IR prep

* mod 2

1. what do you mean by IR models?

→ A retrieval model can be ~~described~~ description of either the computational process or the human process of retrieval: The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

An IR model is a quadruple [D, Q, F, R(qi, dj)] where

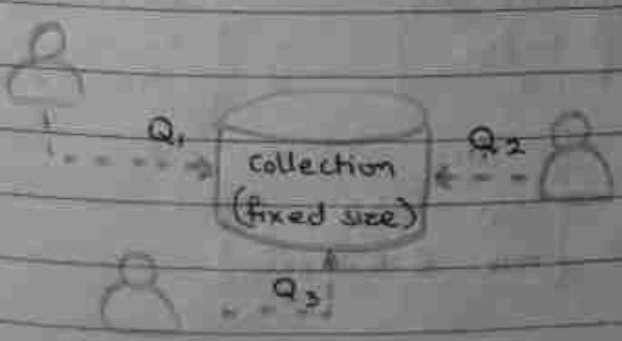
i. D is set of logical view for docs in collection
ii. Q ——————— user queries
iii. F is Framework for modelling docs. & queries
iv. R(qi, dj) is a ranking function

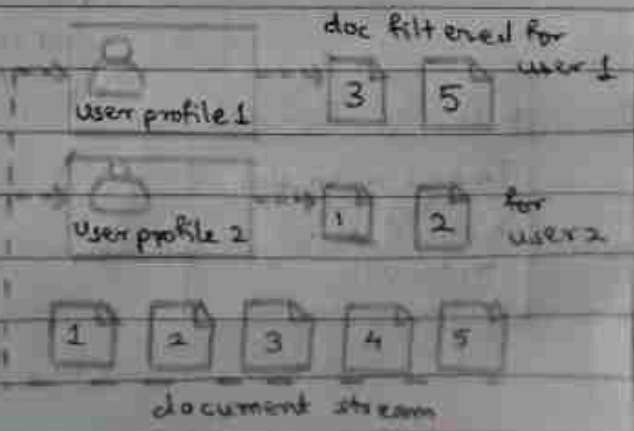
2. Differentiate between Adhoc & Filtering retrieval

→

	Adhoc	Filtering
Nature of collection	- Static doc. collection	- Dynamic doc. collection
Query/User profile	- Uses specific queries	- Uses a user profile built over time
User interaction	- Requires active querying by user	- Passive, based on predefined preferences
Doc processing	- Searches within a static collection	- Processes incoming doc. stream
Time frame	- Typically addresses immediate info. needs	- An ongoing process, continuously filtering documents.
Primary task	- Retrieve info based on query	- Selection & ranking of relevant incoming docs.
Adaptability	- Static doc. collection with changing queries.	- Static user profile with dynamic doc. stream.
Example	- Searching db for relevant doc.	- Filtering emails based on user preferences

Diagram -





3. Formal characterization of IR models

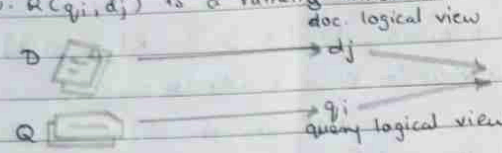
→ An IR model is a quadruple [D, Q, F, R(qi, dj)] where

i. D is a set of logical view for docs. in collection

ii. Q ——"—— for user queries

iii. F is framework for modeling docs and queries

iv. R(qi, dj) is a ranking function



4. 3 classic models in IRS?

