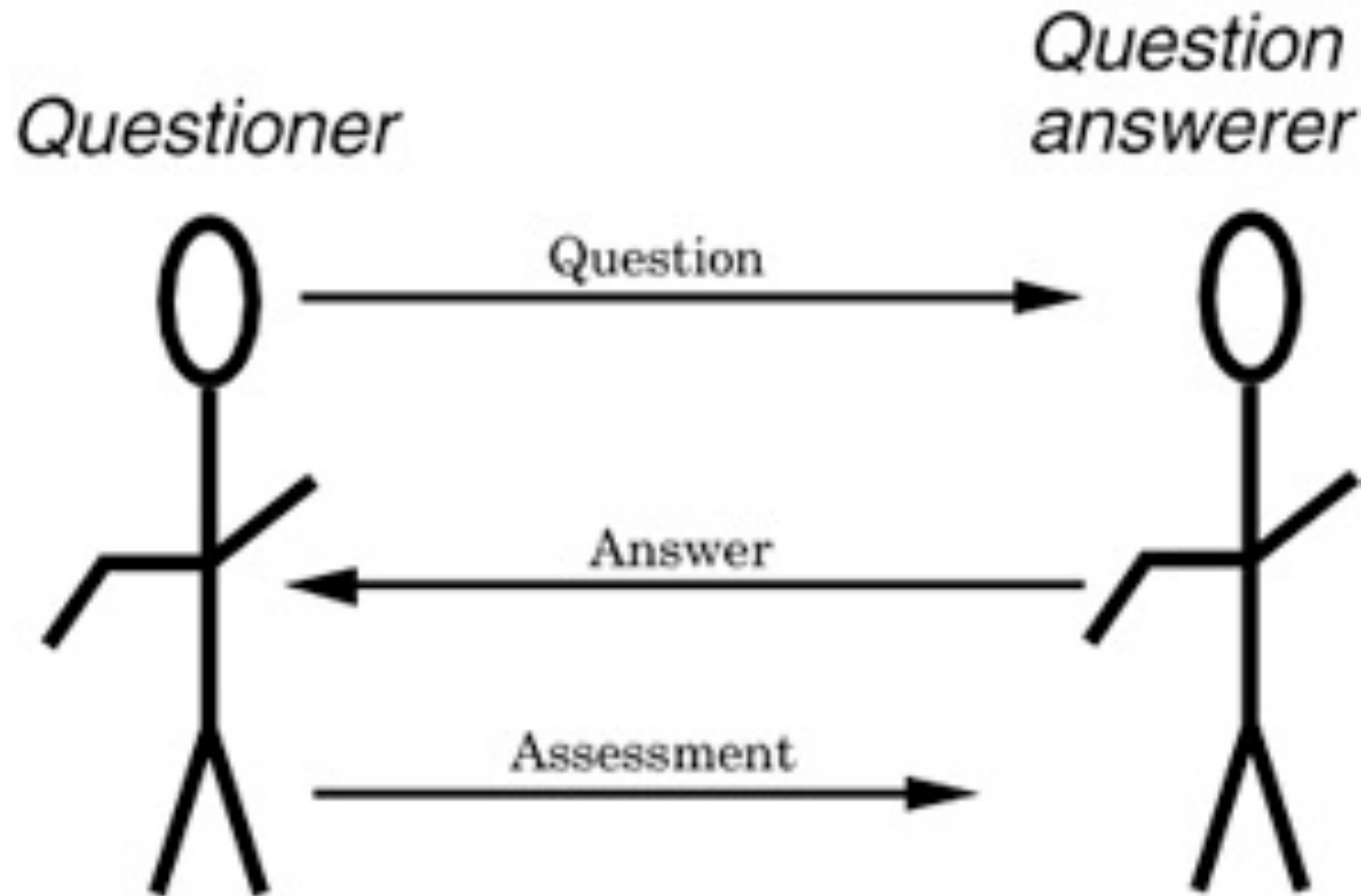
Document Proxies

- ❖ surrogates for the real document
 - ❖ abridged representations
 - ❖ catalog, abstract
 - ❖ pointers
 - ❖ bibliographical citation, URL
 - ❖ different media
 - ❖ microfiches
 - ❖ digital representations

Indexing

- ❖ a vocabulary of keywords is assigned to all documents of a corpus
- ❖ an index maps each document doc_i to the set of keywords $\{kw_j\}$ it is about
 $Index: doc_i \rightarrow^{about} \{kw_j\}$
 $Index^{-1}: \{kw_j\} \rightarrow^{describes} doc_i$
- ❖ indexing of a document / corpus
 - ❖ manual: humans select appropriate keywords
 - ❖ automatic: a computer program selects the keywords
- ❖ building the index relation between documents

FOA Conversation Loop



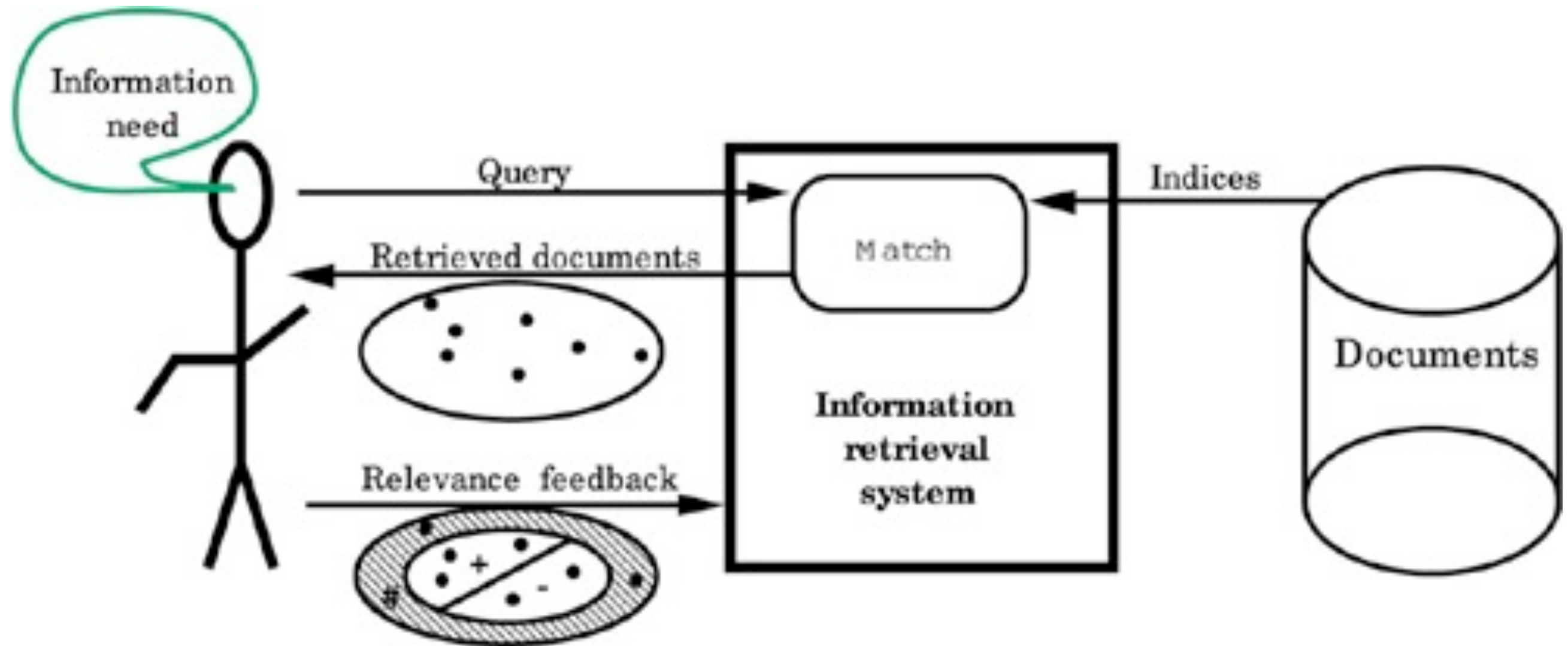
Data Retrieval

- ❖ access to specific data items
- ❖ access via address, field, name
- ❖ typically used in data bases
- ❖ user asks for items with specific features
 - ❖ absence or presence of features
 - ❖ values
- ❖ system returns data items
 - ❖ no irrelevant items
- ❖ deterministic retrieval method

Information Retrieval (IR)

- ❖ access to documents
 - ❖ also referred to as document retrieval
- ❖ access via keywords
- ❖ IR aspects
 - ❖ parsing
 - ❖ matching against indices
 - ❖ retrieval assessment

Diagram Search Engine



Parsing

- ❖ extraction of lexical features from documents
 - ❖ mostly words
- ❖ may require some manipulation of the extracted features
 - ❖ e.g. stemming of words
- ❖ used as the basis for automatic compilation of indices

Parsing Tools

- ❖ Montytagger <http://web.media.mit.edu/~hugo/montytagger/>
 - ❖ python and Java
- ❖ fnTBL (C++) <http://nlp.cs.jhu.edu/~rflorian/fntbl/>
 - ❖ fast
- ❖ Brill Tagger (C) <http://www.cs.jhu.edu/~brill/>
 - ❖ the original; influenced several later ones
- ❖ Natural Language Toolkit: <http://nltk.sourceforge.net/>
 - ❖ good starting point for basics of NLP algorithms

