Knowledge Retrieval

Franz J. Kurfess

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

Acknowledgements

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."

[Belew 2000]

Finding Out About

Keywords
Queries
Documents
Indexing

[Belew 2000]

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
  - choice of keywords

[Belew 2000]

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
    - often Boolean operators (AND, OR)
    - sometimes regular expressions

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

[Belew 2000]

Structural Aspects of Documents

- documents may be composed of other smaller pieces, or other 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

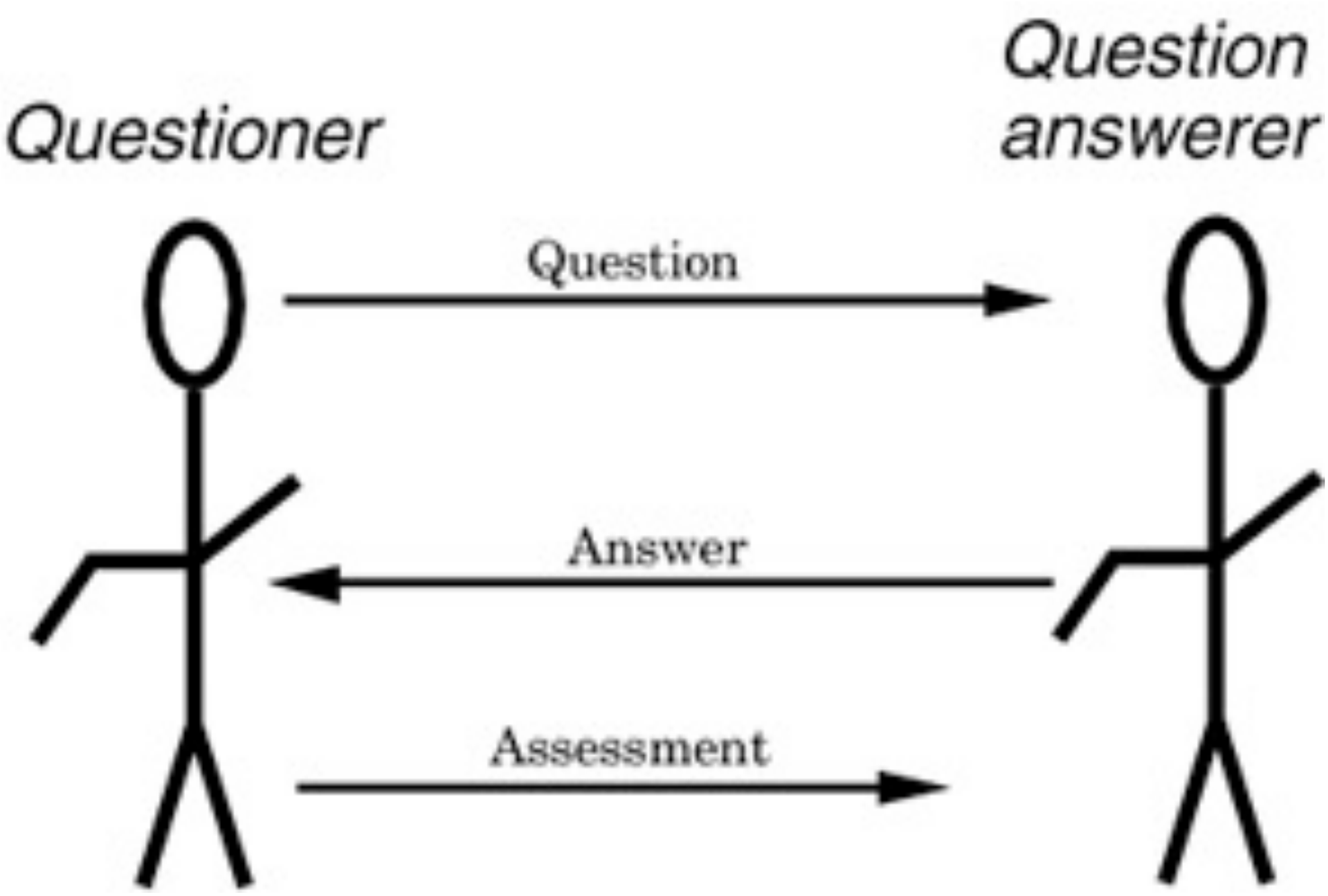
- surrogates for the real document
  - abridged representations
    - catalog, abstract
  - pointers
    - bibliographical citation, URL
  - different media
    - microfiches
    - digital representations

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 and sets of keywords is critical for information retrieval

[Belew 2000]

FOA Conversation Loop



[Belew 2000]

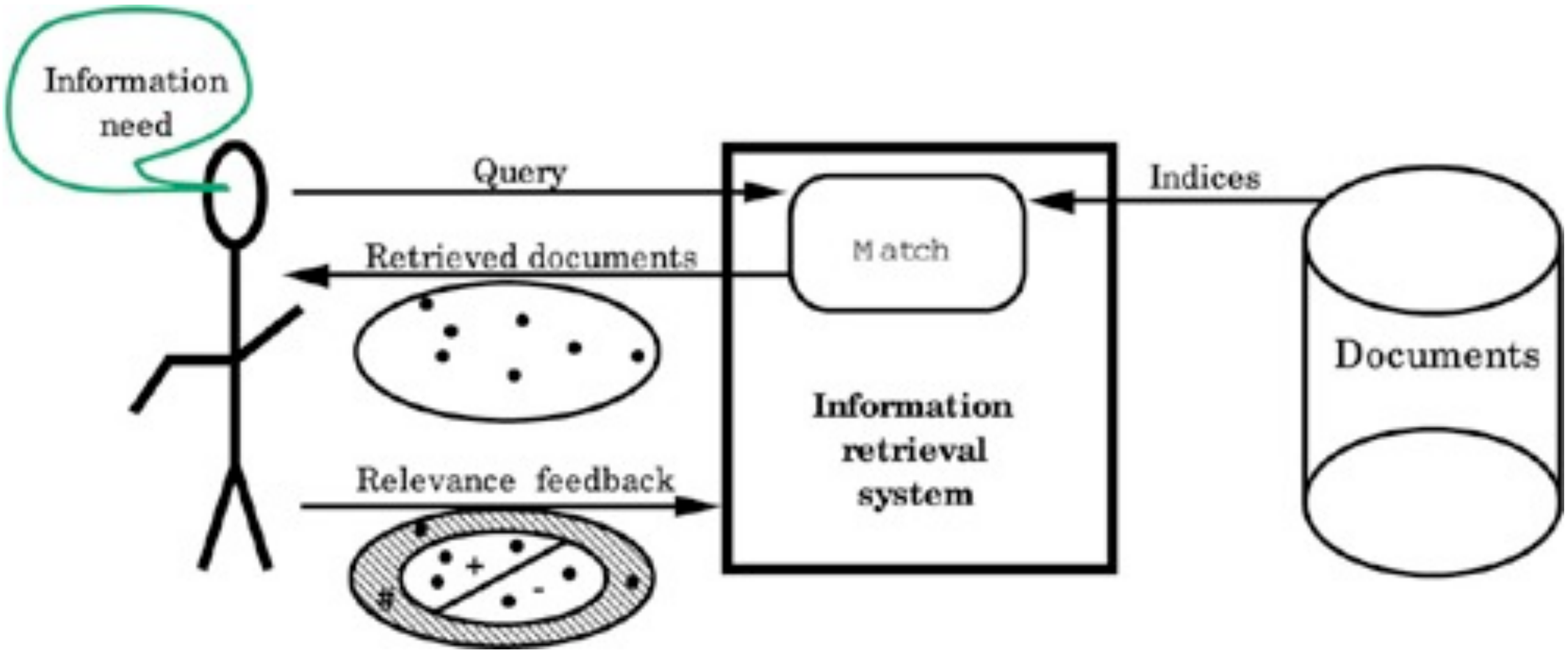
Data Retrieval

- 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



[Belew 2000]

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

[Belew 2000]

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

[Belew 2000]