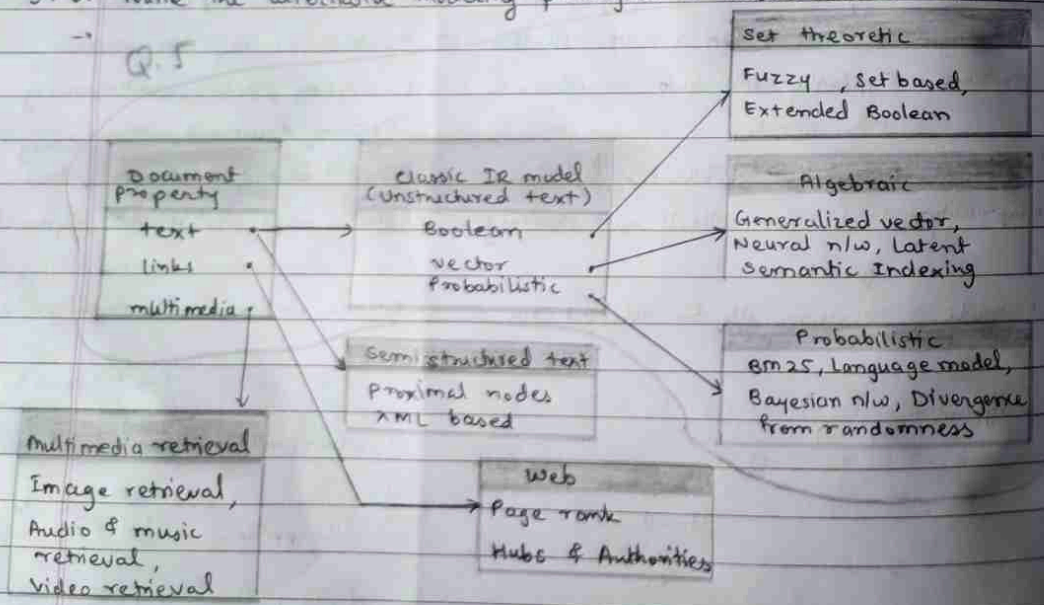
→ i. Boolean model: Simple model based on set theory & boolean algebra.

ii. Vector space model: Assigns non-binary weights to index terms in queries and in documents.

iii. Probabilistic model: Captures IR problem using a probabilistic framework.

5 & 6. Name the alternative modeling paradigms of classic models in IRS.

→



7. Define Boolean model and explain it with an example.

→ It is a simple retrieval model based on set theory and boolean algebra. It use binary decision criterion to decide whether a doc. is relevant or not relevant (no partial match). The term document frequency in term document matrix are all binary,

$w_{ij} \in \{0, 1\}$ weight associated with pair (k_i, d_j)

w_{iq} ——"—— (k_i, q)

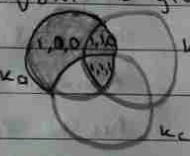
A query q is composed of index terms linked by 3 connectives, NOT, AND, OR. A query is a conventional Boolean expression, which can be represented as a disjunction of conjunctive vectors (disjunct normal form [DNF]).

Consider a query, $q = k_a \wedge (k_b \vee \neg k_c)$

vocabulary $V = \{k_a, k_b, k_c\}$, then q_{DNF} is given as,

$q_{DNF} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$

∴ Conjuctive components of query:



8. Basis for boolean model.

→ Simple model based on set theory & boolean algebra. Documents are sets of terms. Queries are specified as Boolean expression on terms, that can be represented as a disjunction of conjunctive vectors.

9. Advantages of boolean model.

→ Clean formalism, Quite intuitive & precise semantics, Easy to implement, still dominant model for document db system.

10. Disadvantages of boolean model.

→ - No notion for partial matching. - Not simple to translate info. to boolean expression
- Exact matching may lead to retrieval of too few or too many documents.
- Difficult to rank o/p, some are more imp. - more like data retrieval not info. retrieval
- Doesn't use term weights. (all terms are equally weighted)

11. Significance of tf and idf. How can you calculate them in vector model

→ Term frequency (TF) and Inverse document frequency (IDF) are the foundations of most popular term weighting scheme in IR. Luhn Assumption states that, value of w_{ij} is proportional to term frequency $f_{i,j}$, i.e, the more often a term occurs in doc, the higher its weight is. It's based on the observation that high frequency terms are more imp. for describing doc. ($tf_{ij} = f_{i,j}$)
A variant of tf weight used in literature is:

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$ where log is taken in base 2

Inverse doc. frequency (IDF) addresses the challenge of assigning optimal index terms to doc, recognizing that while more terms increase retrieval probability, too many can lead to irrelevant results. IDF incorporates both semantic and statistical definition being the inverse of a term's occurence across doc.
Acc. to Zipf's law, $n(r) \propto (r^{-\alpha}$ $n(r) \rightarrow r^{th}$ largest doc. freq.
Let $\alpha = 1$ (for english), take log $c, \alpha \rightarrow$ Emperical constant
on both sides, $\log n(r) = \log C - \log r$
Normalize assuming $C = N$, where N is no. of docs in collection
$\therefore \log r \approx \log N - \log n(r)$
Let k_i be the term with r^{th} largest doc. freq., i.e., $n(r) = n_i$
Then $idf_i = \frac{N}{n_i}$ where idf_i is inverse doc. freq. of term k_i