Matching Against Indices

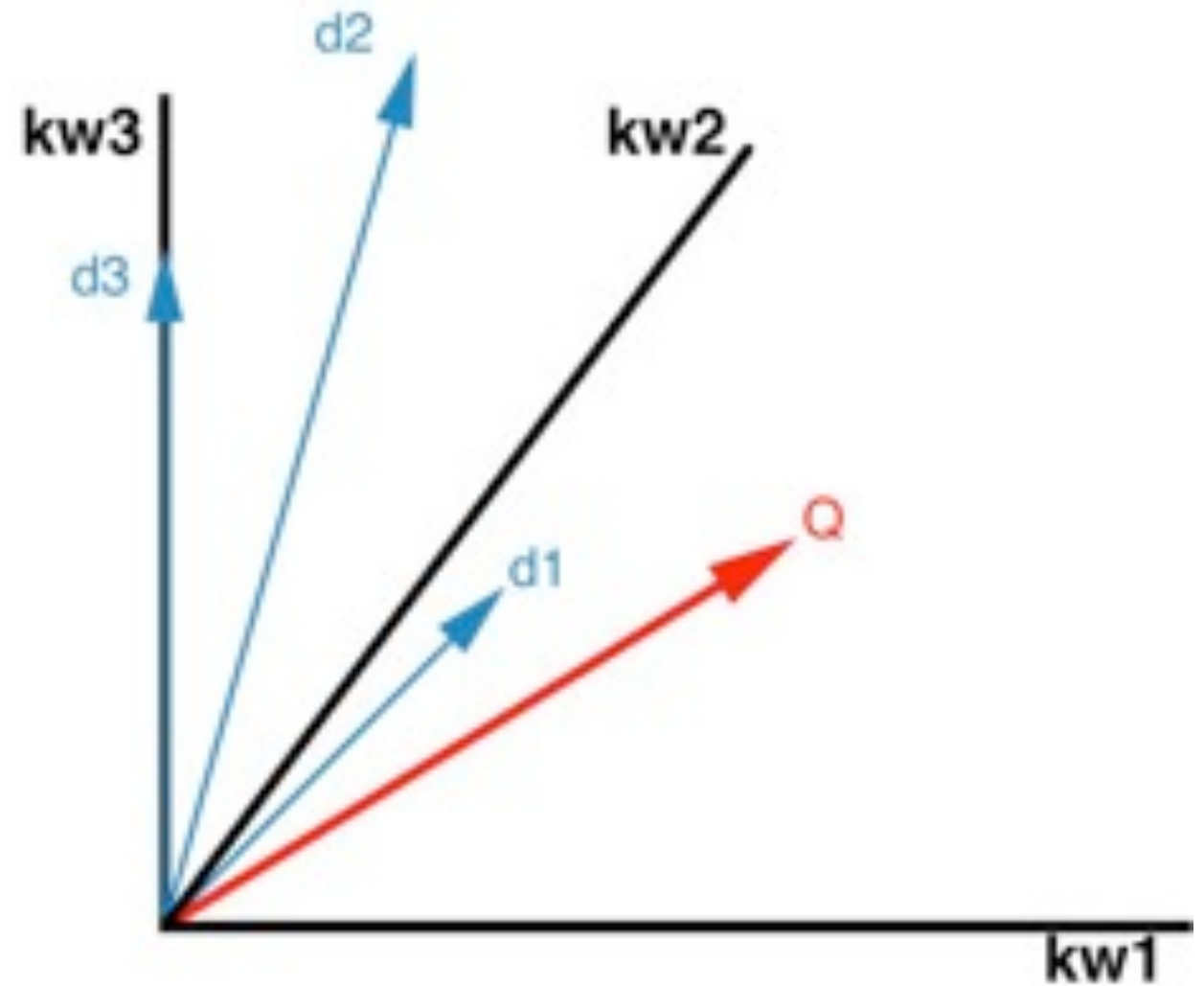
- ❖ identification of documents that are relevant for a particular query
- ❖ keywords of the query are compared against the keywords that appear in the document
 - ❖ either in the data or meta-data of the document
- ❖ in addition to queries, other features of documents may be used
 - ❖ descriptive features provided by the author or cataloger
 - ❖ usually meta-data
 - ❖ derived features computed from the contents of the document

Vector Space

- ❖ interpretation of the index matrix
 - ❖ relates documents and keywords
- ❖ can grow extremely large
 - ❖ binary matrix of 100,000 words * 1,000,000 documents
 - ❖ sparsely populated: most entries will be 0
- ❖ can be used to determine similarity of documents
 - ❖ overlap in keywords
 - ❖ proximity in the (virtual) vector space
- ❖ associative memories can be used as hardware implementation

Vector Space Diagram

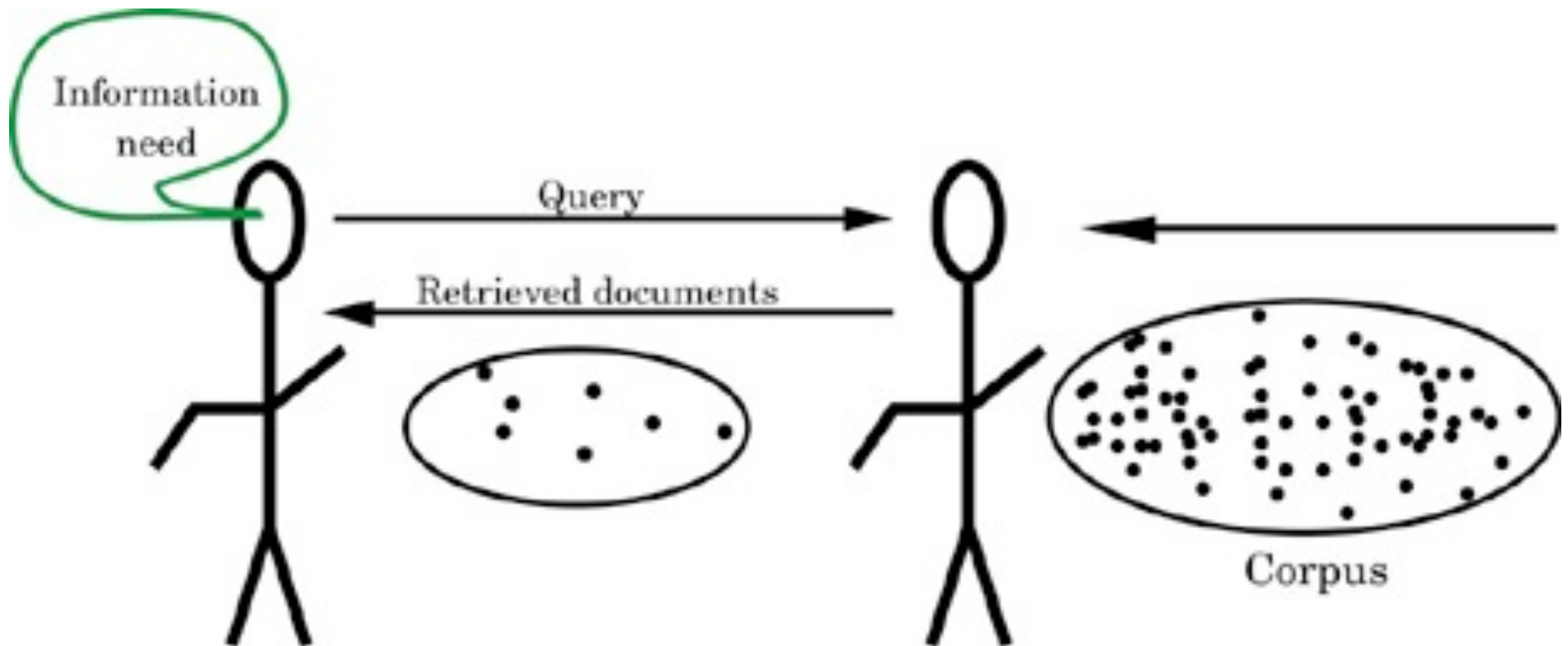
	kw1	kw2	kw3	...	kw _m
doc1	1	0	1	...	0
doc2	0	1	1	...	0
doc3	0	0	1	...	1
...					
docn	1	1	0	...	0
Q	1	0	0	...	1



Measuring Retrieval

- ❖ ideally, all relevant documents should be retrieved
 - ❖ relative to the query posed by the user
 - ❖ relative to the set of documents available (corpus)
 - ❖ relevance can be subjective
- ❖ precision and recall
 - ❖ relevant documents vs. retrieved documents

