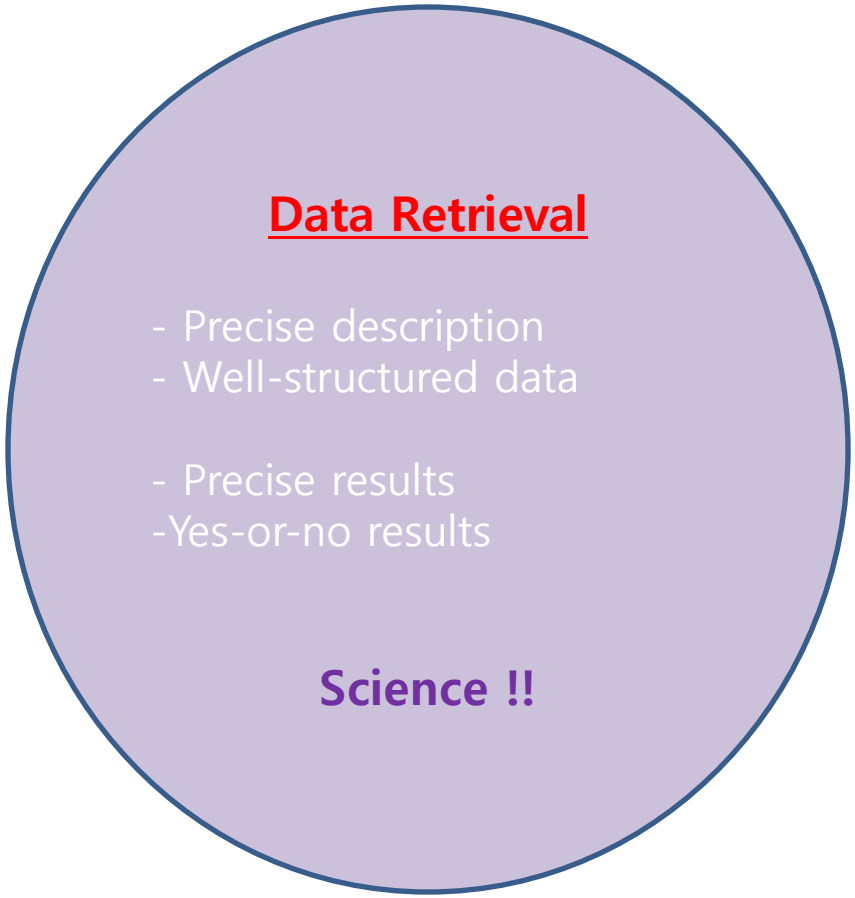
Information Retrieval

- Overview

Contents

Data vs. Information Retrieval



Definitions

Collection : is a set of documents

Volume : is a subset of documents

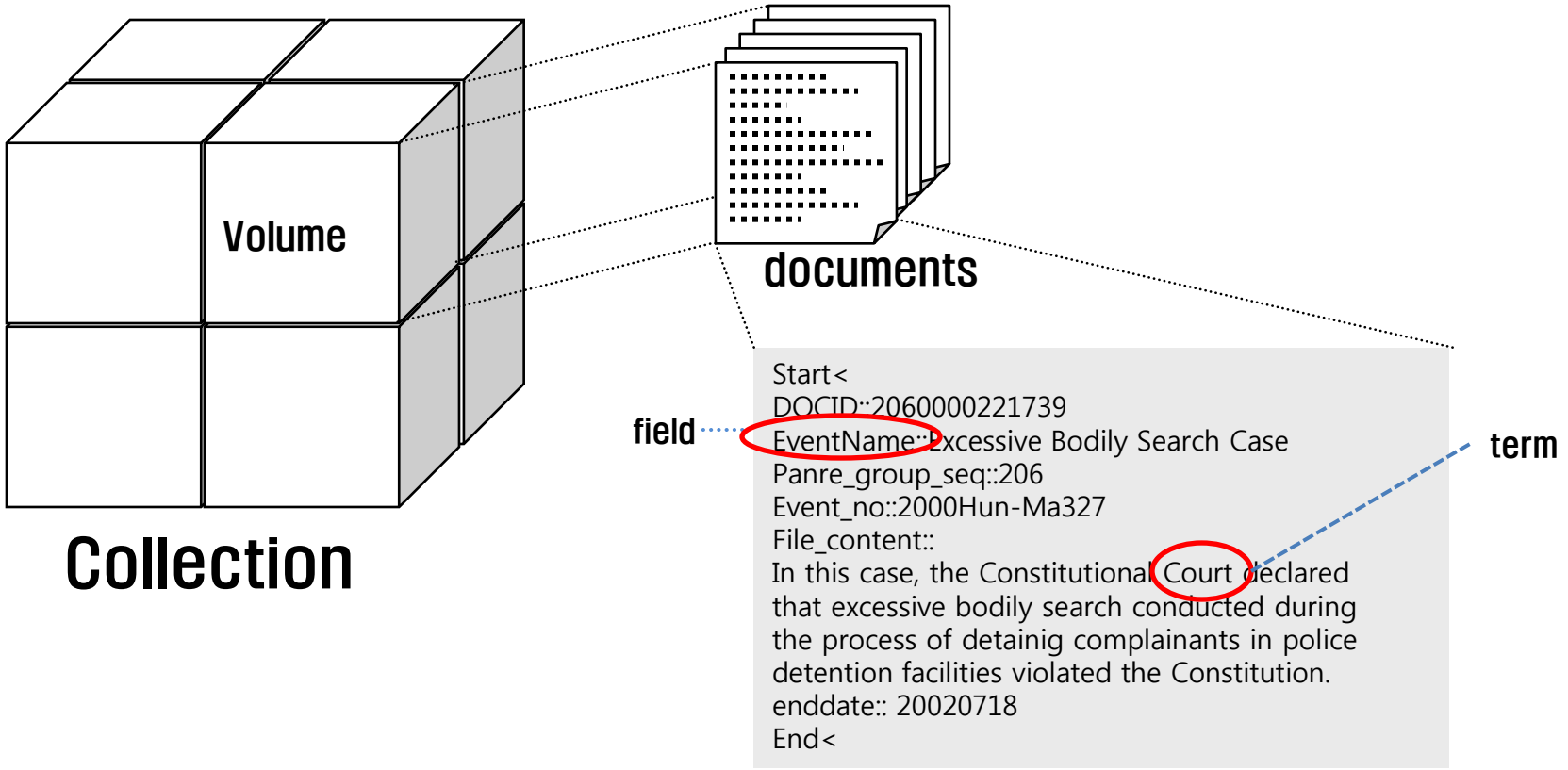
Document : is a sequence of terms