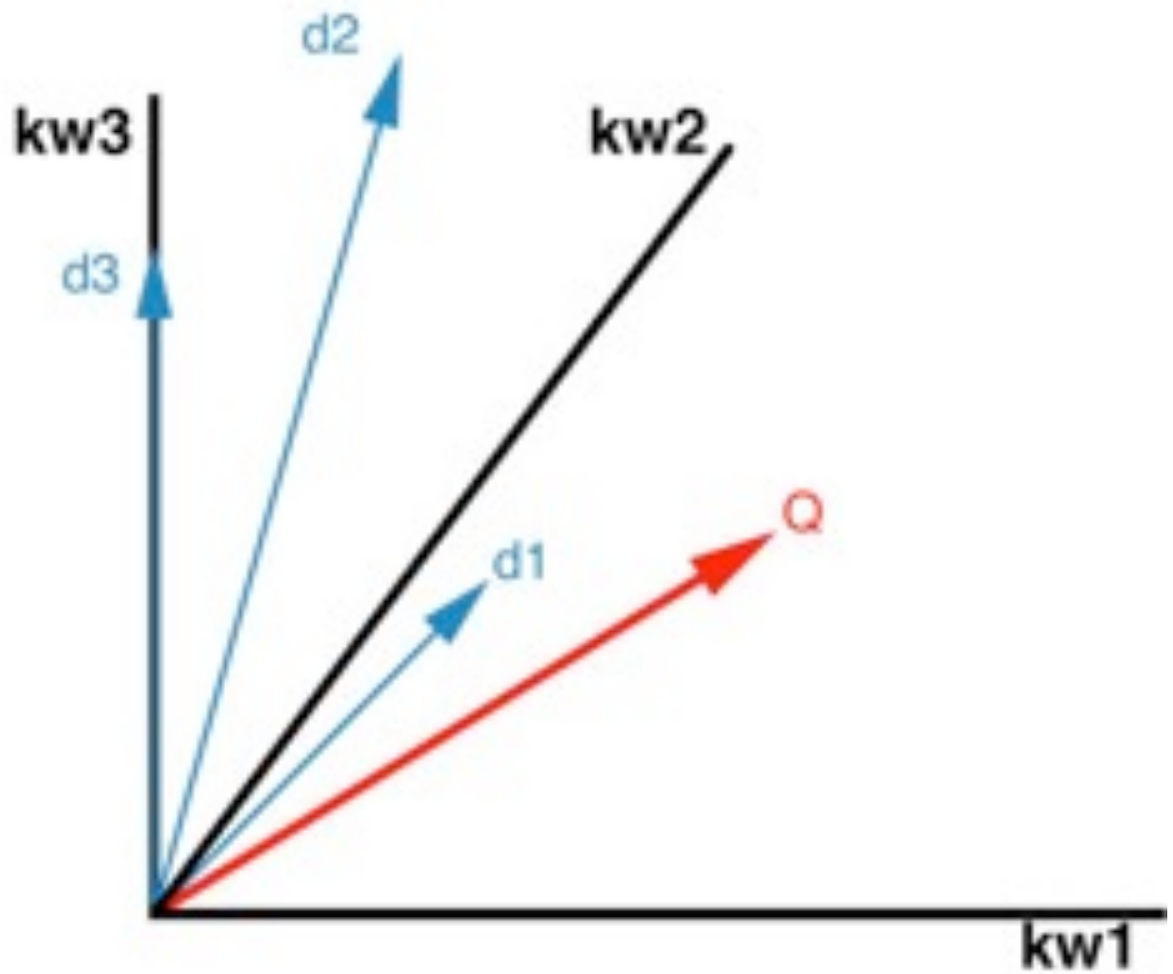
Vector Space

- 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
  - extremely fast, but expensive to build

[Belew 2000]

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

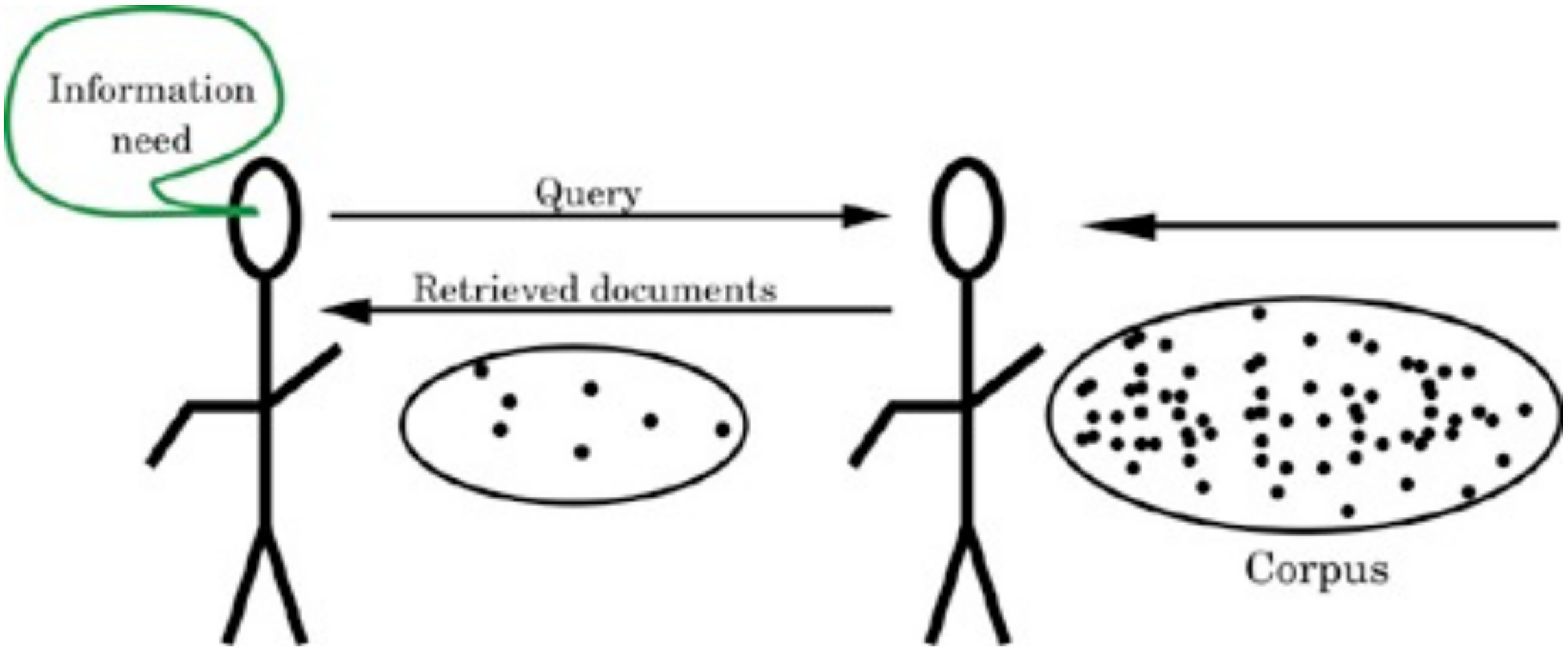


[Belew 2000]

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

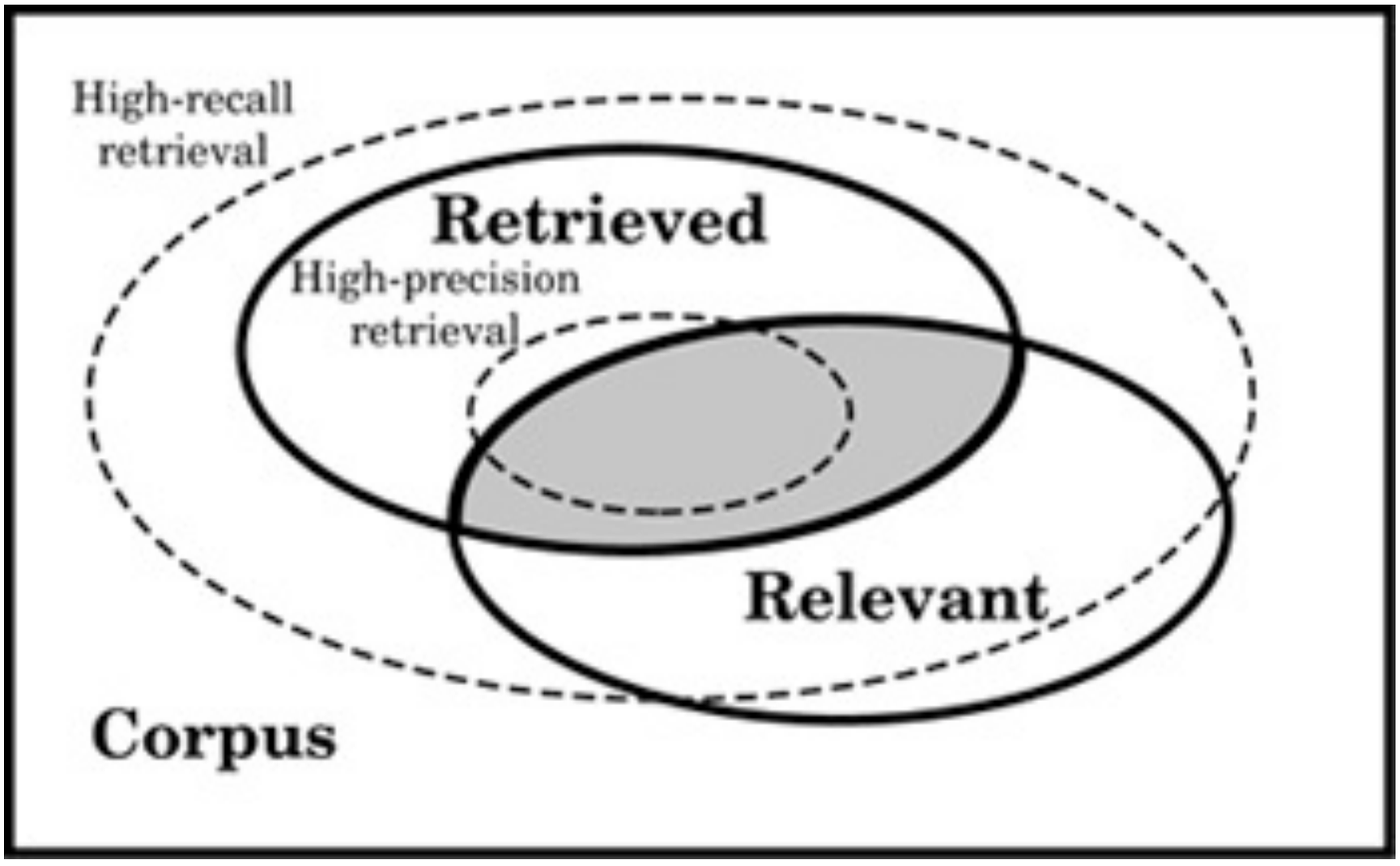


[Belew 2000]