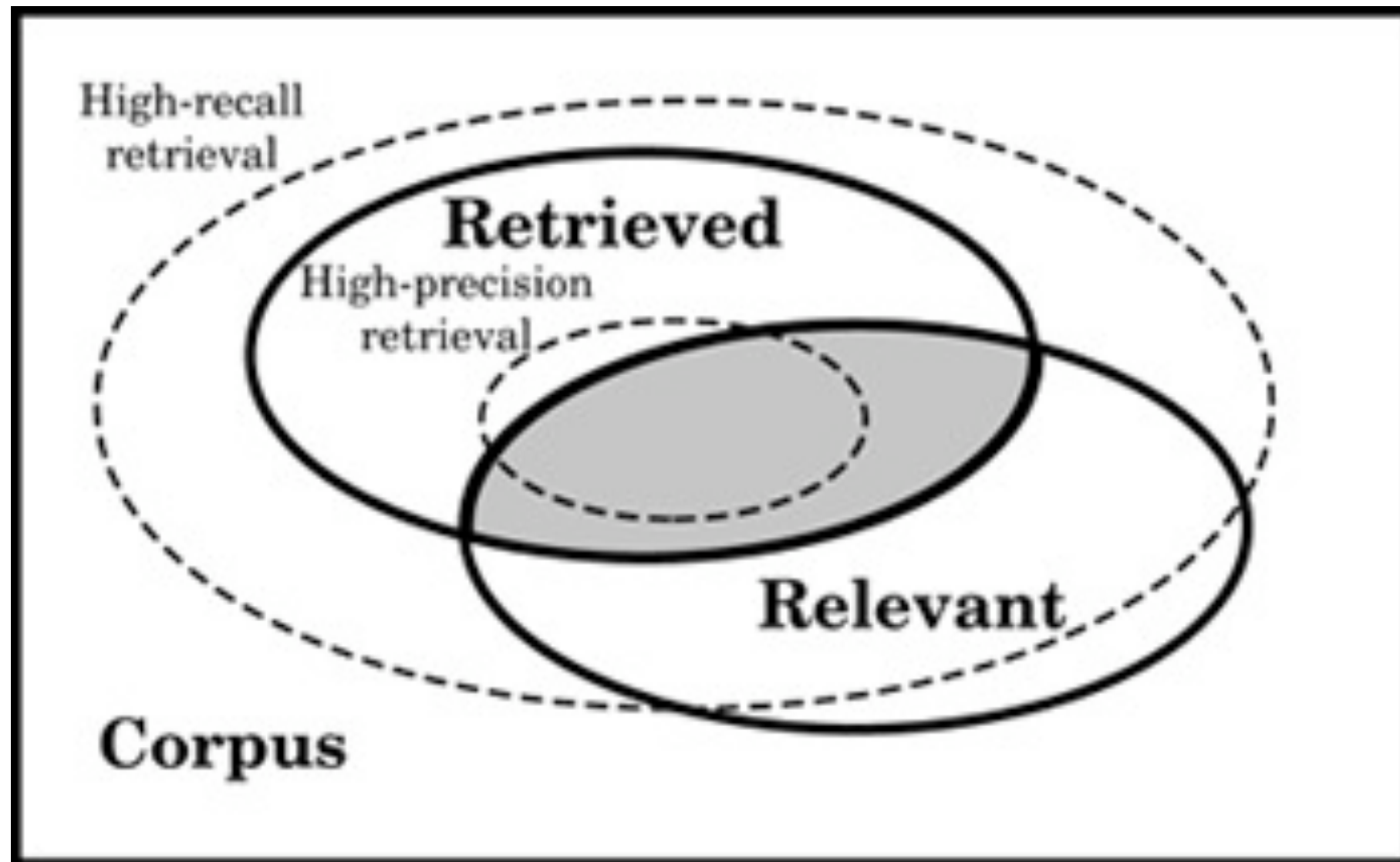
Document Retrieval



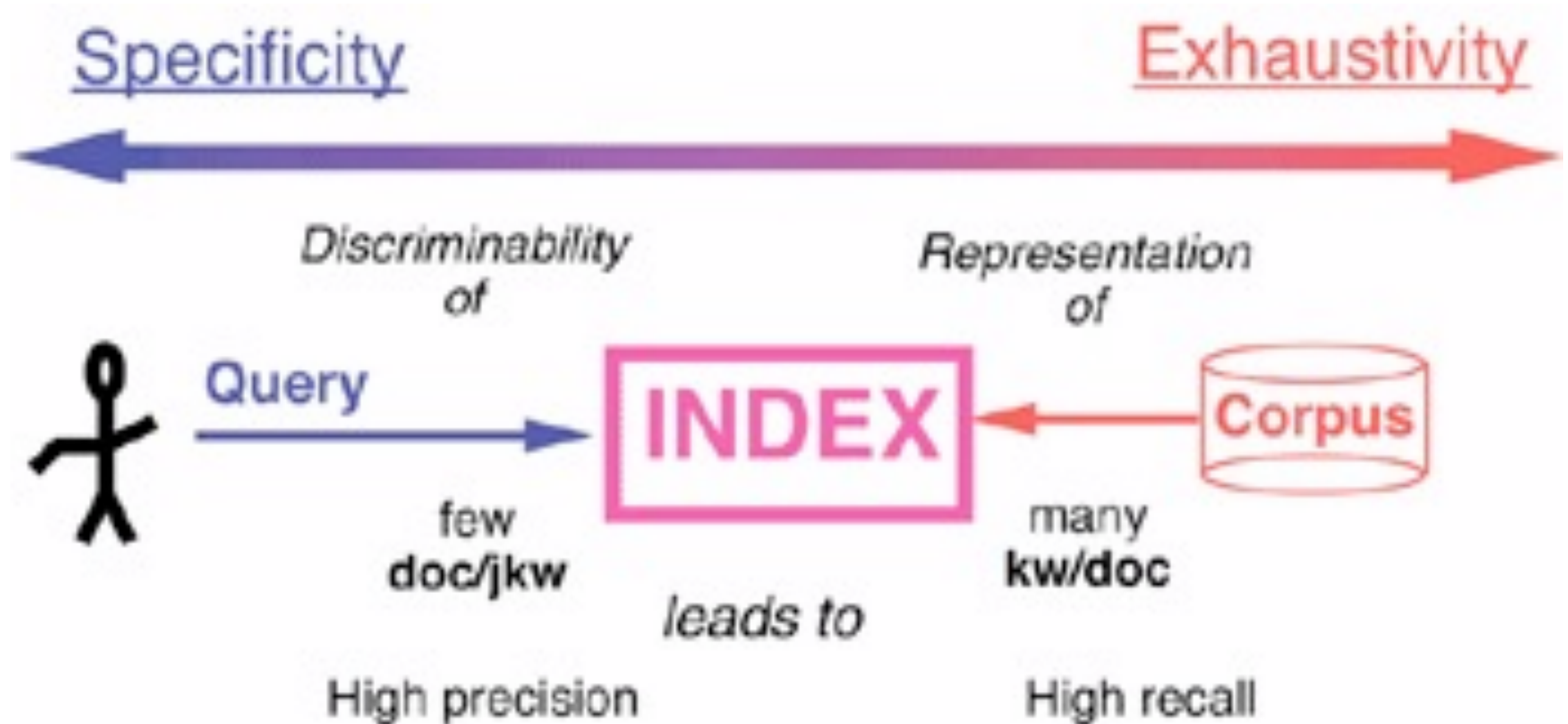
Precision and Recall

recall $\equiv |retrieved \cap relevant| / |relevant|$

precision $\equiv |retrieved \cap relevant| / |retrieved|$



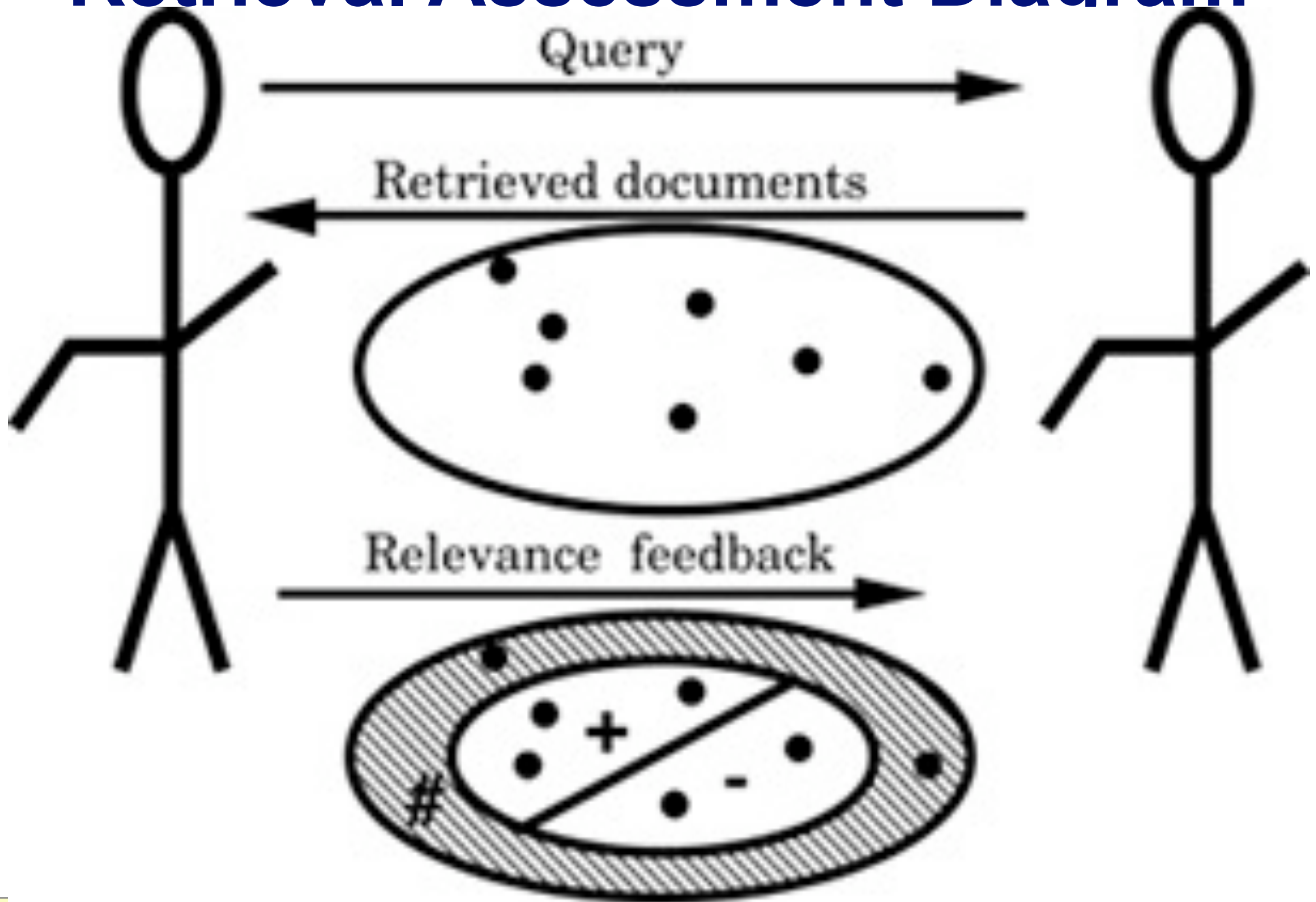
Specificity vs. Exhaustivity



Retrieval Assessment

- ❖ subjective assessment
 - ❖ how well do the retrieved documents satisfy the request of the user
- ❖ objective assessment
 - ❖ idealized omniscient expert determines the quality of the response