Term : is a semantic unit
ex) word, phrase, root of word

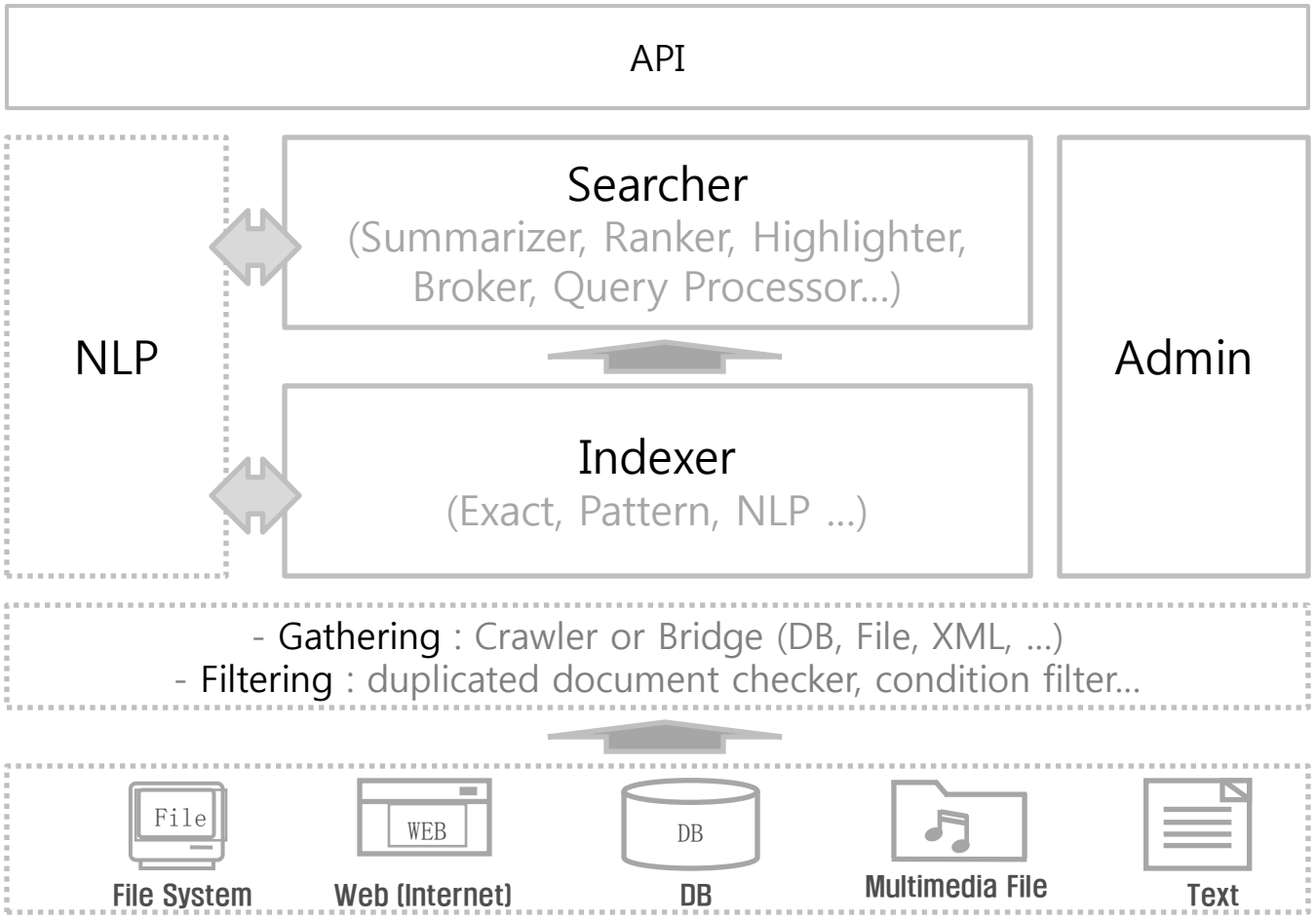
Query : is a request for documents pertaining to some topic

Information Retrieval (IR) System : attempt to find relevant documents to respond to a user's request