12. Parameters in calculating weight for doc term or query term.

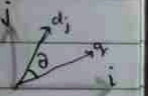
→ Term frequency (tf): No. of times a term i appears in doc j (tf_{ij})
Document frequency (df): No. of docs a term i appears in (df_i)
Inverse document frequency (idf): A discriminating measure for a term i in collection. $idf_i = \log_2 (N/df_i)$

13. Assumptions of vector space model.

→ - Index terms are assumed to be independent
- Degree of matching can be used to rank-order docs
- Rank ordering corresponds to how well a doc satisfying user's info. needs

14,15. Vector space model

→ Vector model proposes a framework in which partial matching is possible. It is accomplished by assigning non-binary weights to index terms in queries & in docs. These are used to compute degree of similarity (D.O.S) between query & doc & docs are ranked in decreasing order of D.O.S. weight associated with a pair (k_i, d_j) is positive & non-binary. Index terms are assumed to be mutually independent. They are represented as unit vector in t-dimensional space where t is no. of index terms.



$\vec{d_j} = (w_{1j}, w_{2j}, \ldots w_{tj})$, $\vec{q} = (w_{1q}, w_{2q}, \ldots w_{tq})$

$$\cos\theta = \frac{d_j * q}{|\vec{d_j}| \times |\vec{q}|} \quad \therefore sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

$\because w_{ij} > 0$ and $w_{iq} > 0$, we have $0 \leq sim(d_j, q) \leq 1$
weights in vector model are basically tf-idf weights.

$$w_{iq} = (1 + \log f_{iq}) \times \log N/n_i$$
$$w_{ij} = (1 + \log f_{ij}) \times \log N/n_i$$
} These eq are applied when $tf_i > 0$ else, respective weight is 0.

16. Advantages of vector model.

→ - Simple model based on linear algebra - Term weights are non-binary
- Allows computing a continuous degree of similarity between queries & doc.
- Allows ranking docs acc. to their possible relevance - Allows partial matching
- Allows efficient implementation for large doc. collection

17. Disadvantages of vector model.

→ - Index terms assumed to be mutually independent - weighting is not very formal
- Long docs are poorly represented - Searched keywords must precisely match doc term
- Order in which terms appear, is lost in vector space representation

18. What is link analysis?

→ Goal of IR is to find all docs that are relevant for a user query in a collection of docs. with advent of web, new source of info. became available, one of them being hyperlink between docs & records of user behavior. Hyperlinks provide a valuable source of info. for web info. retrieval. This area of info retrieval is commonly called link analysis.

21. Define probabilistic model or Binary independence retrieval.

→ Objective of probabilistic model is to capture IR problem using a probabilistic framework. Given a user query, the ideal answer set, referred to as R, should maximize the probability of relevance.

weight variables are all binary, i.e, $w_{i,j} \in \{0,1\}$ & $w_{i,q} \in \{0,1\}$

where q is query that's a subset of index terms.

Let, R be the set of relevant doc to query q.

$\bar{R}$ —||— non-relevant —||—

$P(R|\vec{d_j})$ be the probability that d_j is relevant to query q

$P(\bar{R}|\vec{d_j})$ —||— non-relevant —||—

$\therefore sim(d_j, q) = P(R|\vec{d_j}) \Big/ P(\bar{R}|\vec{d_j})$

22. What are the fundamental assumptions for probabilistic principles.

→ Model assumes, relevance depends only on query and doc representation

Suppose, $q \rightarrow$ user query, $d_j \rightarrow$ docs in collection,

$R \rightarrow$ ideal answer set, relevant to q

$\bar{R} \rightarrow$ ideal answer set, non-relevant to q

Similarity to query ratio is that, probabilistic ranking is computed as, Ratio = P(dj relevant to q) / P(dj non-relevant to q)

The rank minimizes probability of erroneous judgement.

23. Adv: Docs are ranked in descending order of their probability of relevance

Disadv:- All weights are binary. - Need to guess initial separation of docs into relevant & non-relevant sets.

- Guess inital value of $P(k_i|R)$
- Adoption of independent assumption for index terms
- method does not take into account tf & idf factors

24. Discuss structured text retrieval models in detail.

→ Retrieval models which combine info. on text content with info. on the doc structure are called structured text retrieval models. Some imp. terms include: match point refers to position in text, of a sequence of words which match the user's query. Region refers to a contiguous portion of text. Node refers to a structural component of doc, such as chapter, section or a subsection.

i) Non-overlapping lists: Divide the whole text of each doc in non-overlapping text regions which are collected in a list. Text regions in same list have no overlapping, but text regions from distinct lists might overlap.

ii) Proximal nodes: A model which allows definition of independent hierarchial indexing structures over the same document text. Each of these index structures is a strict hierarchy composed of chapters, sections, paragraphs, pages and lines which are called nodes.

