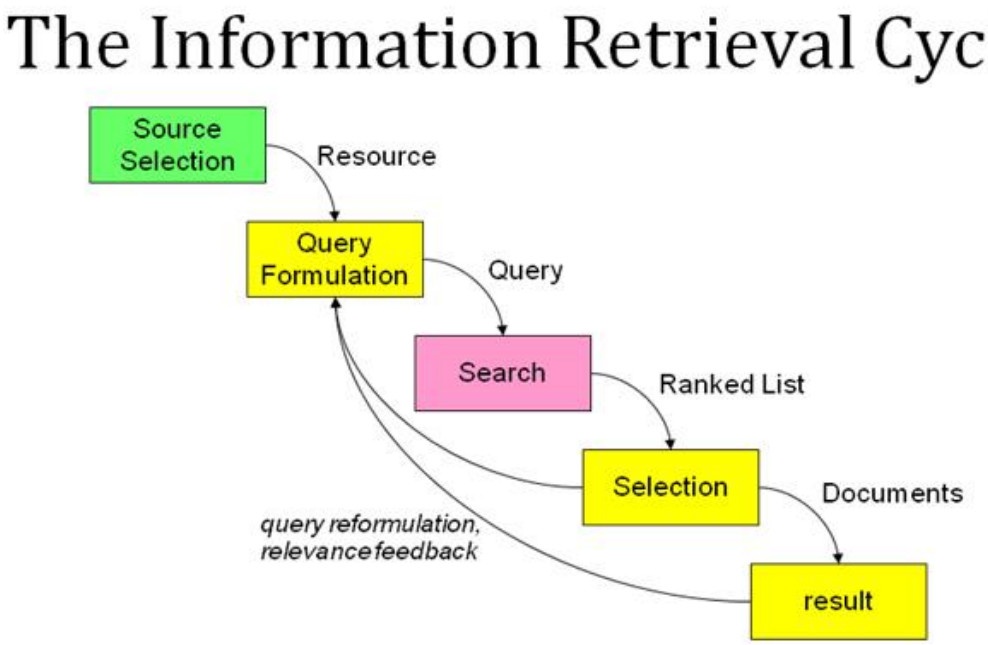
**Query Formulation Process:**

**Definition of Query:**
Query is defined as any question, especially one expressing doubt or requesting information or to check its validity or accuracy of information.

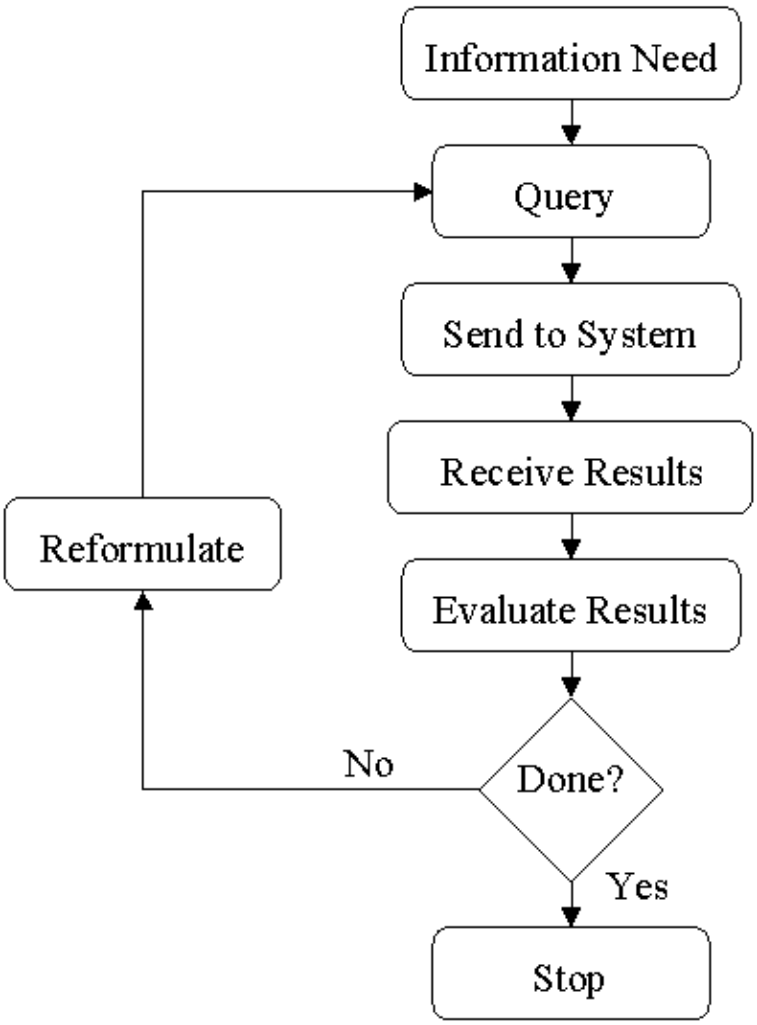
**Query formulation and Information and information retrieval:**
“Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines are employed to carry out the operation.”

Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

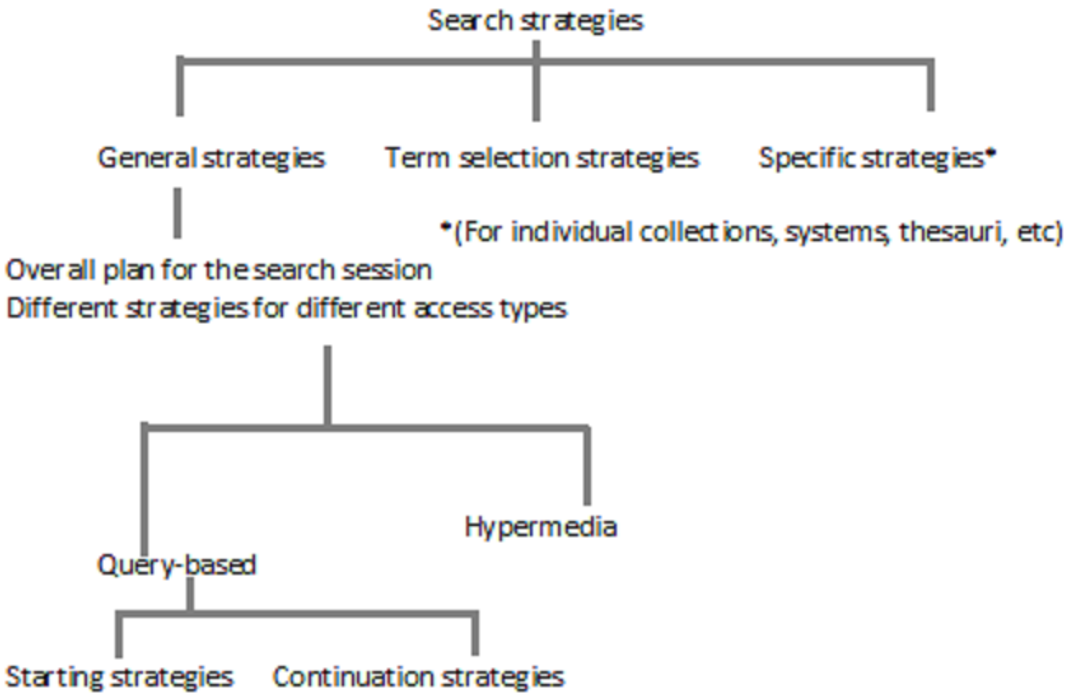



Information Retrieval is a research-driven theoretical and experimental discipline. The focus is on different aspects of the information–seeking process, depending on the researcher’s background or interest.

**The Standard Retrieval Interaction Model**




Query is one of the components in the IR cycle. The information search process has the following steps such as Enrich query formulation, Expand result management, Enable long-term effort, Enhance collaboration.




**Starting strategies**

Select-Break complex query into topics and deal with each topic separately

Exhaust-Include most elements of the query in the initial query formulation

**Continuation strategies**

Building blocks - Combination of discrete topics

Pearl growing - Small relevant set expanded gradually

Successive fractions - Large relevant set refined gradually

**Query Formulation:**

Most standard information retrieval models use a single source of information (e.g., the retrieval corpus) for query formulation tasks such as term and phrase weighting and query expansion

***Query formulation – a process during which the original keyword query issued by the user is transformed into a structured query representation that is consumed by the search engine.***

**Query Formulation Processing:**

The process of query formulation (also referred to as query rewriting or query transformation) modifies the original keyword query submitted by the user to the search engine in order to better represent the underlying intent of the query.

The formulated query is then used as an input to the search engine's ranking algorithm. Thus, the primary goal of query formulation is to improve the overall quality of the ranking presented to the user in response to their query.

Query formulation is usually divided into two main processing stages.

I. The first processing stage, which is usually referred to as query refinement, alters the query on the morphological level (e.g., tokenization, spelling corrections,stemming, etc.).

**Query term processing:**

**Tokenization**

- Cut character sequence into word tokens
  - Deal with ***"John's"***, ***a state-of-the-art solution***

**Normalization**

Map text and query term to same form

- You want ***U.S.A.*** and ***USA*** to match

**Stemming**

- We may wish different forms of a root to match
  - ***authorize***, ***authorization***

**Stop words**

- We may omit very common words (or not)
  - ***the, a, to, of***

II. After the query refinement stage is completed, the second processing stage alters the query on the structural level.

Such structural alterations may include, among other actions, segmenting the query into atomic concepts (i.e., combinations of terms), assigning weights to these concepts, or expanding the query with related weighted concepts.