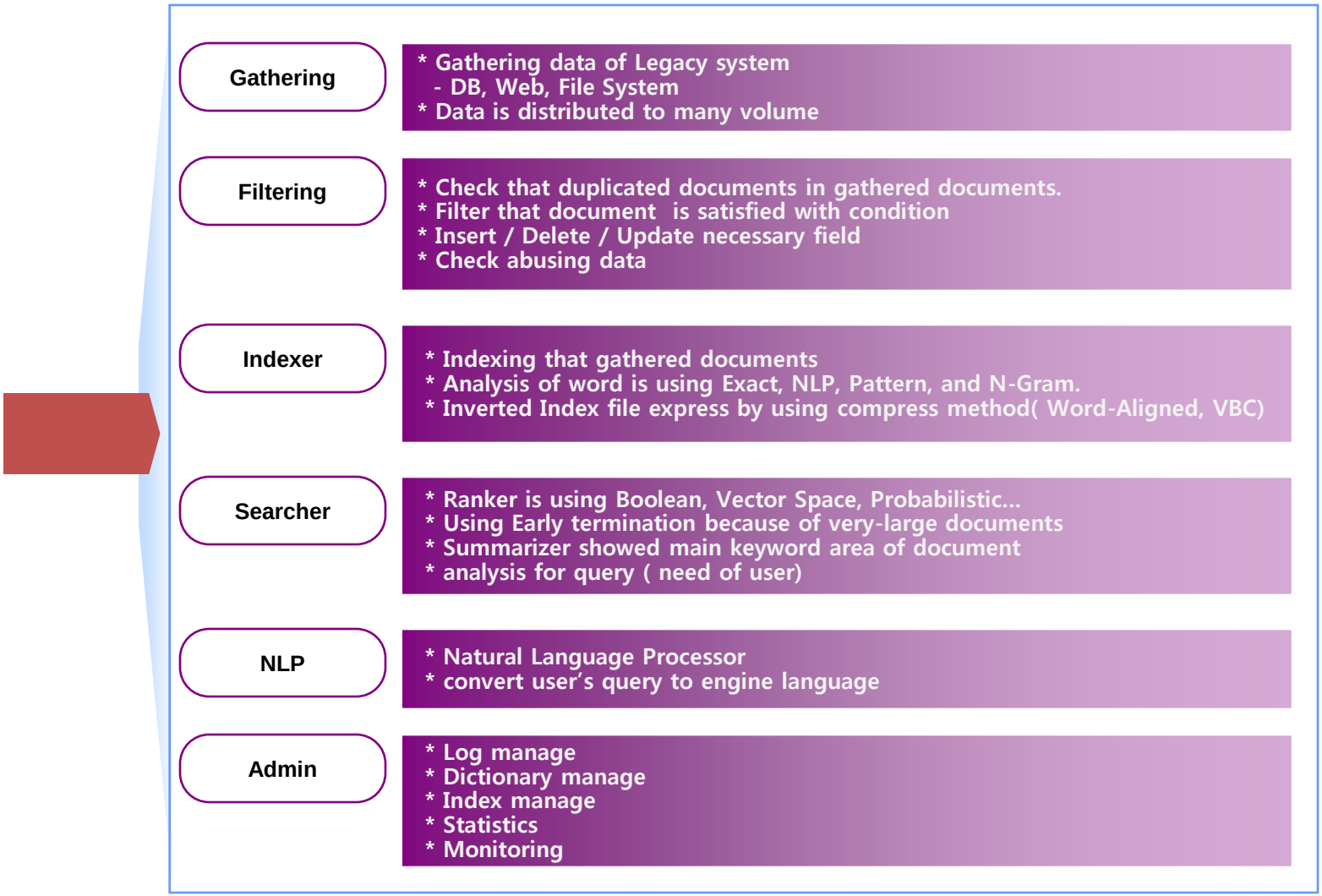
Data Class


Volume
Collection
documents
field
term
Start<
DOCID::2060000221739
EventName::Excessive Bodily Search Case
Panre_group_seq::206
Event_no::2000Hun-Ma327
File_content::
In this case, the Constitutional Court declared that excessive bodily search conducted during the process of detainig complainants in police detention facilities violated the Constitution.
enddate:: 20020718
End<

Concept

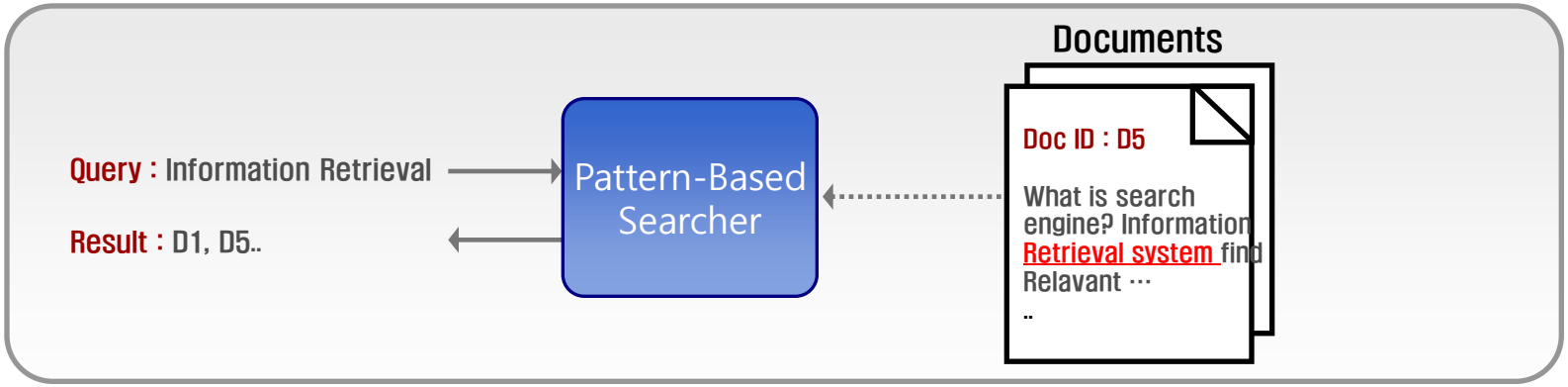

API
NLP
Searcher
(Summarizer, Ranker, Highlighter,
Broker, Query Processor...)
Indexer
(Exact, Pattern, NLP ...)
Admin
- Gathering : Crawler or Bridge (DB, File, XML, ...)
- Filtering : duplicated document checker, condition filter...
File
WEB
DB
File System
Web (Internet)
DB
Multimedia File
Text