20. TF and Normalized TF.

→ Let n_i be no. of docs in which index term k_i appears. Let $f_{i,j}$ be the raw frequency of term k_i in doc d_j (no. of time k_i is mentioned in doc d_j). Then normalized frequency $tf_{i,j}$ of term k_i in doc d_j is given by, $tf_{i,j} = \frac{f_{i,j}}{\max_i f_{i,j}}$

where max is computed over all terms which are mentioned in text doc d_j. If k_i doesn't appear in d_j then $tf_{i,j} = 0$.

* mod 1

1. Define information retrieval.

→ IR is the activity of obtaining info. system resources that are relevant to an info.

IR is finding material (docs) of an unstructured nature (text) that satisfies an info. need from within large collections. It is the activity of obtaining information relevant to the need, from a collection of resources. searches can be based on text or other content based on indexing. IR deals with Representation, Storage, Organization and Access of information items.

2. Applications of IR.

→ i. Search engines: mobile searches, web search, social search, site search
ii. Information filtering: Reccomendation system
iii. Query processing iv. Publish/subscribe system

3.	Data Retrieval	Information Retrieval
Definition	- Process of identifying and retrieving the data from the db, based on query provided by user or app.	- s/w program that deals with representation, storage, organization & access of information items.
working	- Determines keywords in user query and retrieves data.	- Retrieves information about a subject.
Error sensitivity	- Single error means total failure. Sensitive.	- Small errors are likely to go unnoticed. Insensitive.
Structuring	- well defined structure & semantics.	- Not always well structured & semantically ambiguous.
matching	- Exact matching	- Partial / best match
Results	- Not ordered by relevance	- ordered by relevance
model type	- Deterministic	- Probabilistic
Provides solution?	- provides solution to db system user	- Doesn't provide solution to db sys user

4. Explain general objective of IRS.

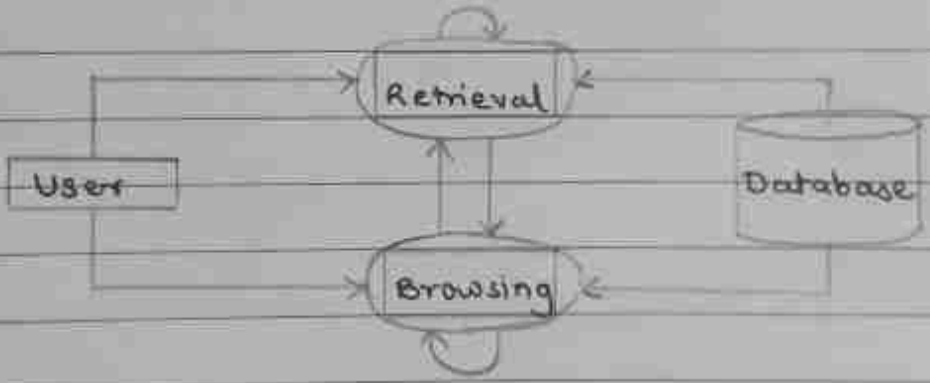
→ The objective of an IRS is to enable users to find relevant info. from an organized collection of documents in response to the user query. To minimize irrelevant info. and provide user with relevant info. in least time and least efforts. Handle various types of data. Scale effectively to handle large volumes of data and concurrent users. Accomodate different query types and formats.

5. Define relevance

→ Relevance appears to be a subjective quality, unique between the individual and a given doc. supporting the assumption that relevance can only be judged by the info. user. subjectivity and fluidity make it difficult to use as a measuring tool for system performance.

6. List and explain IR block diagram.

→



Retrieval and Browsing in www are both pulling actions. Retrieval (searching) is classic info. search process where clear objectives are defined. Consider a user who seeks info on a topic of their interest. This user first translates their info. need into a query, which requires specifying words that compose the query. In this case, we say that the user is retrieving or querying for info. of their interest.

Browsing is a process where one's main objectives are not clearly defined and might change during their interaction with system. Consider the user has an interest that's poorly defined or broad, eg. they want to browse documents on F1 racing and has an interest in car racing. In this case, we say that the user is browsing or navigating the documents of collection.

7. what is conflation?