**Query expansion:**

Known relevant documents contain terms that can be used to describe a larger cluster of relevant documents. Query expansion based on Thesauri, Lexical/statistic analysis of text / context and concept formation and Relevance feedback.

1. Global strategy: all documents in the collection used to determine a global thesaurus-like structure which defines term relationships. This is shown to the user who selects terms for query expansion.
2. 
3. Local strategy: local set for a query are examined at query time to determine terms for query expansion.

**Query refinement**

Encourage users dissatisfied with the top search results to query again.

User types a search query. Result is search results **plus** possible new search queries, which when clicked on issue a new query whose results to replace the current results seen by the user.

A mechanism that recommends query modifications to reduce false positives
Incremental process of transforming a query into a new query that more accurately reflects the user's information need.

**Different forms of queries:**

**Query-By-Form** is the simplest querying method, but it is neither flexible nor expressive.

**Query-By-Example** A known approach in databases, where users formulate queries as filling

**Conceptual Queries**
As many databases are modeled at the conceptual level using EER, ORM or UML diagrams, one can query these databases starting from their diagrams. Users can select part of a given diagram, and their selection is translated into SQL, ConQuer & Mquery etc.

**Natural Language Queries** allow people to write their queries as natural language sentences, and then translate these sentences into a formal language (e.g., SQL , XQuery ).

**Visualize queries**

Several Semantic Web approaches (Isparql, RDFAuthor, GRQL, Nitelight) propose to formulate a SPARQL query by visualizing its triple patterns as ellipses connected with arrows, so that one would need less technical skills to formulate a query.

**Interactive Queries**
Asking and answering method.

**Enrich query formulation**

Query formulation enriched by the following aspects:

Previous & Similar, Structured input, Spell check, Query previews, Finding Aids, Limit:(Time, Geography, Language, Sources, Media etc )

**Supporting query formulation:**

Spelling correction of the query, Suggestions with definitions to the query, Searching for disambiguated concepts, Implicit information about users used for search, Geo location and user language.

**Querying is an iterative process**

Expand the original query with new terms.

Re-weight the terms in the expanded query.

Direct feedback from user – filtering.

Information derived from the set of documents initially retrieved (local set).

Global information derived from the whole document collection.

Shields the user from the details of the query formulation process, and permits the construction of useful search statements without intimate knowledge of collection make-up and search environment.

Breaks down the search operation into a sequence of small search steps, designed to approach the wanted subject area gradually.

Provides a controlled query alteration process designed to emphasize some terms (relevant ones) and to de-emphasize others (non-relevant ones).

**Levels of search activities according to Bates 1990**

**Move:** Low-level search function
(e.g. type in search term, view retrieved document)

**Tactic:** several moves to further a search
(e.g. broaden/narrow a query)

**Stratagem**: set of actions on a single domain
(citation database, tables of contents of journals)

**Strategy:** complete plan for satisfying an information need
(e.g. subject search, browse relevant journals, find referenced articles

**Examlpes for query formulation process:**

**Boolean Queries**

The Boolean retrieval model is being able to ask a query that is a Boolean expression.

Boolean Queries are queries using *AND, OR* and *NOT* to join query terms. Many search systems you still use are Boolean:

– Email, library catalog, Mac OS X Spotlight

**Westlaw**

Largest commercial (paying subscribers) legal search service (started 1975; ranking added 1992; new federated search added 2010)

**Phrase queries**

We want to be able to answer queries such as ***"stanford university"*** – as a phrase***.***

Thus the sentence *"I went to university at Stanford"* is not a match.

The concept of phrase queries has proven easily understood by users; one of the few "advanced search" ideas that works

**Advantages of query formulation process:**

Query easy to specify

The output is ranked based on the estimated relevance of the documents to the query

A wide variety of theoretical models exist

Very precise queries can be specified

Very easy to implement (in the simple form)

**Disadvantages of query formulation process:**

Specifying the query may be difficult for casual users

Lack of control over the size of the retrieved set

Query less precise (although weighting can be used)

**Summary:**
Large relevant set refined gradually

Matching between the document and the query in the abstracted space of the set of index words is very imprecise.

Relational Databases and Query language exemplify data retrieval due to semantic clarity and precision.

Document querying can be transitioned to data retrieval if only we could re-author all the docs so as to provide formalnsemantics to them and implement sufficiently powerful query language.

Semantic Web (RDF and SPARQL) is a step in that direction.

Text operations generate a logic view of the query and the documents (after stopword elimination, stemming, etc).

Query operations may formulate or modify the query and can involve query expansion, relevance feedback, etc.