Retrieval Assessment Diagram

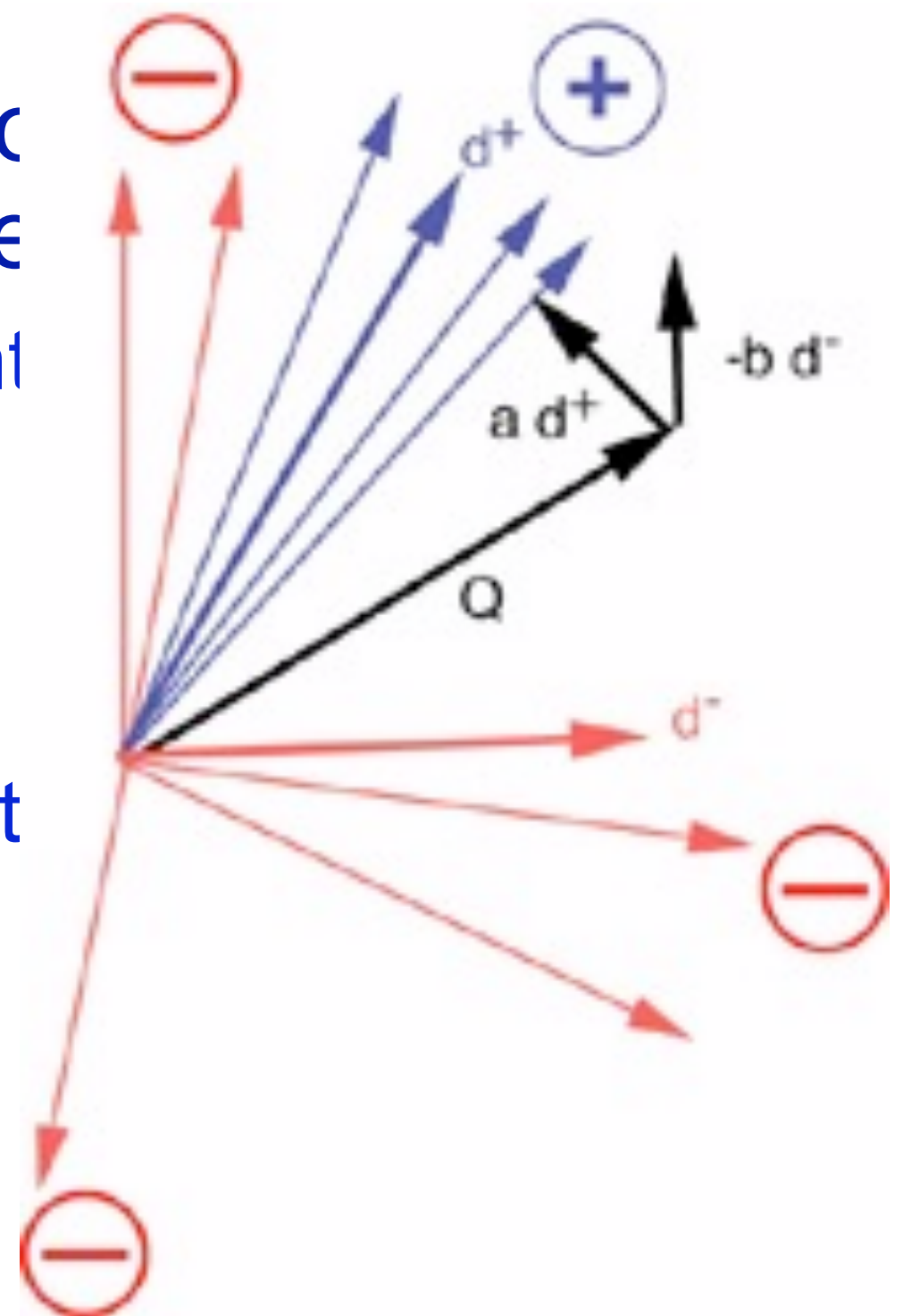


Relevance Feedback

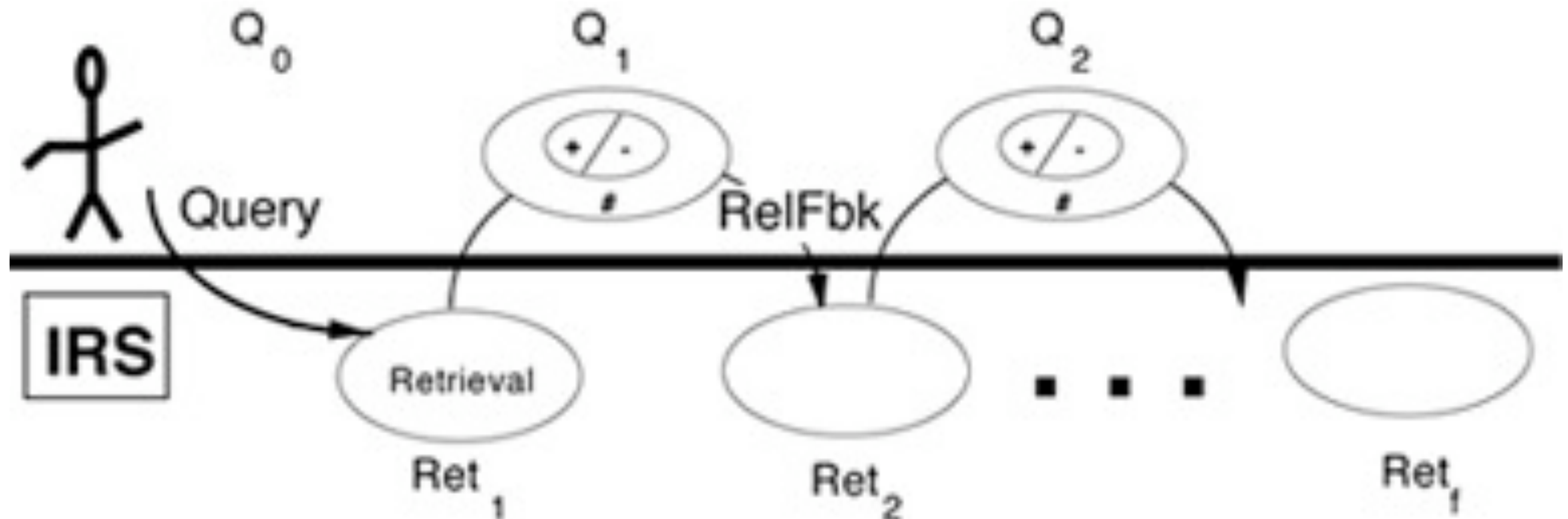
- ❖ subjective assessment of retrieval results
- ❖ often used to iteratively improve retrieval results
- ❖ may be collected by the retrieval system for statistical evaluation
- ❖ can be viewed as a variant of object recognition
 - ❖ the object to be recognized is the prototypical document the user is looking for
 - ❖ this document may or may not exist
 - ❖ the difference between the retrieved document(s) and the idealized prototype indicates the quality of the retrieval results

Relevance Feedback in Vector Space

- ❖ relevance feedback is used to move the query vector towards the cluster of positive documents and away from the cluster of negative documents
- ❖ moving away from bad documents improves the results
- ❖ it can also be used as a filter for a stream of documents
- ❖ as in news channels or similar situations



Query Session Example



Consensual Relevance

- ❖ relevance feedback from multiple users
 - ❖ identifies documents that many users found useful or interesting
 - ❖ used by some Web sites
 - ❖ related to collaborative filtering
 - ❖ can also be used as an evaluation method for search engines
 - ❖ performance criteria must be carefully considered
 - ❖ precision and recall, plus many others