Precision and Recall

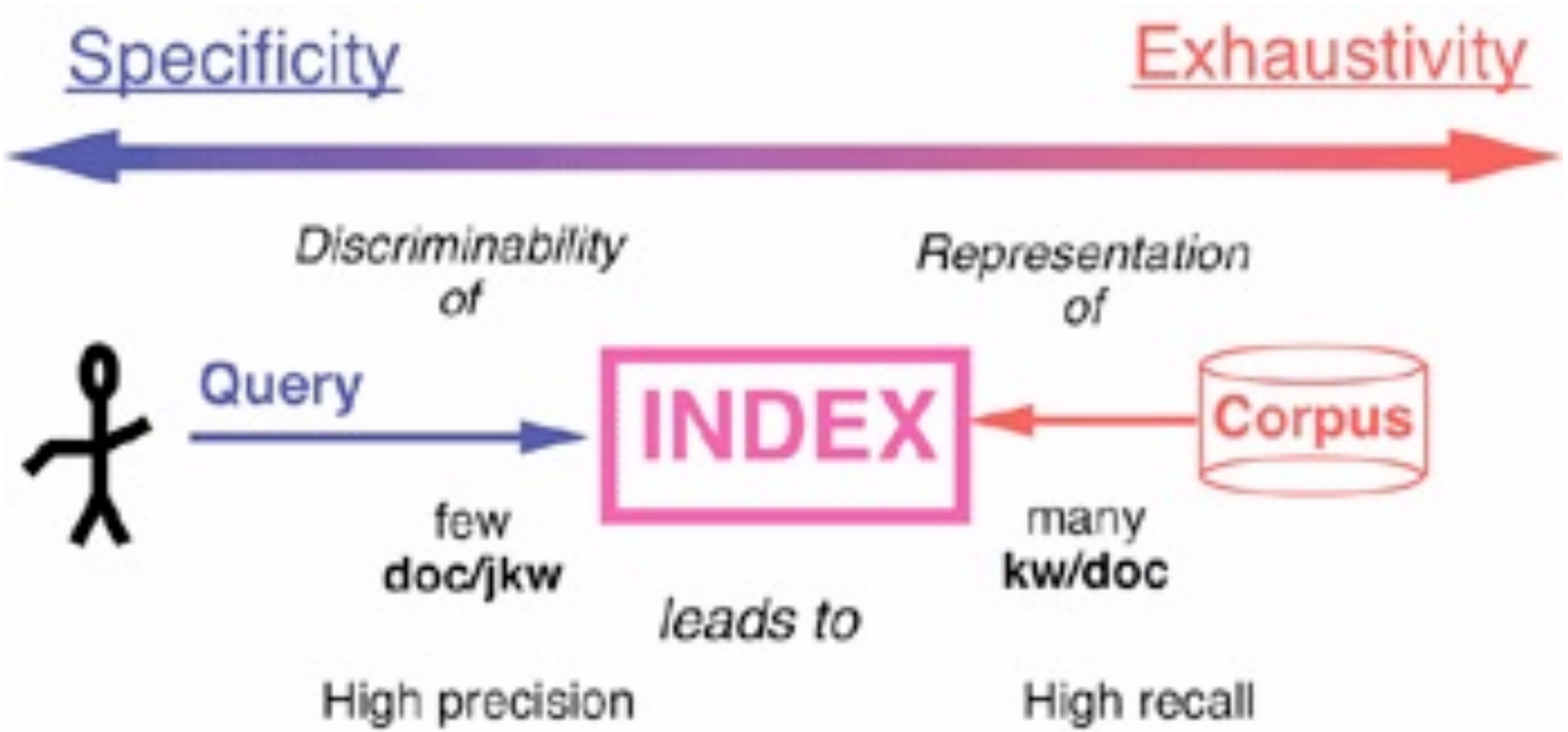
recall ≡ |*retrieved* ∩ *relevant*| / |*relevant*|

precision ≡ |*retrieved* ∩ *relevant*| / |*retrieved*|



[Belew 2000]

Specificity vs. Exhaustivity



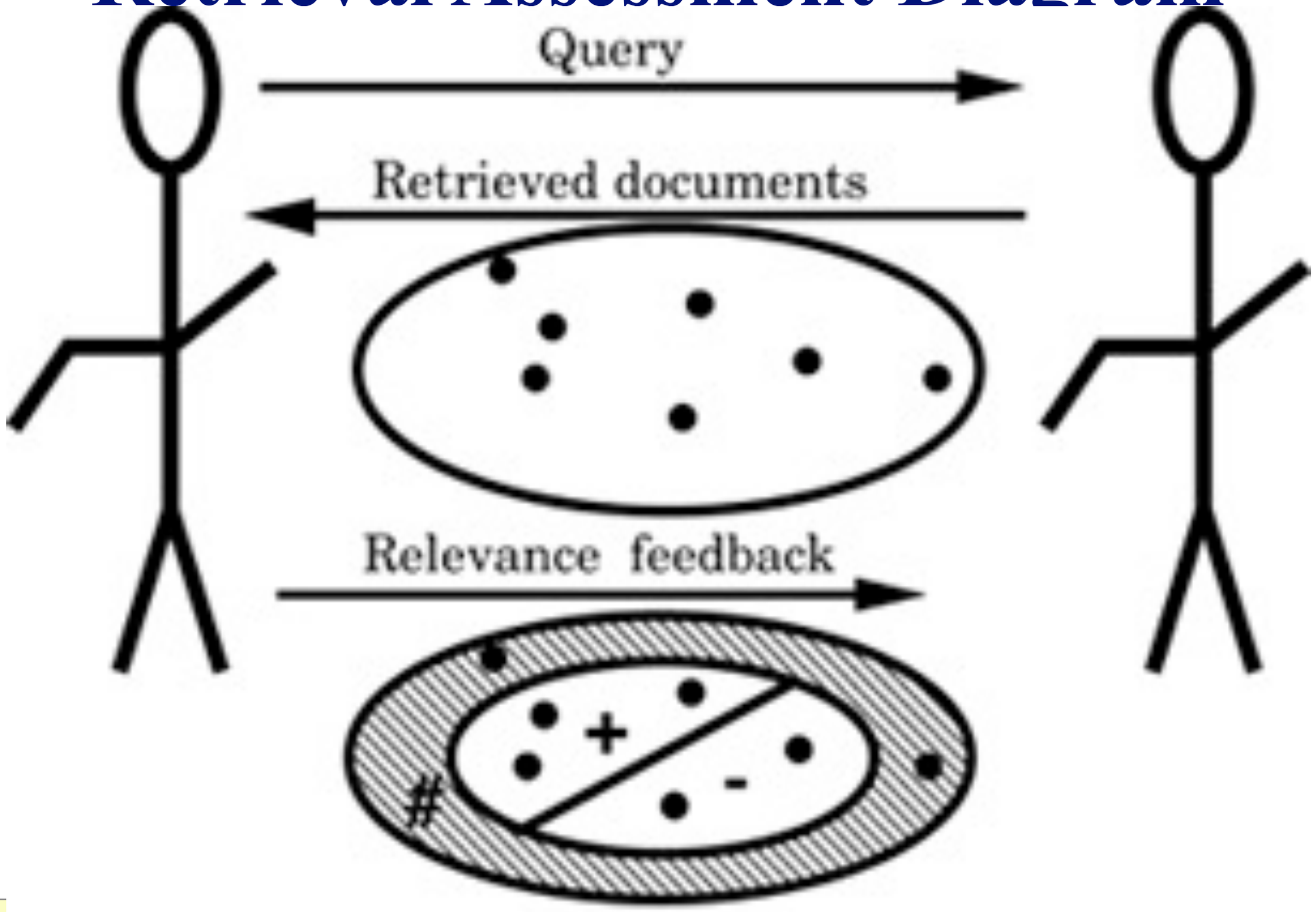
[Belew 2000]

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

[Belew 2000]

Retrieval Assessment Diagram



[Belew 2000]