→ Stemming is the process for reducing inflected words (change form of word) to their stem, base or root form, generally a written word form. This process is also called conflation.

9. what is search engine?

→ A search engine is a doc retrieval system designed to help find info. stored in computer system, such as on WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria.

10. Functions of IRS

→ i. Identify info. sources relevant to areas of interest of target user (community)

ii. Analyze contents of the source (doc.)

iii. Represent contents of the analyzed sources in a way that will be suitable for matching user queries.

iv. Analyze user's query & represent them in a form suitable for matching with db.

v. match search statement with stored db

vi. Retrieve info. that is relevant

vii. make necessary adjustments based on user feedback.

11. Issues in IRS

→ i. Assist user to clarify & analyze problem and determine info needs

ii. Knowing how people use & process info

iii. Assemble a package of info that enables to group the users and come closer to solution of the problem.

iv. knowledge representation

v. Procedures for processing knowledge/info

vi. Human - computer interface

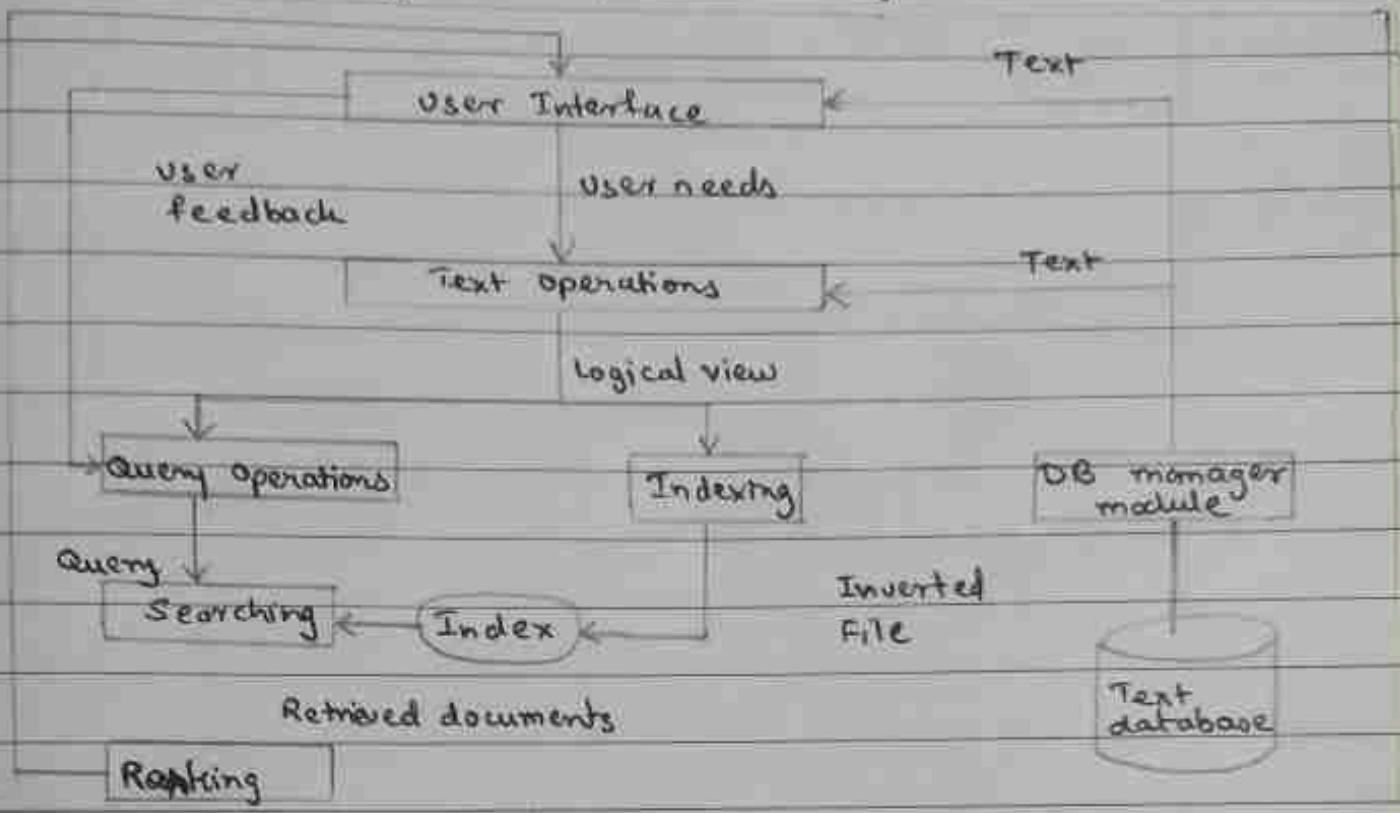
vii. Designing integrated workbench system