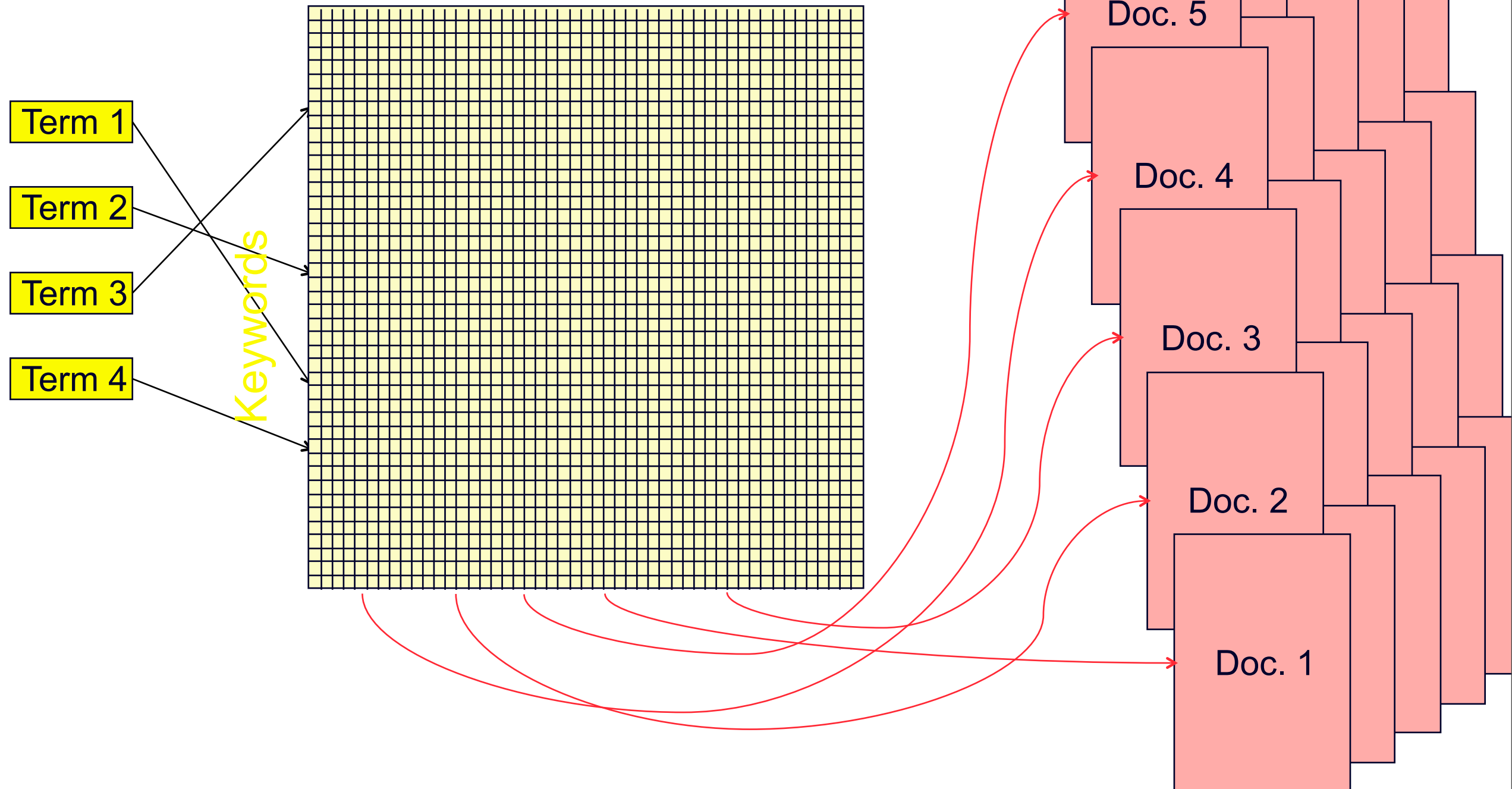
IR Diagram

Index

Query

Documents

Corpus



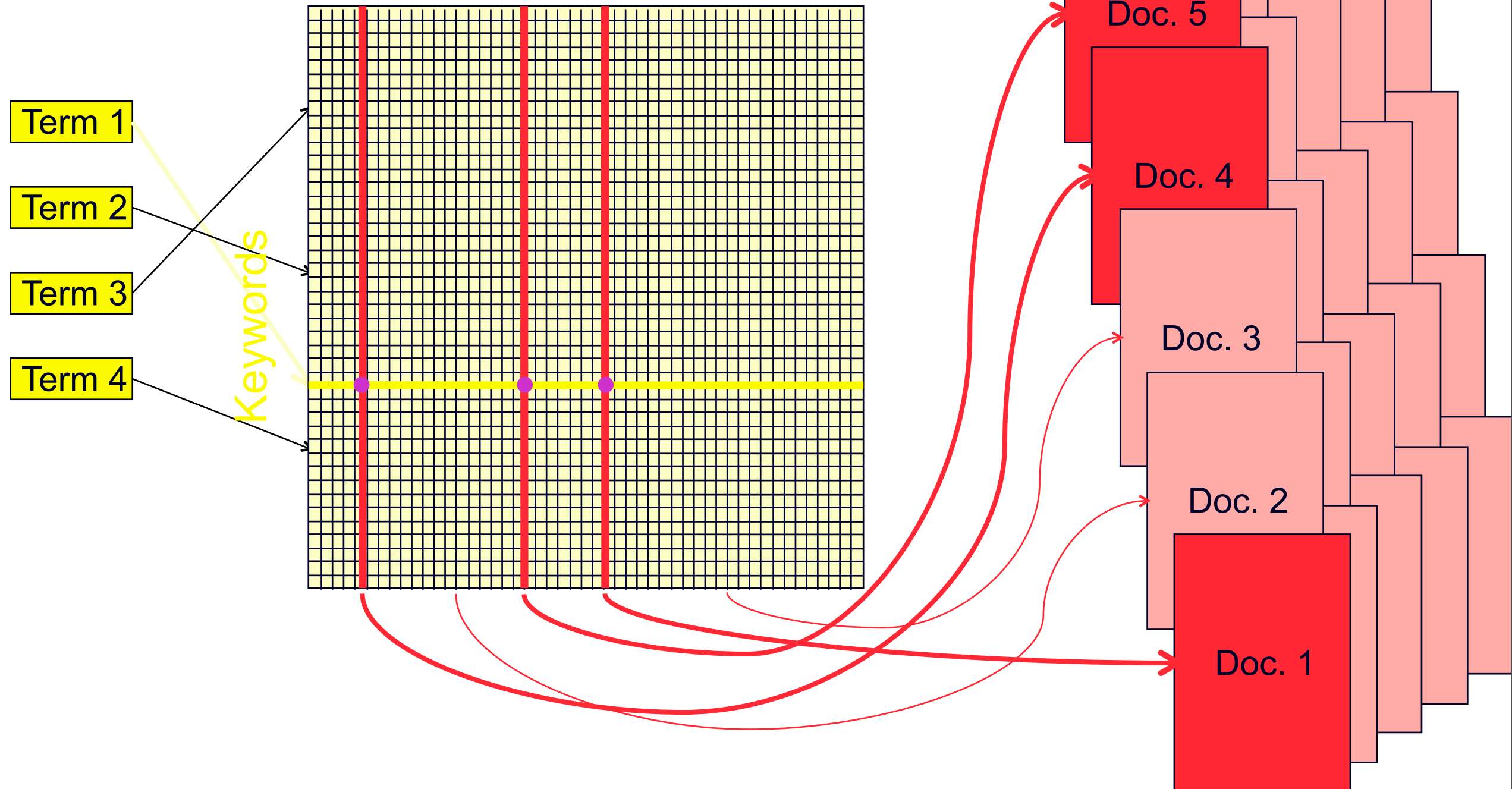
IR Diagram

Index

Query

Documents

Corpus



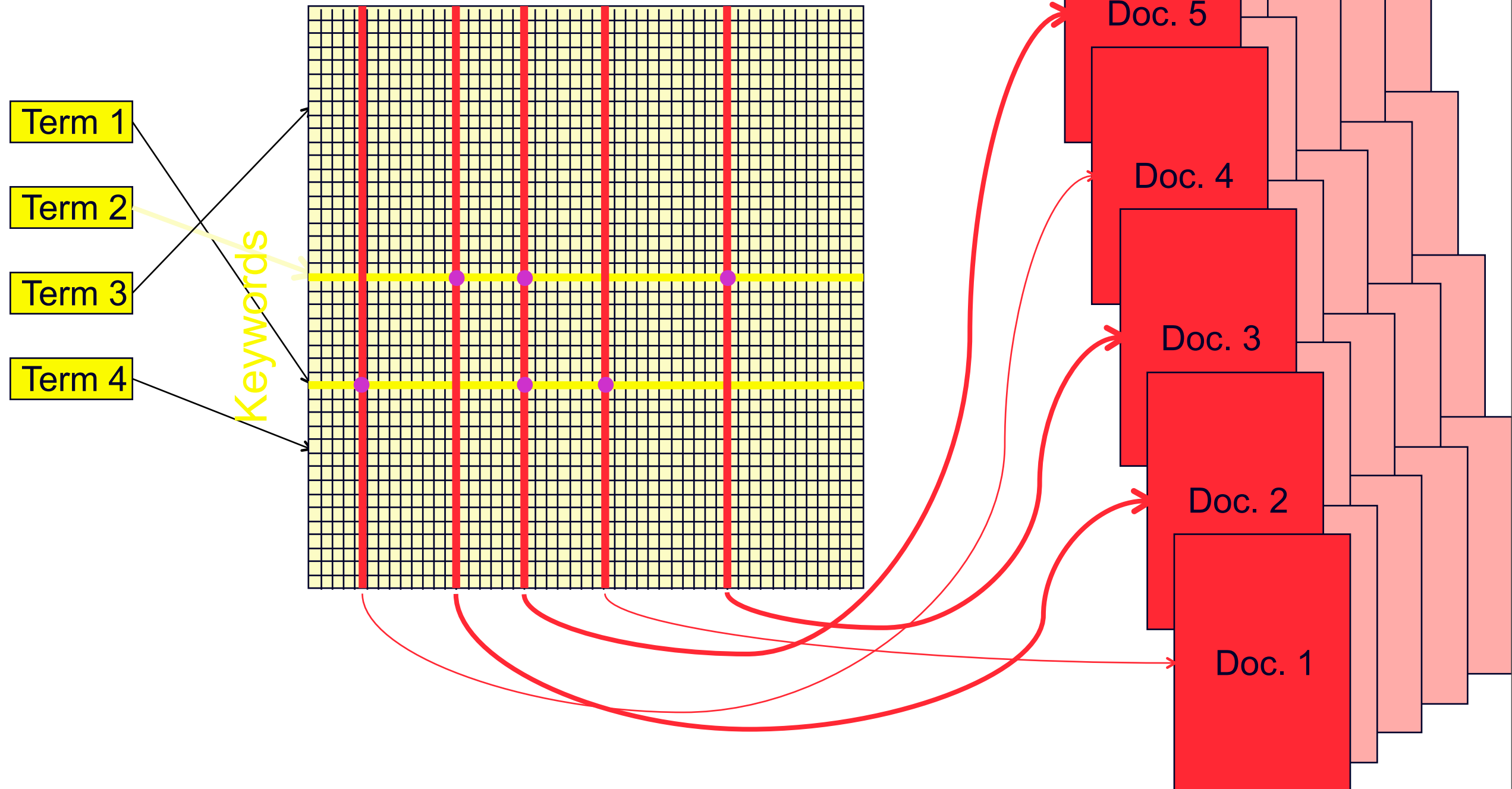
IR Diagram

Query

Index

Documents

Corpus



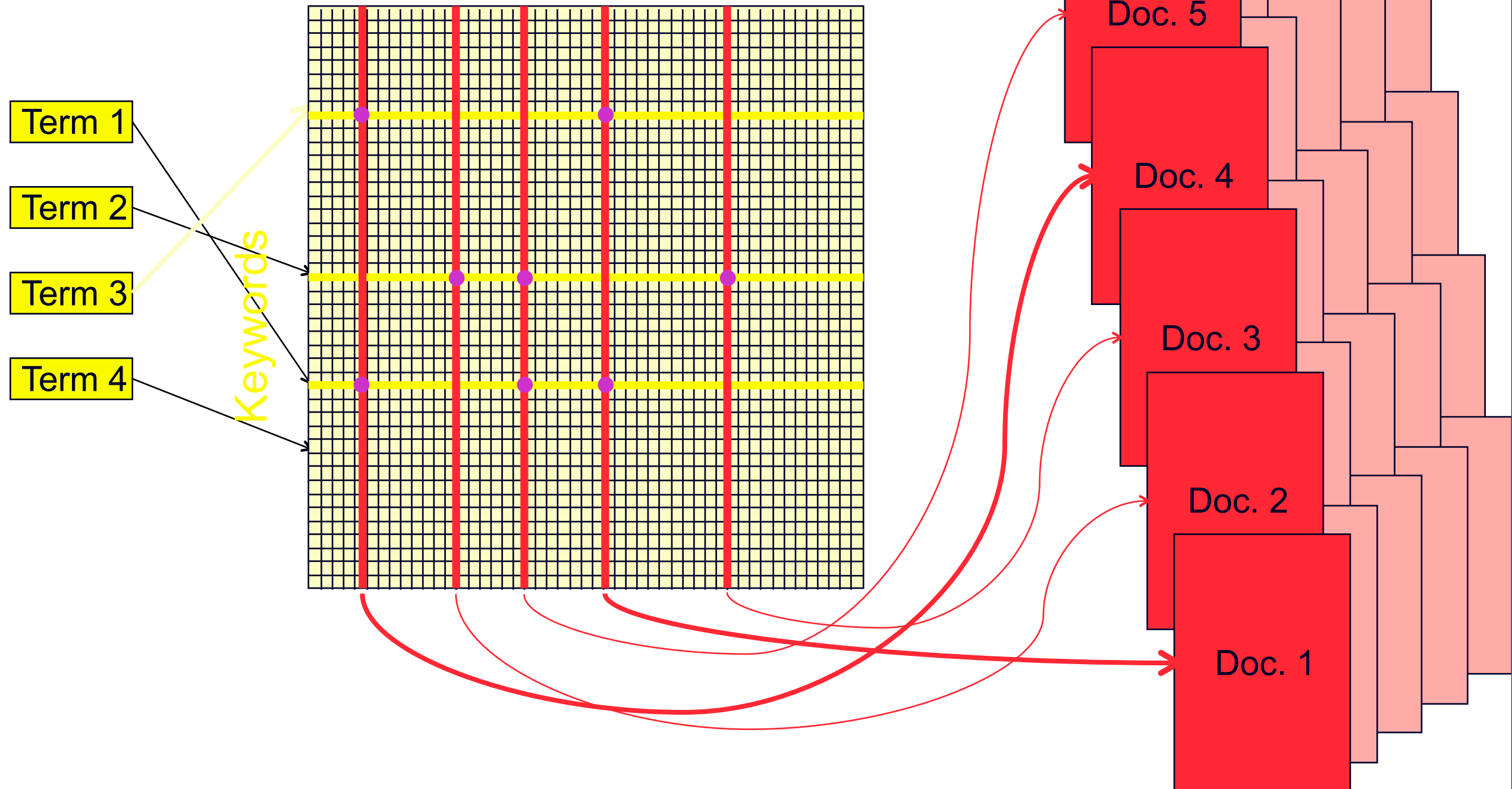
IR Diagram

Index

Query

Documents

Corpus



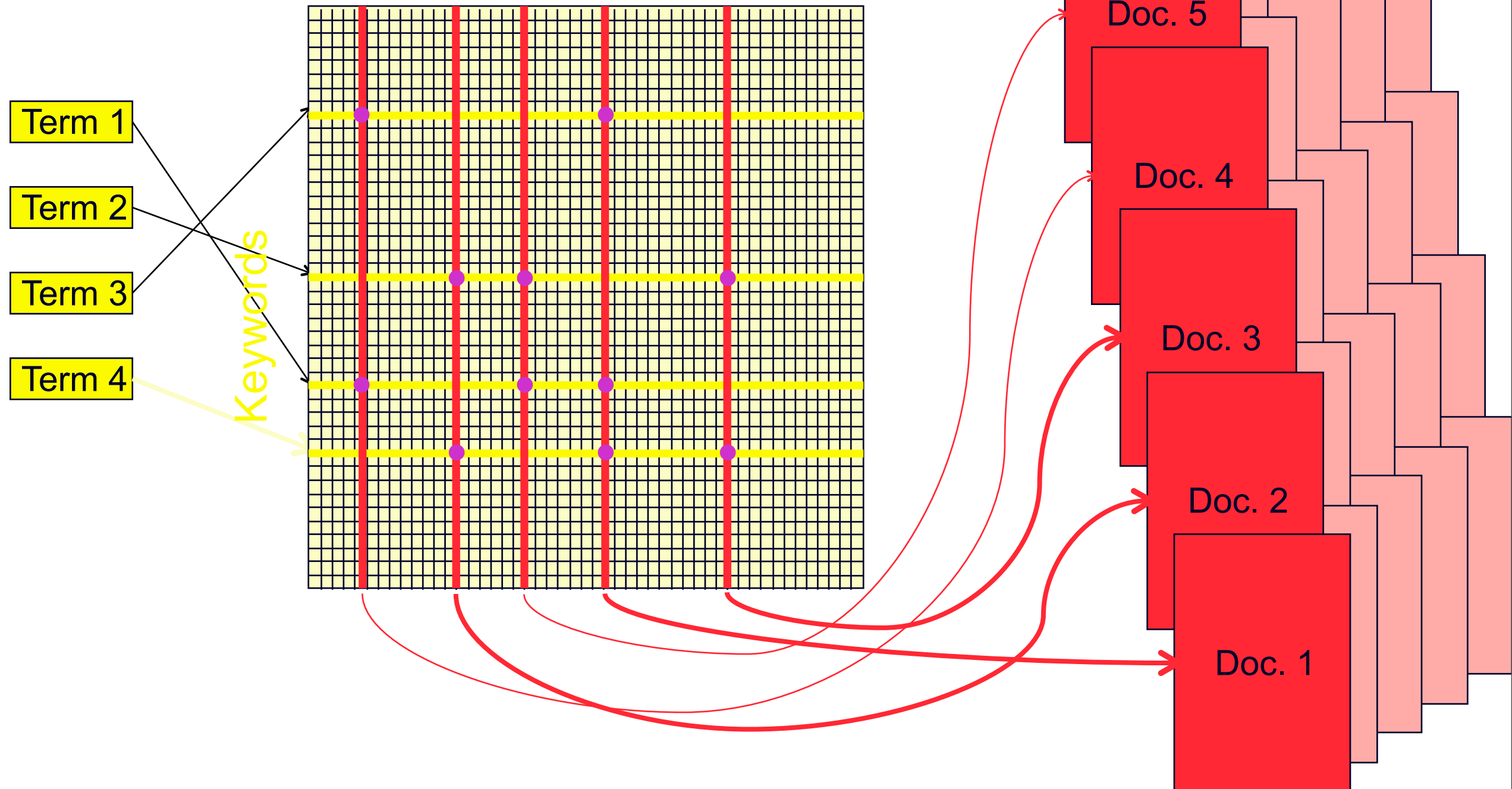
IR Diagram

Index

Query

Documents

Corpus



Knowledge Retrieval

- ❖ Context
- ❖ Usage
 - ❖ exploratory search
 - ❖ faceted search

Context in Knowledge Retrieval

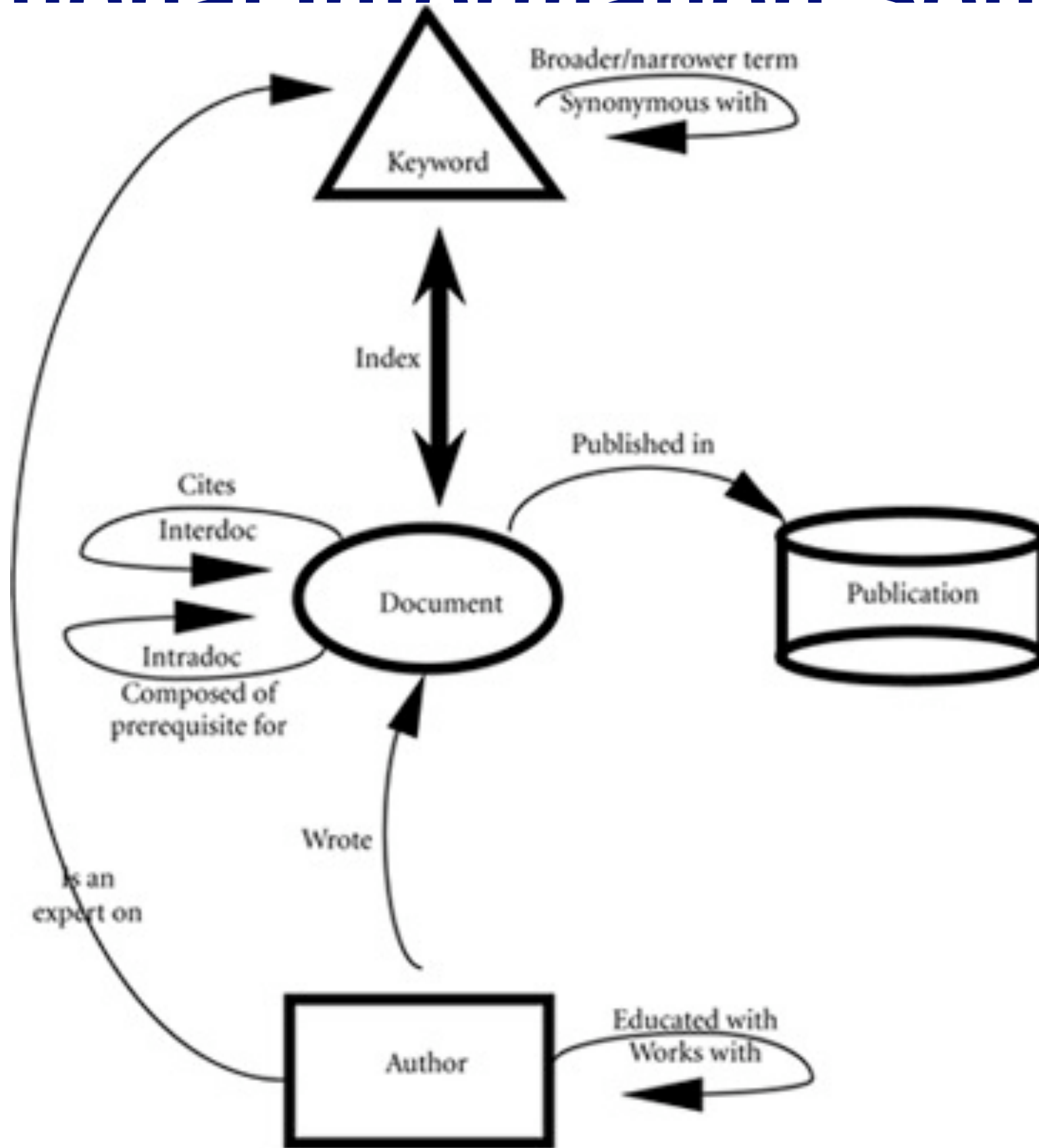
- ❖ in addition to keywords, relationships between keywords and documents are exploited
- ❖ explicit links
 - ❖ hypertext
- ❖ related concepts
 - ❖ thesaurus, ontology
- ❖ proximity

- ❖ spatial: place, directory
- ❖ temporal: creation date/time
- ❖ intermediate relations
 - ❖ author/creator
 - ❖ organization
 - ❖ project

Inference beyond the Index

- ❖ determines relationships between documents
- ❖ citations are explicit references to relevant documents
 - ❖ bibliographic references
 - ❖ legal citations
 - ❖ hypertext
- ❖ examples
 - ❖ NEC CiteSeer <<http://citeseer.nj.nec.com>>
 - ❖ Google Scholar <http://scholar.google.com>

Additional Information Sources



[Belew 2000, after Kochen 1975]