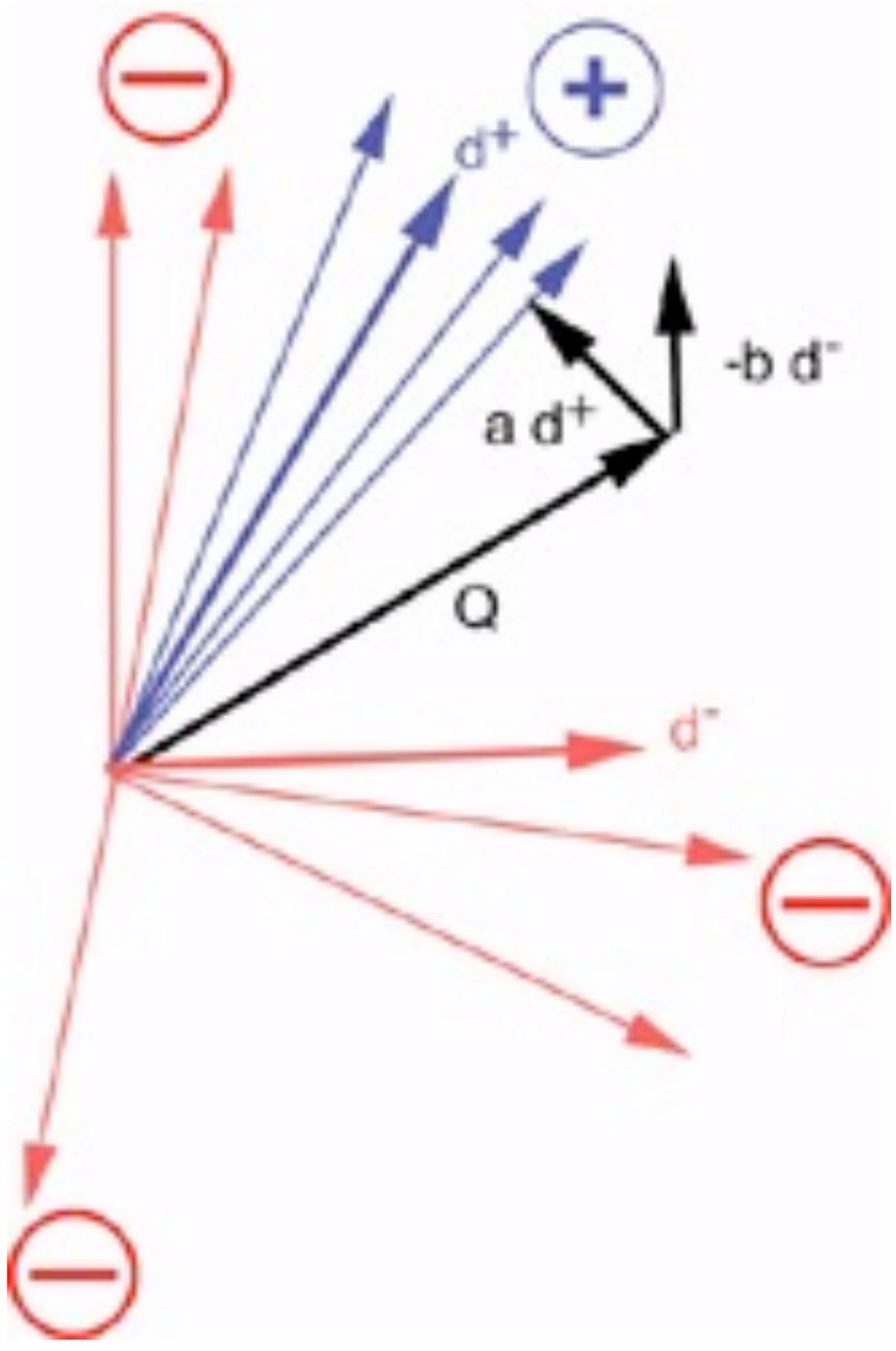
Relevance Feedback

- 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

[Belew 2000]

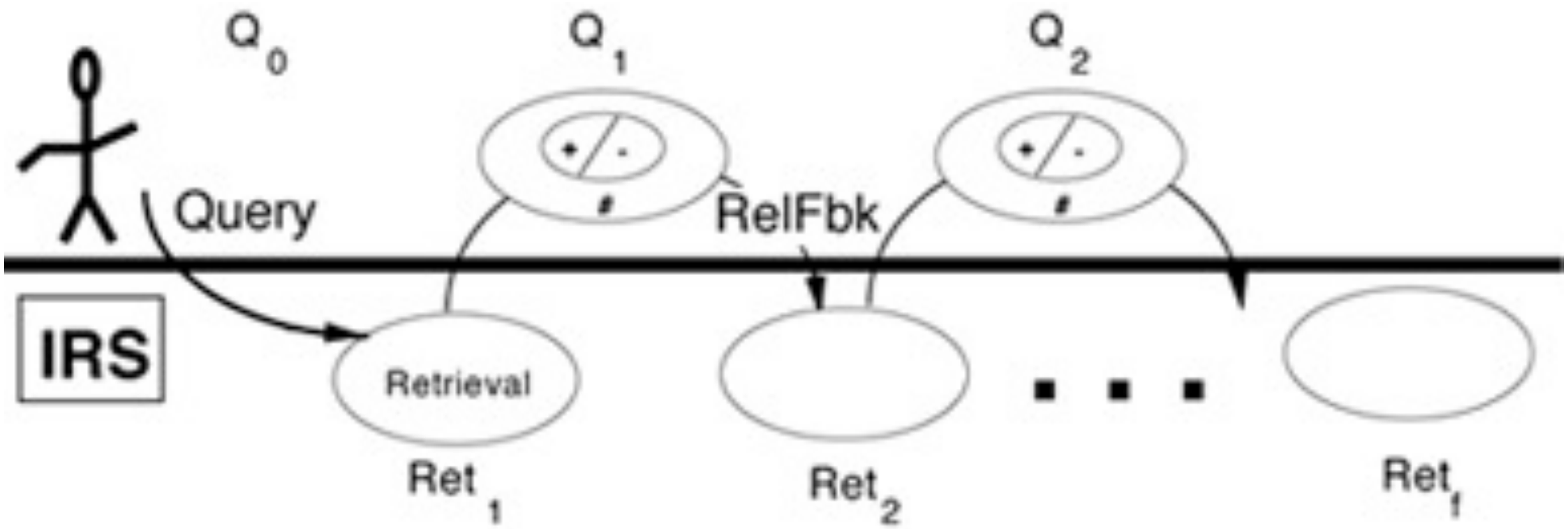
Relevance Feedback in Vector Space

- moves the query towards the cluster of positive documents
  - moving away from bad documents does not necessarily improve the results
- can also be used as a filter for a constant stream of documents
  - as in news channels or similar situations



[Belew 2000]

Query Session Example



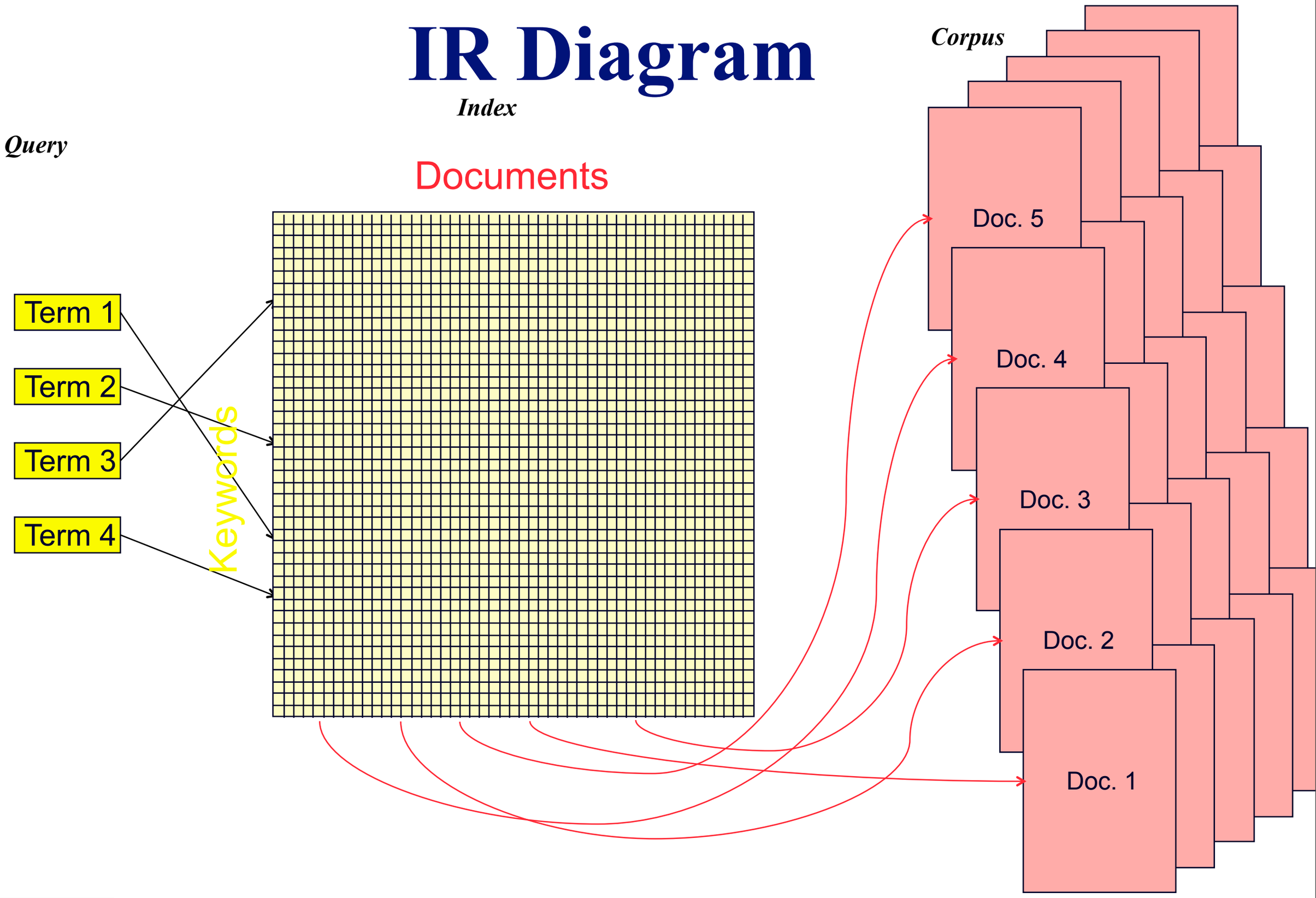
[Belew 2000]