The role of each component

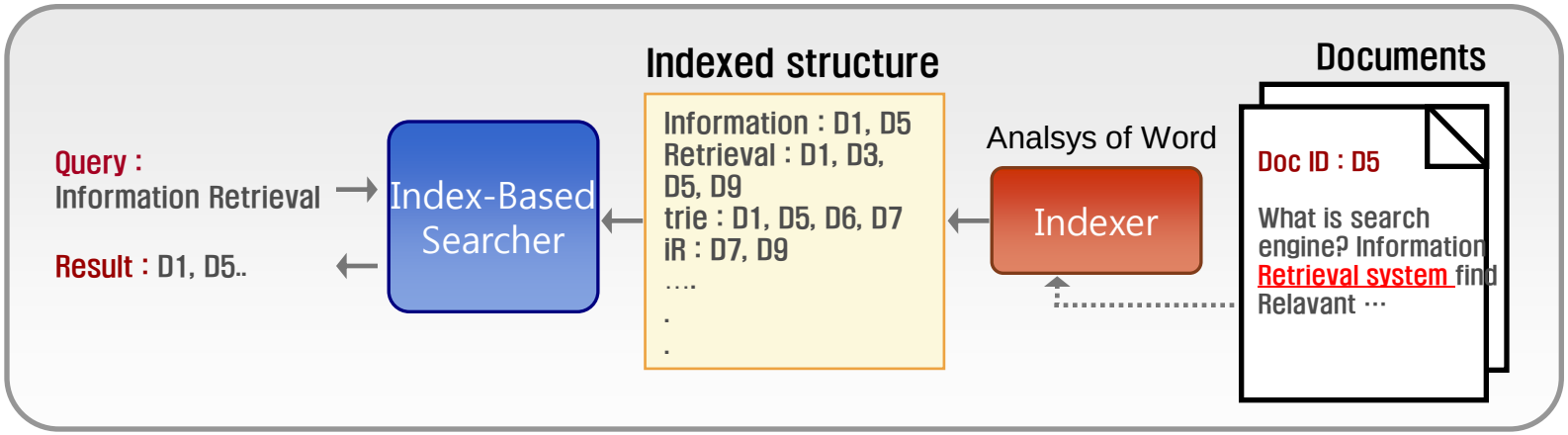


Basic Method for searching

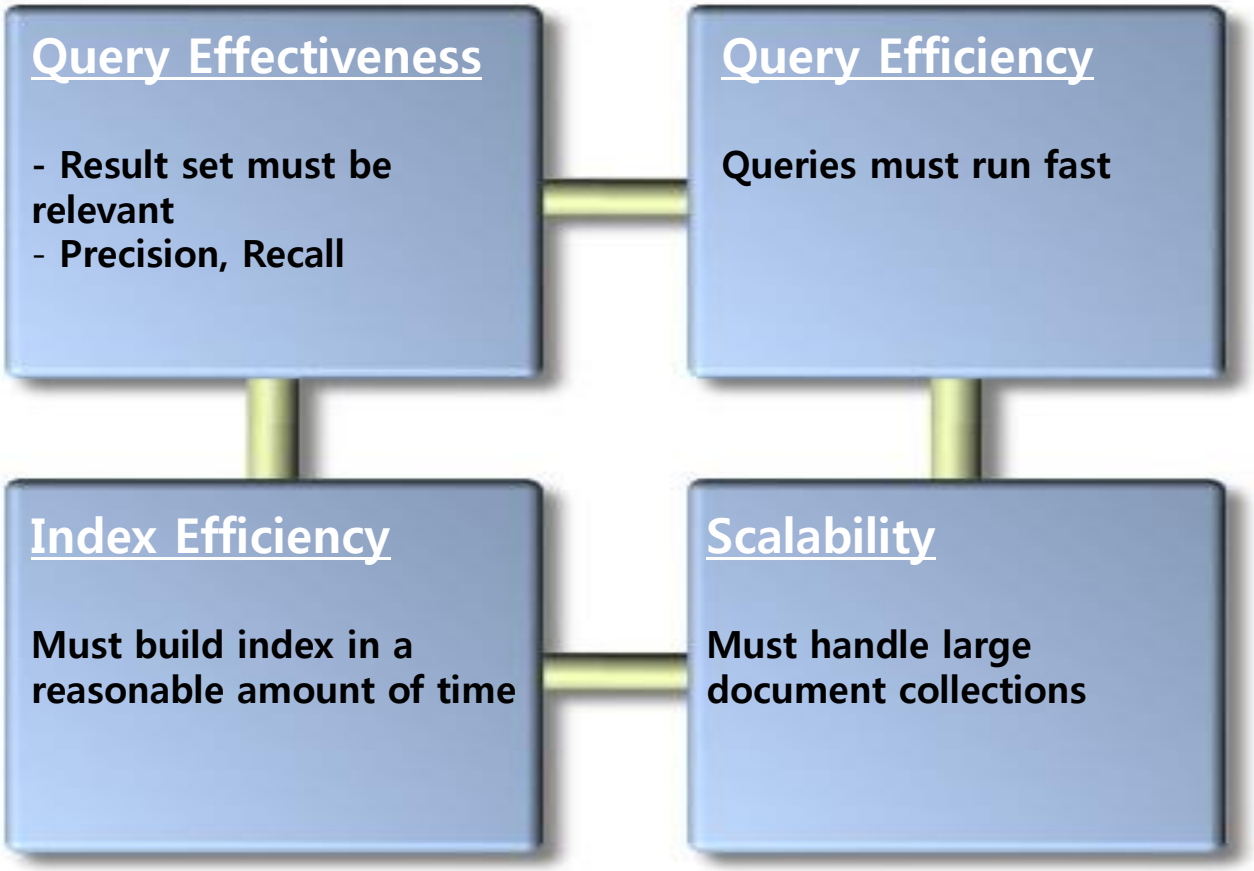
- **Patten-Based Matching Method : searching sentences of document – small collection**