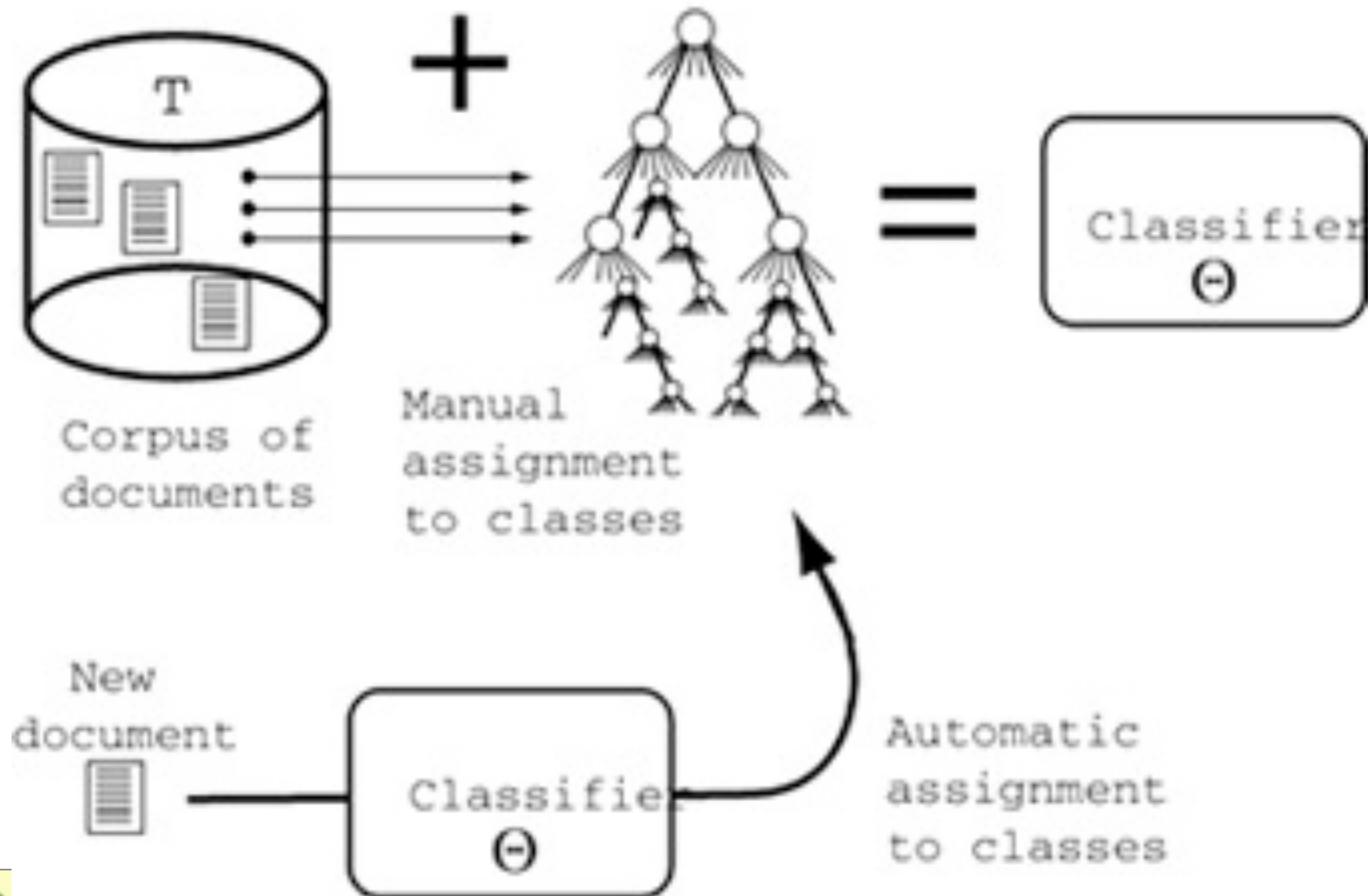
Hypertext

- ❖ inter-document links provide explicit relationships between documents
 - ❖ can be used to determine the relevance of a document for a query
 - ❖ example:
Google <<http://www.google.com>>
- ❖ intra-document links may offer additional context information for some terms
 - ❖ footnotes, glossaries, related terms

Adaptive Retrieval Techniques

- ❖ fine-tuning the matching between queries and retrieved documents
- ❖ learning of relationships between terms
 - ❖ training with term pairs (thesaurus)
 - ❖ pattern detection in past queries
 - ❖ automatic grouping of documents according to common features
- ❖ clustering of similar documents
 - ❖ pre-defined categories
 - ❖ metadata
 - ❖ overlap in keywords
 - ❖ consensual relevance
 - ❖ source

Document Classification



Query Model

- ❖ query types (templates)
 - ❖ frequently used types of queries
 - ❖ e.g. problem/solution, symptoms/diagnosis, problem/further checks, ...
- ❖ category types
 - ❖ abstractions of query types
 - ❖ used to determine categories or topics for the grouping of search results
- ❖ context information
 - ❖ current working document/directory
 - ❖ previous queries

Terminology Model

- ❖ individual terms are connected to related terms
 - ❖ thesaurus/ontology
 - ❖ synonyms, super-/sub-classes, related terms
- ❖ identifies labels for the category types

Matching

❖ categorizer

- ❖ determines the categories to be selected for the grouping of results
- ❖ assigns retrieved documents to the categories

❖ organizer

- ❖ arranges categories into a hierarchy
 - ❖ should be balanced and easy to browse by the user
- ❖ depends on the distribution of the search results

Results

- ❖ retrieved documents are grouped into hierarchically arranged categories meaningful for the user
- ❖ the categories are related to the query
- ❖ the categories are related to each other
- ❖ all categories have similar size
 - ❖ not always achievable due to the distribution of documents
- ❖ reduced search times
- ❖ higher user satisfaction

DynaCat

- ❖ knowledge-based approach to the organization of search results
- ❖ categorizes results into meaningful groups that correspond to the user's query
 - ❖ uses knowledge of query types and of the domain terminology to generate hierarchical categories
- ❖ applied to the domain of medicine
 - ❖ MEDLINE is an on-line repository of medical abstracts
 - ❖ 9.2 million bibliographic entries from 3800 journals
 - ❖ PubMed is a web-based search tool
 - ❖ returns titles as an relevance-ranked list

DyanCat Results

Query: What are the ways to prevent breast cancer...

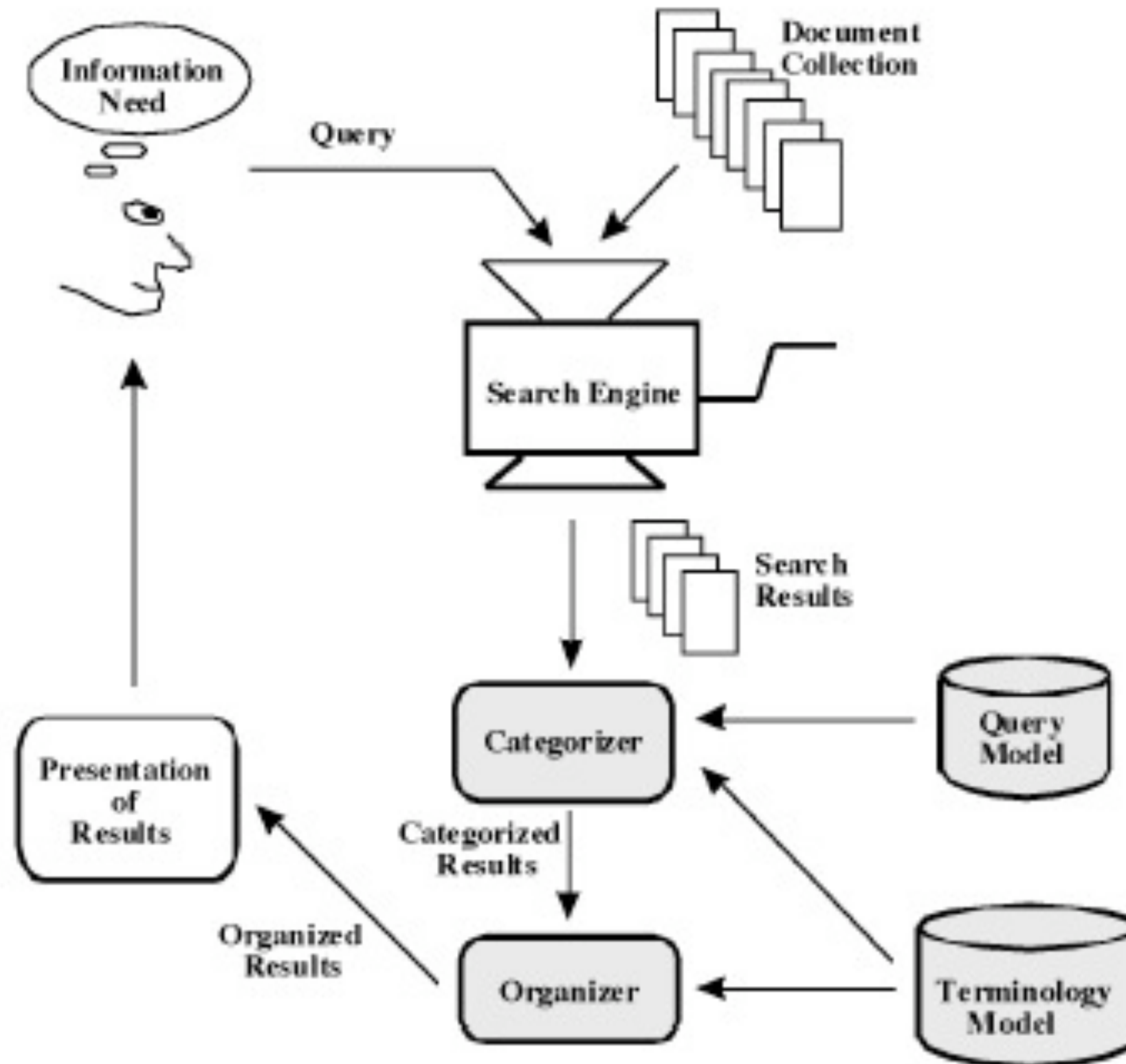
Query: What are the ways to prevent breast cancer?
(83 different references retrieved)

Behavior and Behavior Mechanisms (14 refs)	Biochemical Phenomena, Metabolism, and Nutrition
● Attitude (9 refs) ● Behavior (8 refs) ● Psychology, Social (3 refs)	● Diet ● <u>The etiology of breast cancer--from epidemiology to prevention.</u> ● <u>The role of dietary, genetic and hormonal factors in the development of breast cancer. Importance of adequate vitamin intake. Clinical and instrumental prevention.</u> ● <u>A study of diet and breast cancer prevention in Canada: why healthy women participate in controlled trials.</u> ● <u>Expanding horizons in breast and prostate cancer prevention and early detection. The 1992 Samuel C. Harvey Lecture</u>
Biochemical Phenomena, Metabolism, and Nutrition (5 refs) ● Diet (5 refs)	
Chemicals and Drugs (52 refs) ● Amino Acids, Peptides, and Proteins (2 refs) ● Antineoplastic and Immunosuppressive Agents (18)	

DynaCat Query Types

Query Type	Form of Question
Prevention	
problem—preventive actions	What can be done to prevent <problem>?
problem—risk-factors	What are the risk factors for <problem>?
Diagnosis	
problem—tests	What are the diagnostic tests for <problem>?
problem—symptoms	What are the warning signs and symptoms for <problem>?
symptoms—diagnoses	What are the possible diagnoses for <symptoms>?
Treatment	
problem—treatments	What are the treatments for <problem>?
treatment—problems	What are the adverse effects of <treatment>?
Prognosis	
Problem—prognostic-indicators	What are the factors that influence the prognosis for <problem>?
problem—prognoses	What is the prognosis for <problem>?