Consensual Relevance

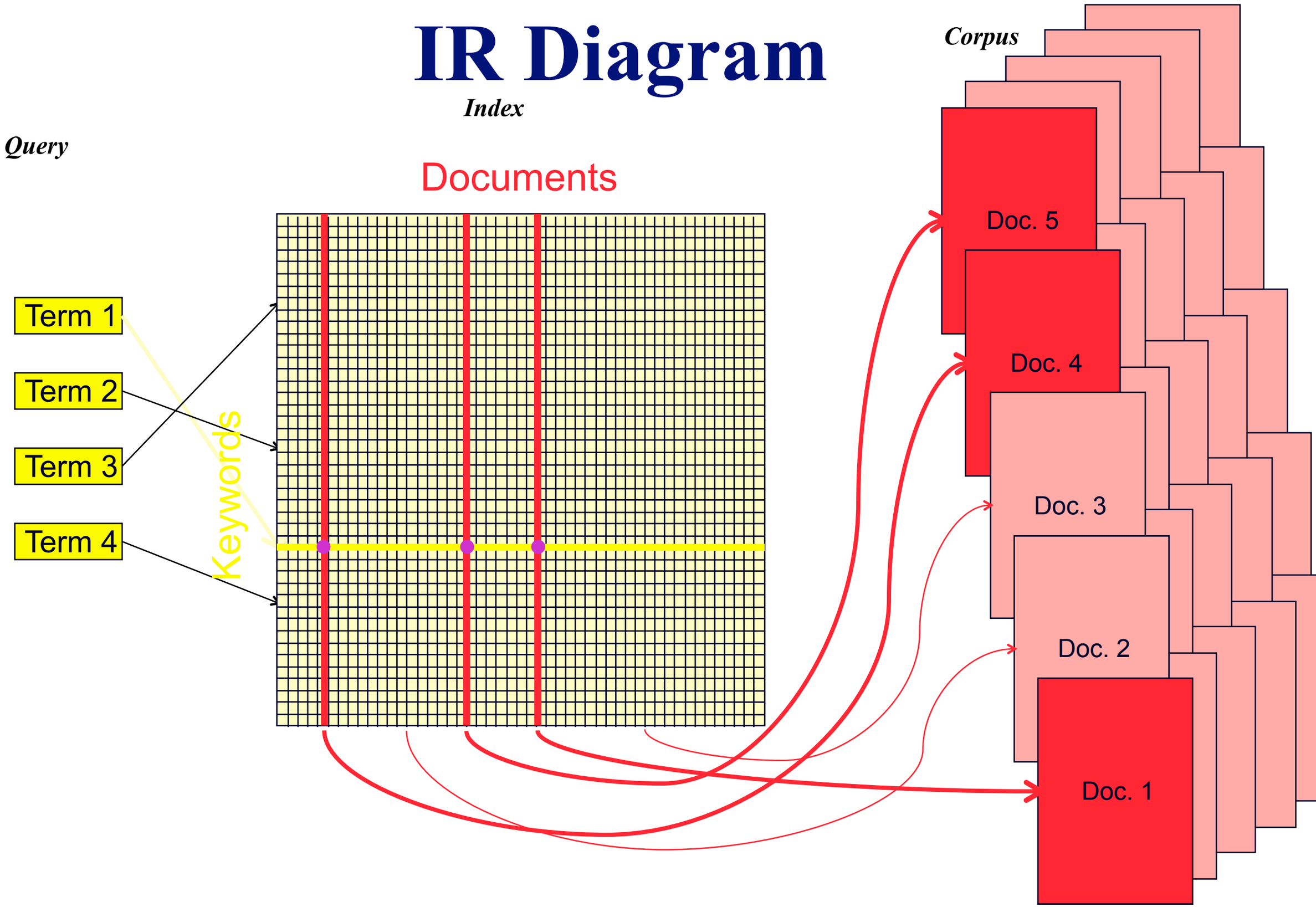
- relevance feedback from multiple users
  - identifies documents that many users found useful or interesting
  - used by some Web sites
  - related to collaborative filtering
- can also be used as an evaluation method for search engines
  - performance criteria must be carefully considered
    - precision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

[Belew 2000]

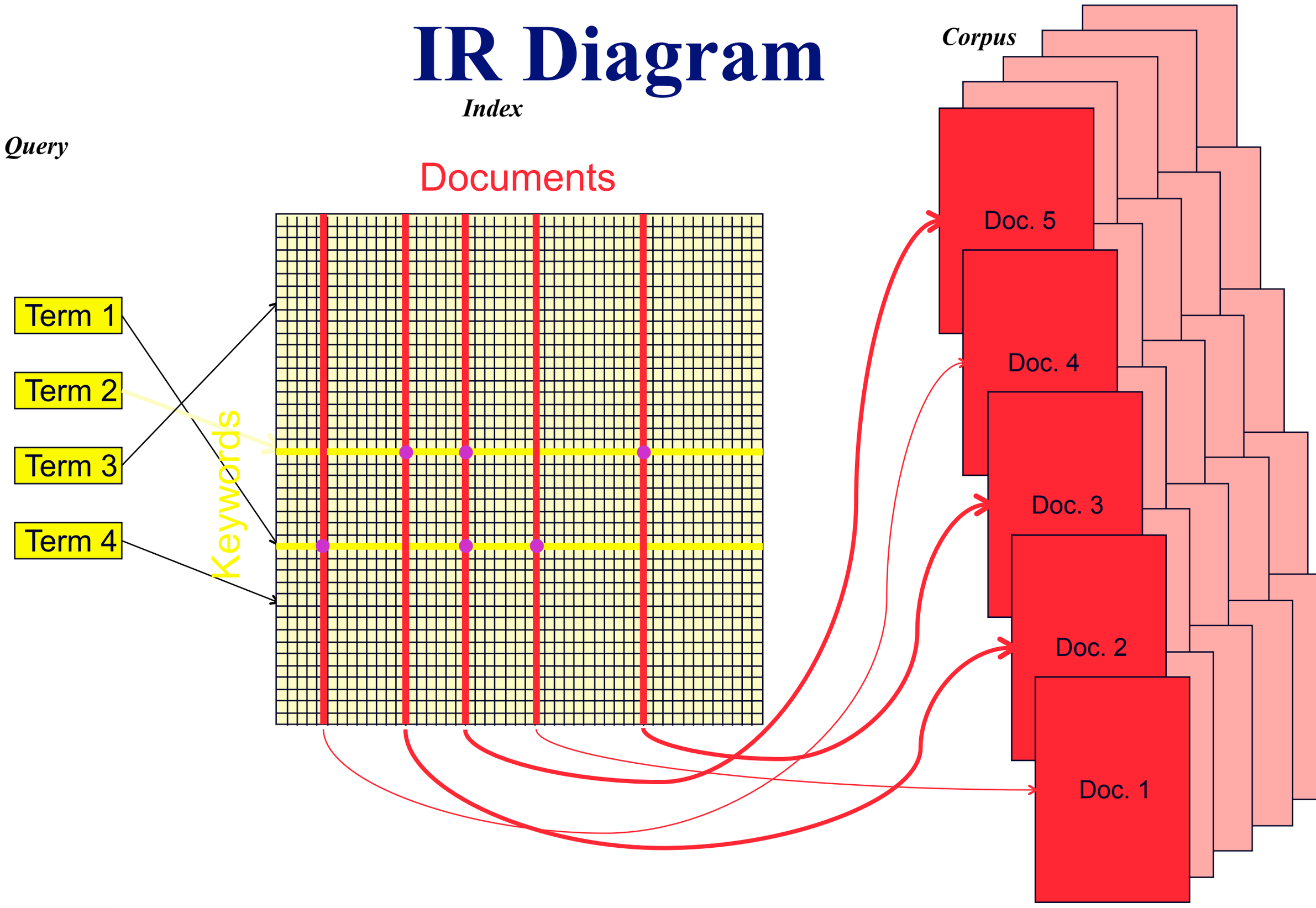
IR Diagram


Index
Query
Documents
Term 1
Term 2
Term 3
Term 4
Keywords
Corpus
Doc. 5
Doc. 4
Doc. 3
Doc. 2
Doc. 1