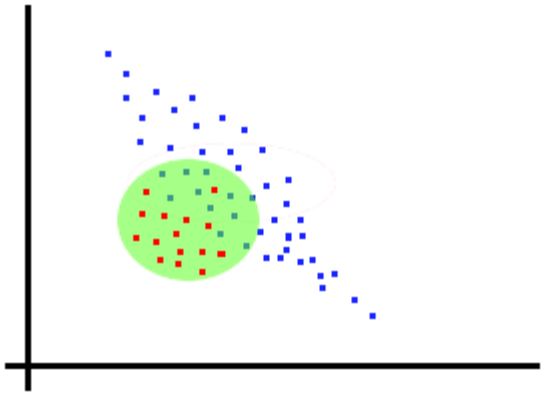
- **Index-Based Method : inverted index structure**



Requirements

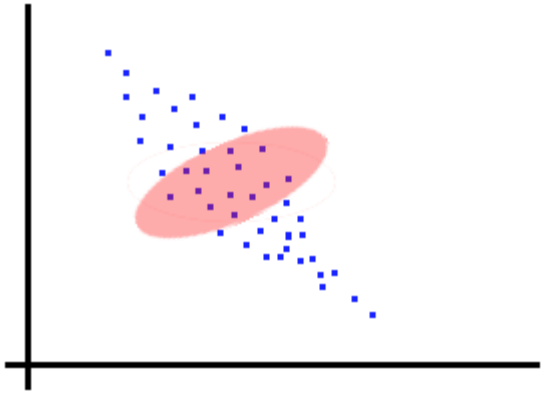


Query Effectiveness



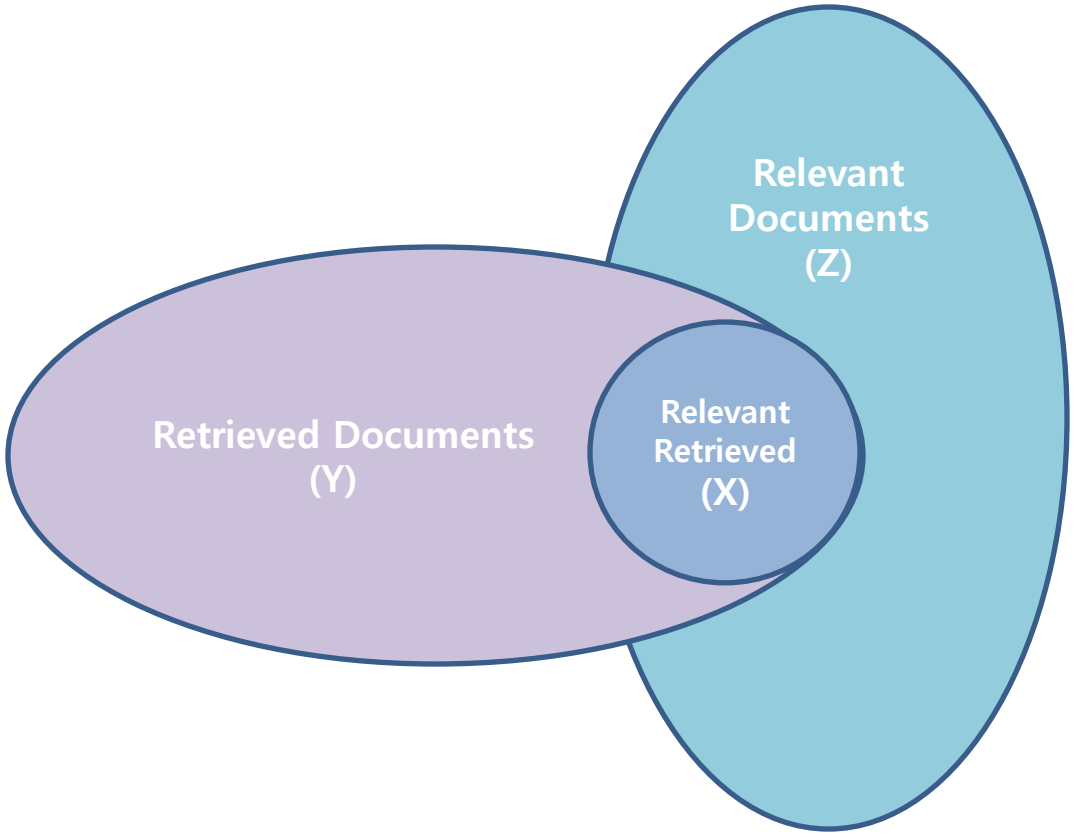
- Relevant documents(X)
- Retrieved but not Relevant documents(y - !X)