viii. Designing user-enhanced info. system

ix. System evaluation

8. IR process with a generic s/w archi. diagram.

→



- Before the retrieval process can be initiated, it is necessary to define text db, which is done by manger of db.
- Manger of db specifies: · Docs to be used · Operations to be performed on text
  · Text model
- Text operations transform original docs & generate a logical view of them
- DB manager (using db manager module) builds an index of text
- Given that doc db is indexed, the retrieval process can be initiated.
- User specifies user needs which is then parsed & transformed to text
- Query operations might be applied to generate actual query, which provide system representation for the user needs.
- Query is then processed to obtain the retrieved docs. Fast query processing is made possible by index structure built earlier.
- Retrieved docs are ranked acc. to likelihood of relevance
- User examines the set of useful info. They might pinpoint a subset of doc as interest and initiate a user feedback cycle. System uses user selected docs to change query formulation. Modified query is better representation of real user need.

* mod 3

1. State different type of queries

→ Query processing is the activity performed in extracting data from the db. It takes various steps to fetch data from db. Steps involved are: Parsing, Translation and optimization. The queries applied on structure and unstructured data stored in db, combined with IR techniques can lead to faster and efficient processing of data. Three major types include: keyword based querying, Pattern matching and structural queries.

2. Explain pattern matching query concept with an example.

→ Pattern matching allows the retrieval of pieces of text that have some property (match a pattern). A pattern is a set of syntactic features that must occur in a text segment.

i) Words: Most basic pattern. String must contain the word in query text.

ii) Prefixes: String must form the beginning of the text word.
Eg. 'inter' in words 'international', 'interactive', etc.

iii) Suffixes: String must form the termination of text word.
Eg. 'dom' in words 'freedom', 'kingdom', etc.

iv) Substring: String can appear within a text word.
Eg. 'pal' in 'palace', 'palm', 'municipality', etc.

v) Ranges: matches any word lying between a pair of strings (alphabetical) in lexicographical order. Eg. 'held' and 'hold' retrieve words such as 'hiss', 'hoax', etc.

vi) Allowing errors

vii) Regular expressions

3. Explain keyword based querying? Discuss content & boolean queries.

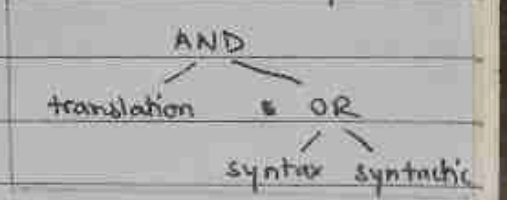
→ Simplest and most widely used kind of IR queries. It requires user to simply enter phrase combinations to retrieve docs. People look for similar docs using keywords. A logical AND operator creates an implied connection between the query keyword terms. When search for 'information retrieval', the first retrieved doc will be the doc containing both words 'info' & 'retrieval'.

Additionally, majority systems also retrieve docs containing either one of those two keywords too. Before delivering the filtered query keywords to IR engine, stopwords are removed in preprocessing.

i) Context queries: Search words in given context, i.e., near other words. words that are close to each other sugger higher possibility of relevance than words that are far apart. It uses phrases & proximity. Phrases are sequence of single word queries that each retrieved doc must contain atleast one instance of. Proximity refers to how close within a record, multiple items should be. Eg. 'enhance" retrieval' should occur within 4 words will match 'enhance the power of retrieval'... word or phrases may not need to be in the same order as they are in the query.

ii) Boolean queries: Use a syntax composed of atoms that retrieve doc and Boolean operators that work on operands. It allows use of AND, OR, (), NOT, +, - boolean operators in combination with keywords.

Eg. translation AND syntax OR syntactic is



AND: requires both terms to be found

OR: requires either term to be found

NOT: record containing second attribute will be removed

( ): Boolean operators can be nested using parentheses

+ : equivalent to AND. '+' should be place directly in front of req. term.

- : equivalent to AND NOT. '-' ——— not req. term.

No ranking is possible as a doc either satisfies the condition or does not (non relevant). A doc is retrieved if the query is logically True as an exact match in doc. Complex queries can be built using operators and their combination and they are evaluated using rules of classic boolean algebra.