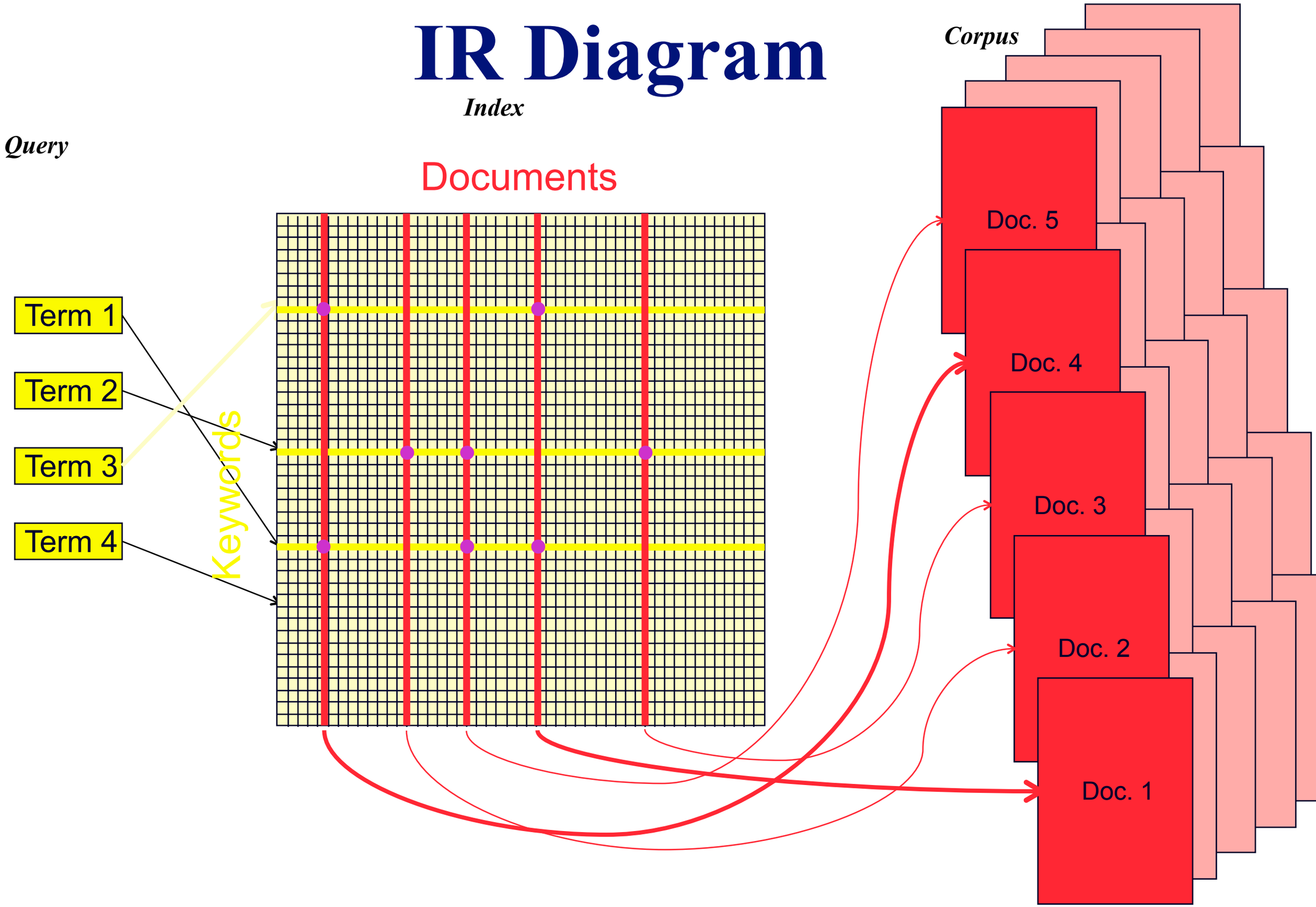
IR Diagram


Index
Query
Documents
Term 1
Term 2
Term 3
Term 4
Keywords
Corpus
Doc. 5
Doc. 4
Doc. 3
Doc. 2
Doc. 1

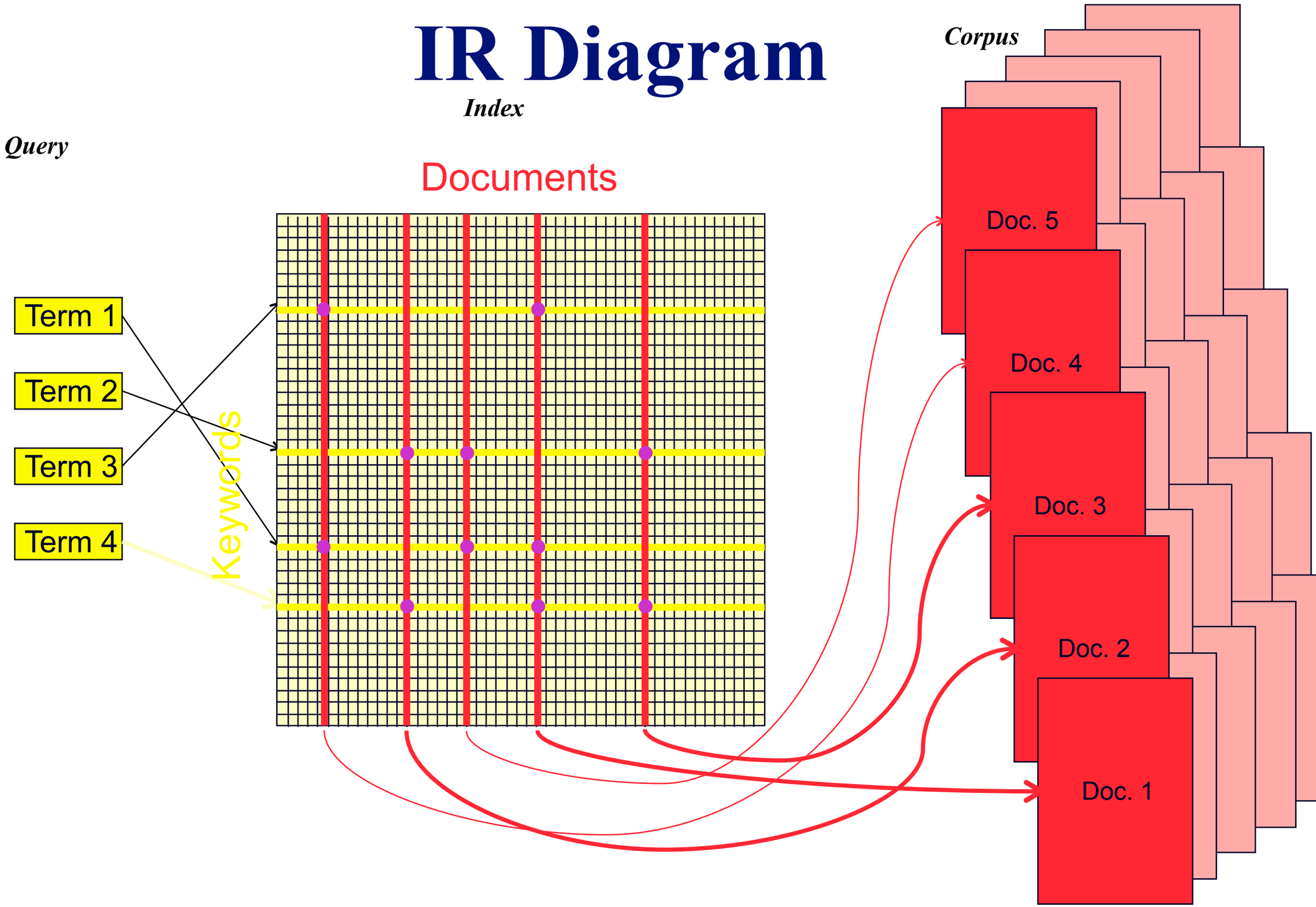
IR Diagram


Index
Query
Documents
Term 1
Term 2
Term 3
Term 4
Keywords
Corpus
Doc. 5
Doc. 4
Doc. 3
Doc. 2
Doc. 1

IR Diagram


Index
Query
Documents
Term 1
Term 2
Term 3
Term 4
Keywords
Corpus
Doc. 5
Doc. 4
Doc. 3
Doc. 2
Doc. 1

IR Diagram


Index
Query
Documents
Term 1
Term 2
Term 3
Term 4
Keywords
Corpus
Doc. 5
Doc. 4
Doc. 3
Doc. 2
Doc. 1