Precision : x / y



- Relevant documents(z)
- Retrieved documents(x)

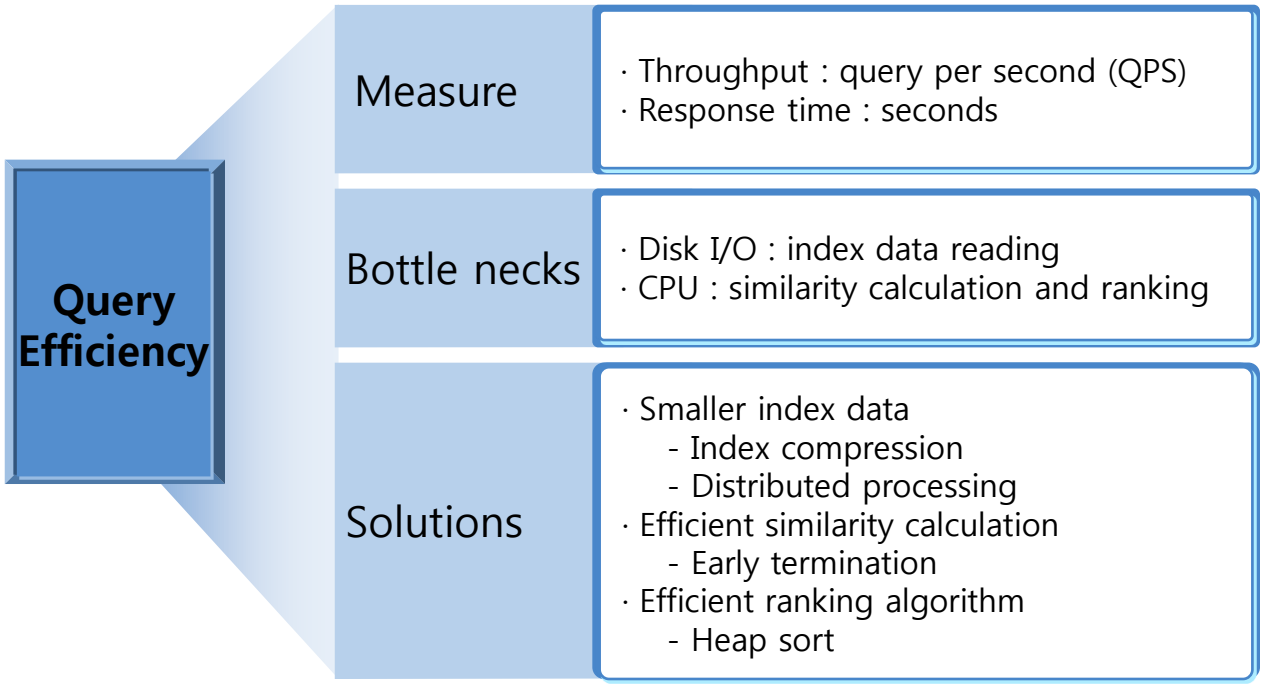
Recall : x / z



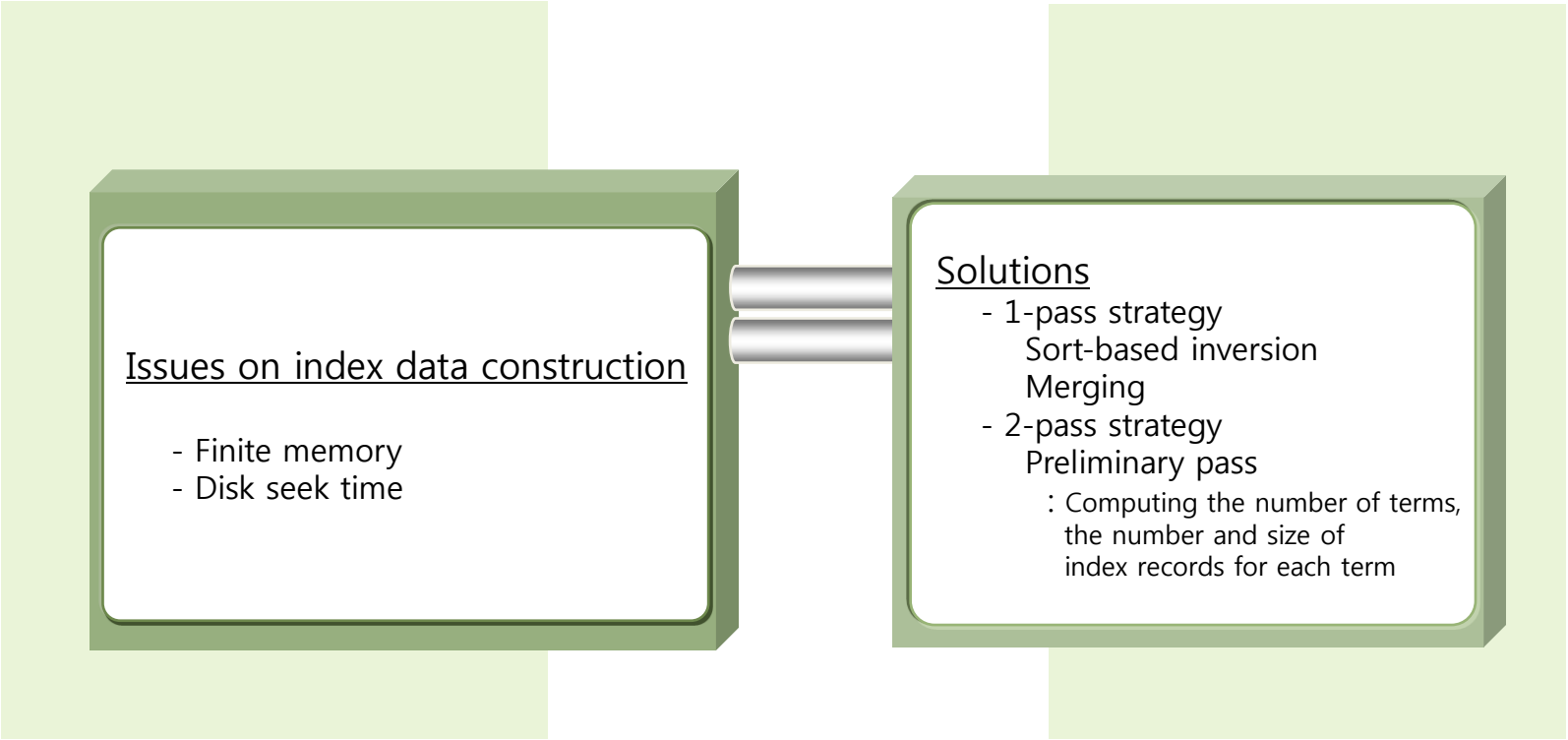
Relationship of Precision and Recall (generally)