DynaCat Search



Information vs. Knowledge Retrieval

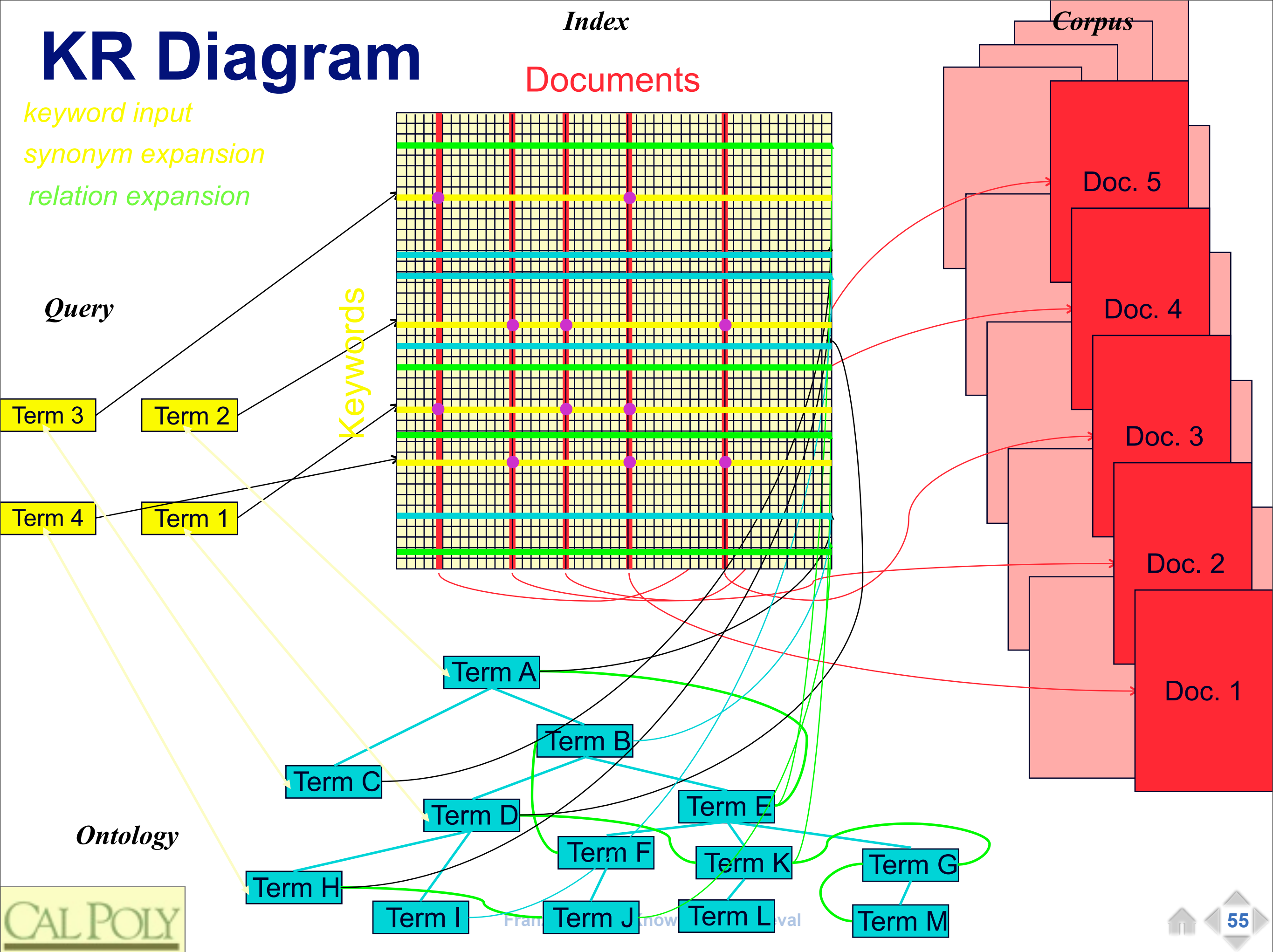
IR	KR
keywords as main components of the query	keywords plus context information for the query
index as match-making facility	index plus ontology for matching query and documents
statistical basis for selection of relevant documents	relationships between keywords and documents influence the selection of relevant documents
(ordered) list of results	results are grouped into meaningful categories

KR Diagram

keyword input

synonym expansion

relation expansion



Exploratory Search

- ❖ finding knowledge through association
- ❖ hypothesis: Human-made associations between knowledge items are valuable for others
- ❖ especially if the associations are made by experts or experienced users

Activity: Modern Exploratory Search

- ❖ What are current concepts, methods and tools that enable exploratory search?

Vannevar Bush: Memex

- ❖ better knowledge management for scientific document collections
- ❖ build, maintain, and share paths through the document space containing knowledge (“knowledge trails”)
- ❖ see Vannevar Bush, “As We May Think”, Atlantic Monthly, July 1945; www.theatlantic.com/194507/bush

Faceted Search

- ❖ exploration of a domain via attributes
 - ❖ select a relevant attribute, and display the elements of the domain ordered according to the attribute

Activity: Faceted Search Outside of Web Browsers

- ❖ What are tools or applications that employ faceted search to display items to the user?

	Name	Album by Year	Artist	Year
☒	Outrage	Now Hear This: Discover New Music – Rock	Capital Lights	2000
☒	My King (Falling Album Version)	Now Hear This: Discover New Music – Rock	Seven Stories Up	2000
☒	Traveler's Song	Now Hear This: Discover New Music – Rock	Future Of Forestry	2000
☒	Hold On	Now Hear This: Discover New Music – Rock	Abandon	2000
☒	I Can't Stand To Fall	Now Hear This: Discover New Music – Rock	Philmont	2000
☒	I'll Love You So	Now Hear This: Discover New Music – Rock	Above The Golden S...	2000
☒	Streetlight	Now Hear This: Discover New Music – Rock	Danyew	2000
☒	Taste And See	Now Hear This: Discover New Music – Pop	Jason Allen Rich	2000
☒	Life Light Up	Now Hear This: Discover New Music – Pop	Christy Nockels	2000
☒	Lord Of All	Now Hear This: Discover New Music – Pop	Kristian Stanfill	2000
☒	Chasing The Daylight	Now Hear This: Discover New Music – Pop	Phillip LaRue	2000
☒	Tell Me	Now Hear This: Discover New Music – Pop	Josh Wilson	2000
☒	Come Save	Now Hear This: Discover New Music – Pop	Sarah Reeves	2000
☒	Walk Tall (feat. Paul Simon)	Family Time	Ziggy Marley	2000
☒	Invisible Cities	Invisible Cities	Nomo	2000
☒	Seasons	Seasons	The Veer Union	2000
☒	Sugarfoot	Sugarfoot	Black Joe Lewis & T...	2000
☒	In My Time Of Dyin'	Time To Grow	The Lovell Sisters	2000
☒	Ninth Place	Lay Your Burden Down	Buckwheat Zydeco	2000
☒	Mama	Mama	Holly Williams	2000
☒	Ride The Nuclear Wave	Be On The Lookout!	The Oranges Band	2000
☒	Shrapnel	Be On The Lookout!	American Steel	2000
☒	Deconstruct/Rebuild	Be On The Lookout!	Small Brown Bike	2000
☒	Face It	Be On The Lookout!	The Reputation	2000
☒	Ear Nose And Throat	Be On The Lookout!	Troubled Hubble	2000
☒	Eating Toothpaste	Be On The Lookout!	Bratmobile	2000
☒	Under The Hedge	Be On The Lookout!	Ted Leo/Pharmacists	2000
☒	Sorry For Freaking Out On The Phone Last Night	Be On The Lookout!	The Mr. T Experience	2000
☒	Are You Gonna Move It For Me?	Be On The Lookout!	The Donnas	2000
☒	Fortune Cookie	The Further Adventures of Los Straitjackets	Los Straitjackets	2000
☒	I Wanna Know Why	The Leaves are Right to Tremble	Justin Rosolino	2000
☒	The Show Is On The Road	The Show Is On The Road	Paleface	2000
☒	Oh My Soul	Little Red	Susan Marshall	2000
☒	Excursion Around The Bay	Great Big Sea	Great Big Sea	2000
☒	Yours For the Taking	Smoking Kills	The Disciplines	2000
☒	Songs in the Night	Songs in the Night	Samantha Crain and...	2000
☒	Little Bit of Heaven	Cheval Sombre	Cheval Sombre	2000
☒	She Loves Everybody	She Loves Everybody	Chester French	2000
☒	Longing For	Longing For	Ballas Hough Band	2000
☒	Perfect	Emo Is Awesome Emo Is Evil 2	Sounds Like Violence	2000

2411 items, 5.9 days, 12.15 GB

Variations on Faceted Search

- ❖ displaying lists of items ordered according to an attribute can get quite boring
- ❖ attributes often lend themselves to alternative presentation methods
 - ❖ visual
 - ❖ static
 - ❖ color, size, shape
 - ❖ dynamic
 - ❖ movement, changes over time
 - ❖ auditory
 - ❖ often for supplementary information

Knowledge Discovery

- ❖ combination of
 - ❖ Data Mining
 - ❖ Knowledge Extraction
 - ❖ Knowledge Fusion

Data Mining

- ❖ identification of interesting “nuggets” in huge quantities of data
 - ❖ often relations between subsets
 - ❖ automatic or semi-automatic
- ❖ techniques
 - ❖ classification, correlation (e.g. temporal, spatial)

Knowledge Extraction

- ❖ conversion of internal representations of knowledge into human-understandable format
- ❖ extraction of rules from neural networks is one example

Knowledge Fusion

- ❖ multiple pieces of information are combined into one
- ❖ redundancy
 - ❖ do several pieces contain the same type of information
- ❖ compatibility
 - ❖ do the individual pieces have similar formats and interpretations
 - ❖ are there mappings to convert values into the same format
- ❖ consistency
 - ❖ are the values of the individual pieces close

Summary Knowledge Retrieval

- ❖ identification, selection, and presentation of documents relevant to a user query
- ❖ utilization of structural information, context, meta-data in addition to keyword search
- ❖ organized presentation of results
 - ❖ categories, visual arrangement
- ❖ internal representations may be converted to human-understandable ones