Knowledge Retrieval

- Context
- Usage
  - exploratory search
  - faceted search

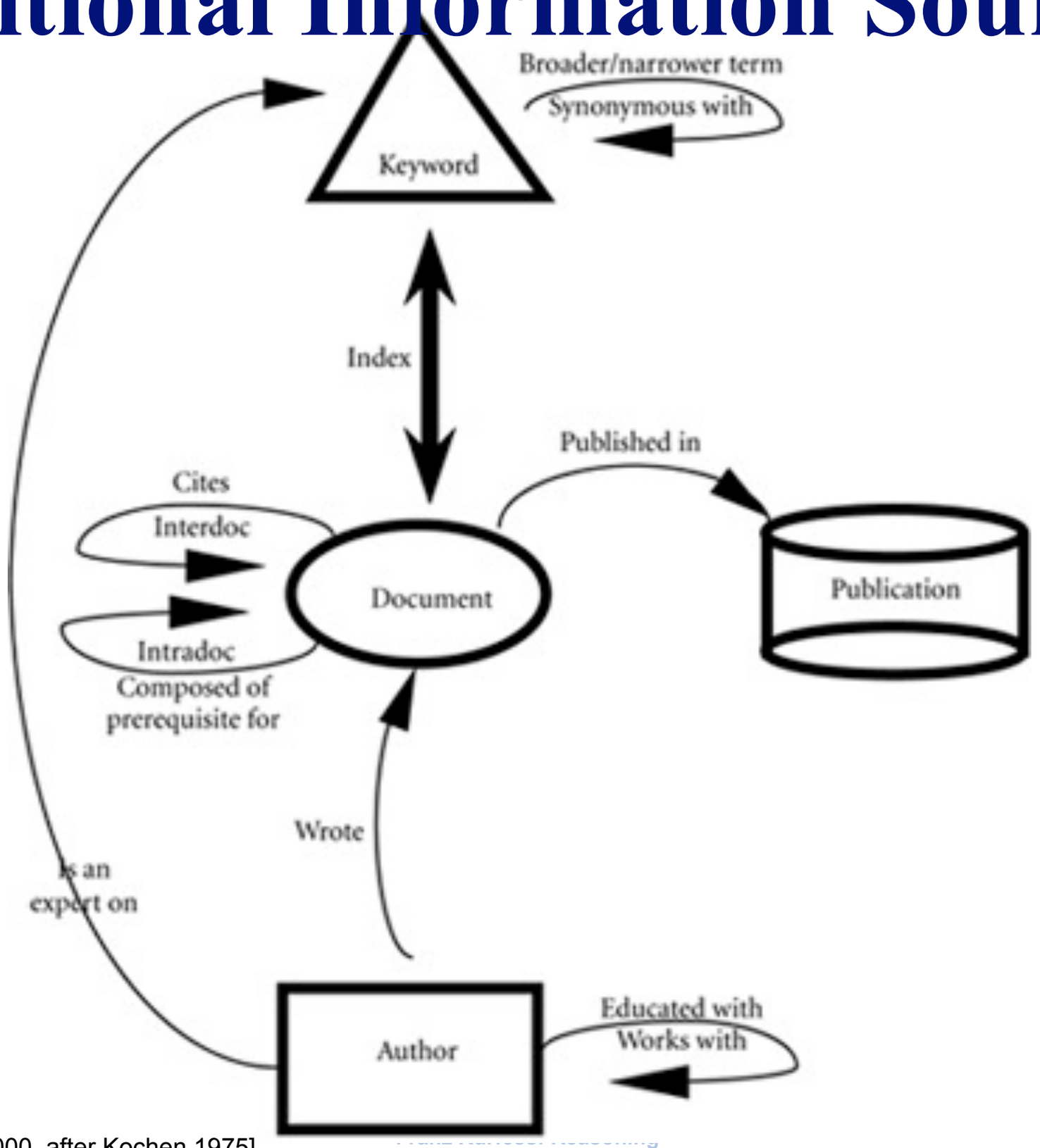
Context in Knowledge Retrieval

- 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]

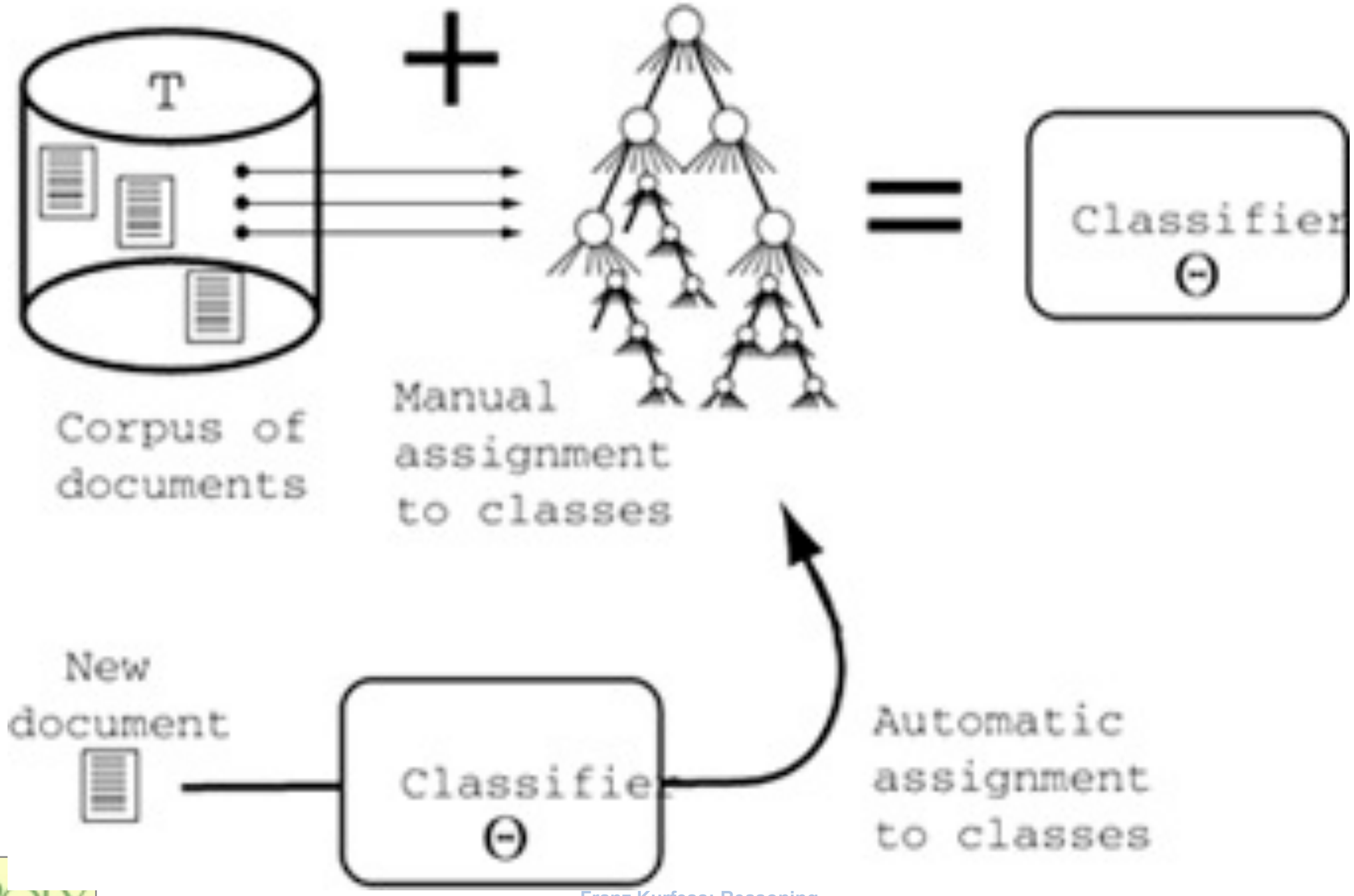
Hypertext

- inter-document links provide explicit relationships between documents
  - can be used to determine the relevance of a document for a query
  - example:
    Google's PageRank algorithm
- 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

[Pratt, Hearst, Fagan 2000]

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

[Pratt, Hearst, Fagan 2000]

Results

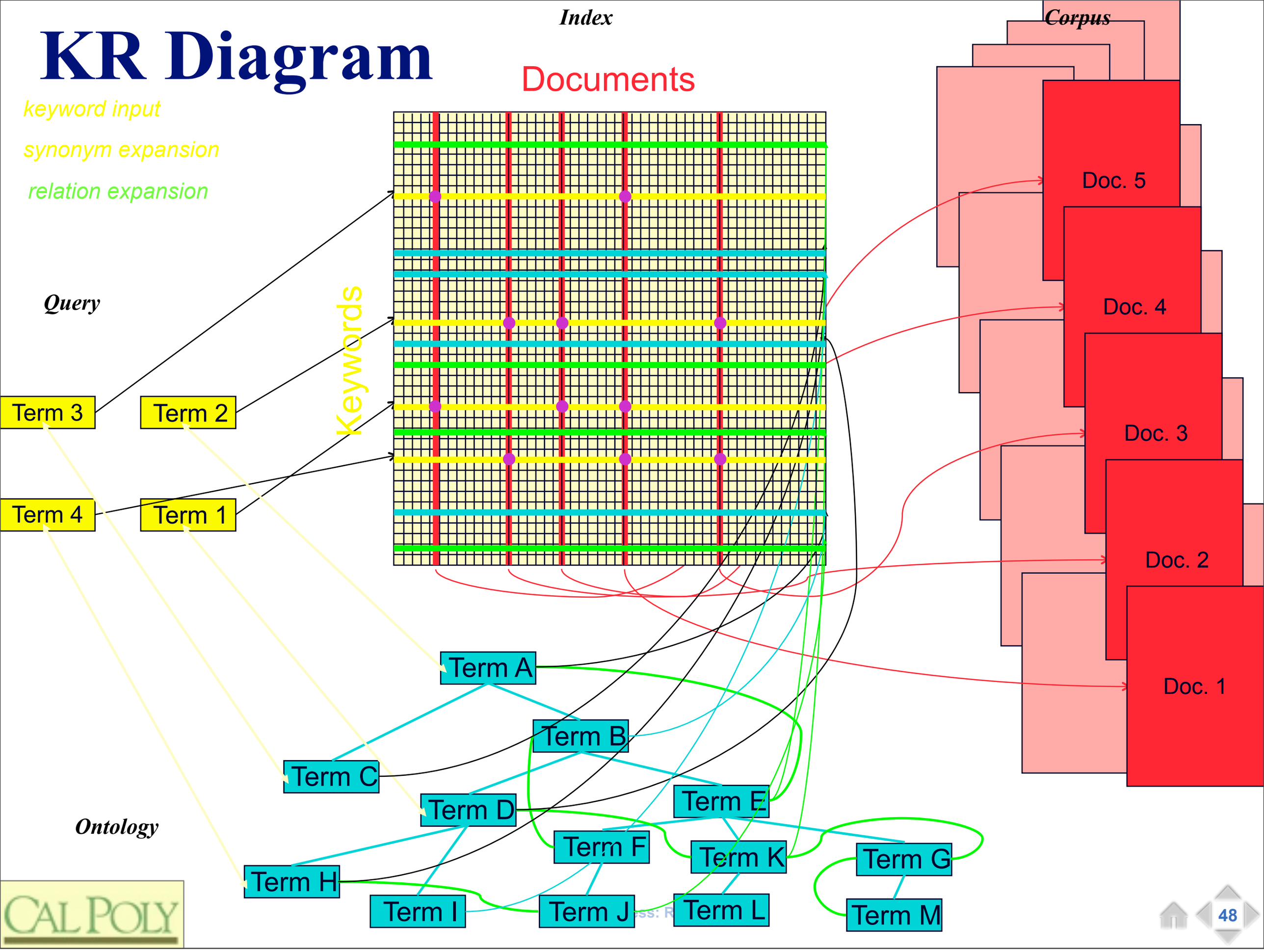
- retrieved documents are grouped into hierarchically arranged categories meaningful for the user
  - the categories are related to the query
  - the categories are related to each other
  - all categories have similar size
    - not always achievable due to the distribution of documents
- reduced search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