- Recall ↑ → Precision ↓
  showed relevant documents But, showed non-relevant documents also.
- Recall ↓ → Precision ↑
  may not showed relevant documents

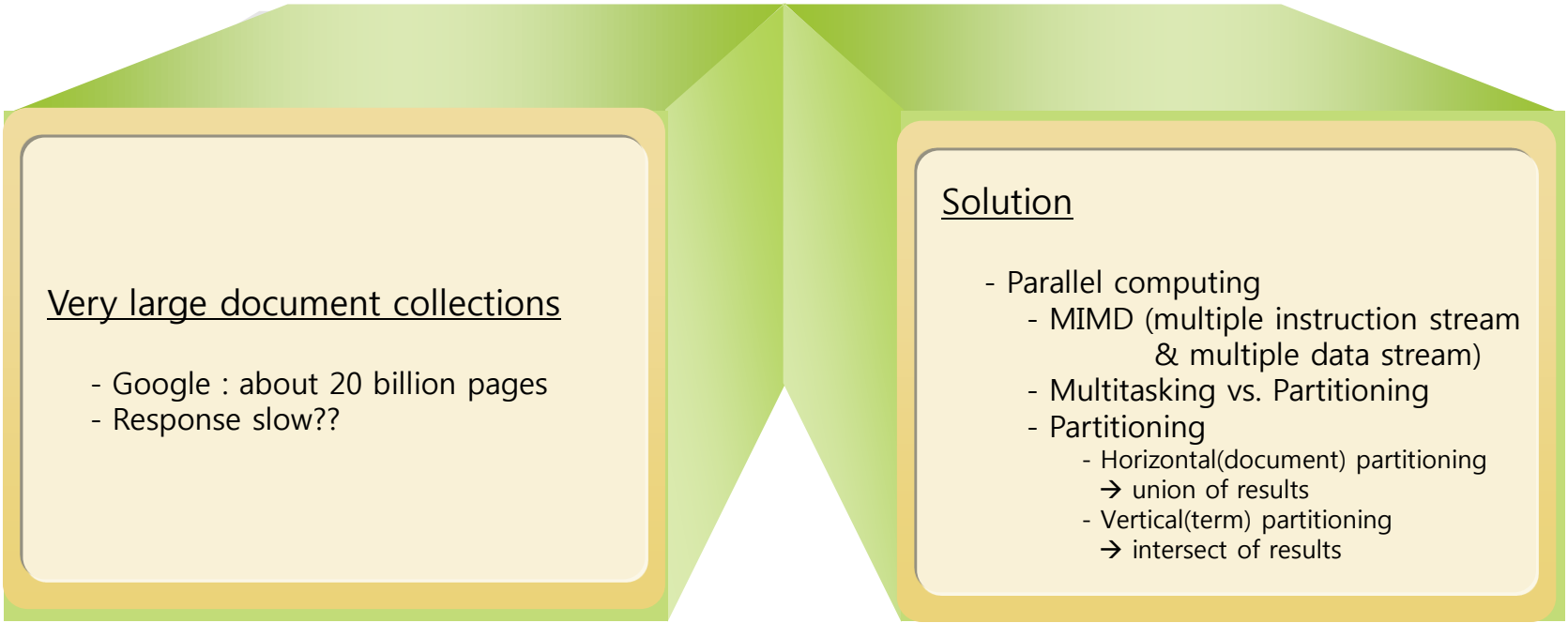
Query Efficiency



Index Efficiency



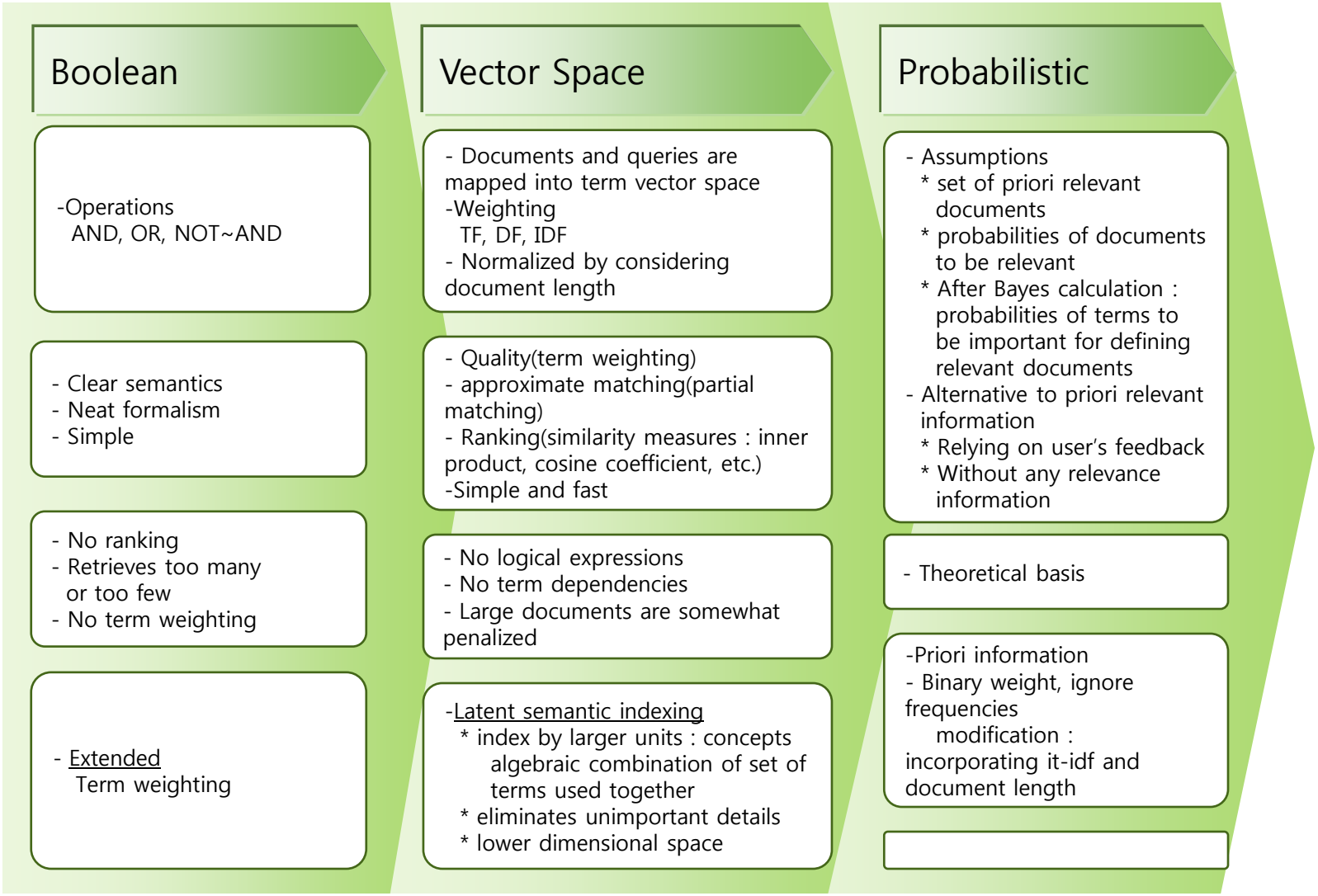
Scalability



Models

Boolean Model	Vector Space Model	Probabilistic Model
- Set theoretic - extended	- Algebraic - Generalized vector - Latent semantic indexing (LSI) - Neutral network	- Inference network - Belief network

Strength and weakness of each Model



Searching and Ranking

Retrieve matched documents on inverted indexes

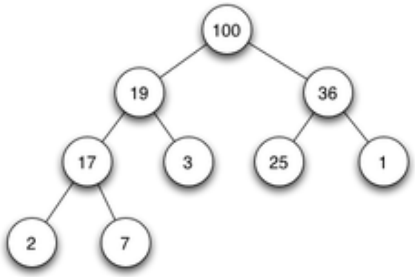
- Single term : locate entry, retrieve documents from list
- Conjunction of terms : intersection of lists
- Disjunction of terms : union of lists
- Negation of terms : complement of lists

Types of Ranking

- Relevance between a query and a document
- Features of documents : date, size, importance of a document, etc.

Ranking with heap

- Heap
  - A specialized tree-based data structure
  - key of a node is less than key of its parent
  - Time complexity : 0 (n * log n)
- Ranking
  - A heap maintains only top m-documents with reversed order
  - Each documents is compared to the document in root node.
  - Replace root node and heapify when new document is bigger.
- Benefit of using heap in ranking
  - Time complexity to get top m-documents : 0 (n * log m)



Query Processing

Wildcard Query

- mon* : find all documents containing any word beginnig "mon"
- Union of terms : retrieve all word in range (mon ≤ w < moo)
- *mon : nearly impossible
- Permuterm index
  - * for word hello index under :
    - * hello$, ello$h, llohe, lohel, o$hell
  - * queries
    - * X : lookup on X$, X* : lookup on $X*
    - * *X : lookup on X$* *X* : lookup on X*
    - * X*Y : lookup on Y$X*
- n-gram index
  - * for word hello index under (bigram index) :
    - * h, he, el, ll, lo, o
  - * queries
    - * hel : $h {AND} he {AND} el {AND} l$, hel* : $h {AND} he {AND} el
    - * *hel : he {AND} el {AND} l$ *hel : he {AND} el
    - * he*lo : $h {AND} he {AND} lo {AND} o$

Query Analysis

- detect user intension from query and expand search request
- ex1) "dentist's near KangNam stop"
  - * collection : local information
  - * keyword : dentist
  - * sorting : distance from KangNam stop → point of interest
- ex2) "latest song of The Beatles"
  - * collection : music
  - * keyword : Beatles
  - * target field : singer
  - * sorting : descend of date