[Pratt, Hearst, Fagan 2000]

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


Index
Documents
keyword input
synonym expansion
relation expansion
Query
Keywords
Corpus
Term 3
Term 2
Term 4
Term 1
Doc. 5
Doc. 4
Doc. 3
Doc. 2
Doc. 1
Term A
Term B
Term C
Term D
Term E
Term F
Term K
Term G
Term H
Term I
Term J
Term L
Term M
Ontology

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

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

Name	Album by Year	Artist	Yea
Outrage	Now Hear This: Discover New Music – Rock	Capital Lights	200
My King (Falling Album Version)	Now Hear This: Discover New Music – Rock	Seven Stories Up	200
Traveler's Song	Now Hear This: Discover New Music – Rock	Future Of Forestry	200
Hold On	Now Hear This: Discover New Music – Rock	Abandon	200
I Can't Stand To Fall	Now Hear This: Discover New Music – Rock	Philmont	200
I'll Love You So	Now Hear This: Discover New Music – Rock	Above The Golden S...	200
Streetlight	Now Hear This: Discover New Music – Rock	Danyew	200
Taste And See	Now Hear This: Discover New Music – Pop	Jason Allen Rich	200
Life Light Up	Now Hear This: Discover New Music – Pop	Christy Nockels	200
Lord Of All	Now Hear This: Discover New Music – Pop	Kristian Stanfill	200
Chasing The Daylight	Now Hear This: Discover New Music – Pop	Phillip LaRue	200
Tell Me	Now Hear This: Discover New Music – Pop	Josh Wilson	200
Come Save	Now Hear This: Discover New Music – Pop	Sarah Reeves	200
Walk Tall (feat. Paul Simon)	Family Time	Ziggy Marley	200
Invisible Cities	Invisible Cities	Nomo	200
Seasons	Seasons	The Veer Union	200
Sugarfoot	Sugarfoot	Black Joe Lewis & T...	200
In My Time Of Dyin'	Time To Grow	The Lovell Sisters	200
Ninth Place	Lay Your Burden Down	Buckwheat Zydeco	200
Mama	Mama	Holly Williams	200
Ride The Nuclear Wave	Be On The Lookout!	The Oranges Band	200
Shrapnel	Be On The Lookout!	American Steel	200
Deconstruct/Rebuild	Be On The Lookout!	Small Brown Bike	200
Face It	Be On The Lookout!	The Reputation	200
Ear Nose And Throat	Be On The Lookout!	Troubled Hubble	200
Eating Toothpaste	Be On The Lookout!	Bratmobile	200
Under The Hedge	Be On The Lookout!	Ted Leo/Pharmacists	200
Sorry For Freaking Out On The Phone Last Night	Be On The Lookout!	The Mr. T Experience	200
Are You Gonna Move It For Me?	Be On The Lookout!	The Donnas	200
Fortune Cookie	The Further Adventures of Los Straitjackets	Los Straitjackets	200
I Wanna Know Why	The Leaves are Right to Tremble	Justin Rosolino	200
The Show Is On The Road	The Show Is On The Road	Paleface	
Oh My Soul	Little Red	Susan Marshall	200
Excursion Around The Bay	Great Big Sea	Great Big Sea	200
Yours For the Taking	Smoking Kills	The Disciplines	200
Songs in the Night	Songs in the Night	Samantha Crain and...	
Little Bit of Heaven	Cheval Sombre	Cheval Sombre	200
She Loves Everybody	She Loves Everybody	Chester French	200
Longing For	Longing For	Ballas Hough Band	200
Perfect	Emo Is Awesome Emo Is Evil 2	Sounds Like Violence	200

112
4
re
Cart
brary
s
lusic
Entertainment
Learning
ructions
o Music
y
os
dded
dded Songs
ayed
st Played
od Selection

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

Non-Textual Retrieval

Image Search
Music Search

Image Search

- contextual
  - meta-data
  - text in the same or close-by documents
    - e.g. on the same Web page, or in the same directory
- content-based
  - analysis and comparison of images
  - feature extraction

Contextual Image Search

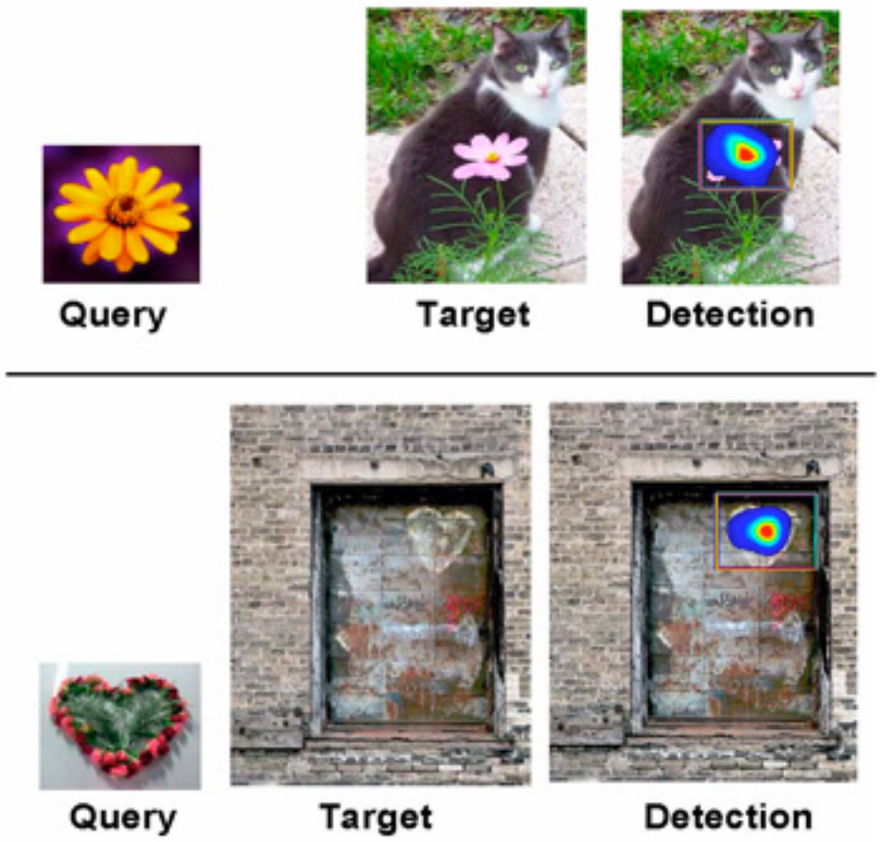
- images are indexed via keywords
  - captions of images
  - metadata (tags)
  - surrounding text
- relies on the assumption that the indexed text is correlated to the image and a good description of its content
- basis for most current image search engines

Content-Based Image Search

- images are compared against query images
  - no text elements as proxies
    - or at least not in a “pure” content-based image search
  - relies on feature extraction or object recognition
    - direct comparison of pictures on a pixel-by-pixel basis is impractical
      - only yields identical pictures, not similar ones
      - computationally very challenging
        - especially if the query image is a subset of a target image
- allows the use of a picture as a “template” to find related pictures

Example: UC Santa Cruz Image Search

- feature extraction and object recognition from images and video



Using a single image as a template, computer software can find similar images in a large database of photos, as shown in these examples. Images courtesy of P. Milanfar.
http://www.physorg.com/newman/gfx/news/hires/newsearchtec.jpg

http://www.physorg.com/news177095786.html

Music Search: Shazam

- identifies musical pieces through “finger prints”
  - emphasis on popular music

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