Information Retrieval

By Sunil Regmi

sunilregmi233@gmail.com

Outlines

1. Introduction to Information Retrieval [4 Hours]

    1.1. Overview of IR and its applications

    1.2. History and evolution of IR

    1.3. Components of IR systems

    1.4. Ethical Considerations in IR (privacy, bias, and fairness)

Before diving into information retrieval, it's crucial to understand the distinction between data and information.

What is Data and Information?

Data is raw, unorganized facts and figures and defined as structured, semi-structured or unstructured information such as text, observations, images, symbols, and descriptions.

In other words, data provides no specific function and has no meaning on its own.

Information ???

Information refers to processed, organized, and structured data.

It gives context for the facts and facilitates decision making.

In other words, information is processed data that makes sense to us.

Data: A list of numbers (e.g., 1, 2, 3, 4, 5)
Information: The average of those numbers is 3.

In terms of NLP...

Data:

- Raw text: This is the fundamental building block of NLP. It could be anything from books, articles, social media posts, emails, or code.
  - Example: "The quick brown fox jumps over the lazy dog."
- Numerical representations: Text can be converted into numerical formats for machine processing.
  - Example: Word embeddings, where words are represented as dense vectors of numbers.
- Structured data: While less common in NLP, structured data can also be used, such as tables of entities and relationships.
- Example: A table of movie reviews with columns for reviewer, rating, and text.

Information

- Semantic meaning: Understanding the underlying meaning of text.
  - Example: Recognizing that "The quick brown fox" refers to animals.
- Sentiment: Determining the emotional tone of text.
  - Example: Identifying that "This movie was terrible" expresses negative sentiment.
- Named entities: Identifying specific entities like people, organizations, or locations.
  - Example: Extracting "Apple" as a company from the sentence "Apple released a new iPhone."
- Relationships: Understanding how entities are connected.
  - Example: Identifying that "Steve Jobs" is the founder of "Apple."
- Knowledge: Derived from processing information, it represents understanding of the world.
  - Example: Knowing that dogs are mammals and that they typically bark.

Types of Data

1. Structured
2. Semi-Structured
3. Unstructured

Structured Data

Highly organized data with a predefined format, typically stored in relational databases. It follows a rigid schema with rows and columns.

Characteristics:

- Easily searchable and queryable.
- Efficient for analysis and reporting.
- Examples: CSV files, Excel spreadsheets, SQL databases.

Use Cases:

- Financial data
- Customer demographics
- Sales figures

Semi-Structured Data

Data that exhibits some organizational properties but doesn't conform to a strict data model like structured data. It often contains tags or markers to separate semantic elements.

Characteristics:

- More flexible than structured data.
- Can be processed using specialized tools.
- Examples: XML, JSON, CSV without predefined schema.

Use Cases:

- Social media feeds
- Sensor data
- Log files

Unstructured Data

Data without a predefined data model or organization. It's raw and unorganized.

Characteristics:

- Challenging to process and analyze.
- Requires advanced techniques like machine learning and natural language processing.
- Examples: Text documents, images, audio, video, social media posts.

Use Cases:

- Text analysis
- Image recognition
- Speech recognition

...

not planned before it happens

- Unstructured data typically refers to **free text.**
- It allows **Adhoc and More sophisticated queries**.
- Keyword queries including operators like **and, or, not** etc.
- More sophisticated **"concept" queries** e.g., find all web pages dealing with football matches.
- Unstructured data means **semantically raw and easy for computer.**
- It may be **flat** or having **structure** like
  - Document structure (headings, paragraphs, lists. . .)
  - Explicit markup formatting (e.g. in HTML, XML. . .)

So, while talking about information retrieval, we generally refer to information extraction from unstructured data...

Sometimes (form structured and semi-structured data)

Just getting a credit card out of your wallet so that you can type in the card number is a form of information retrieval.

Information Retrieval

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).

- deals with the representation, storage, organization of, and access to information items.

- The representation and organization of the information items should provide the user with easy access to the information in which he/she is interested.

- In reality, **almost no data are truly "unstructured,"**
  - For example, text data may seem unstructured, but it typically follows linguistic rules and patterns, such as grammar and syntax, that provide a kind of structure.
  - most text has structure, such as headings, paragraphs, and footnotes, often represented by explicit markup.
- IR facilitates "semistructured" search, such as finding a document where the title contains Java and the body contains threading.

- The field of IR also covers browsing or filtering document collections and further processing retrieved documents.

- Clustering is the task of grouping documents based on their contents, similar to arranging books by topic.

- Classification involves deciding which class(es) a set of documents belongs to, often using a combination of manual classification and automatic methods.

IR systems operate at different scales:

Web search involves searching billions of documents across millions of computers, requiring efficient systems to handle enormous scale and web-specific issues like hypertext exploitation and page ranking manipulation.

Personal information retrieval is integrated into consumer operating systems like Apple's Mac OS X Spotlight and Windows Vista's Instant Search, focusing on handling various document types, lightweight processing, and maintenance-free operation.

Enterprise, institutional, and domain-specific search covers retrieval for collections like internal corporate documents, patent databases, or research articles, usually managed by centralized file systems and dedicated machines.

Difference between IR (Information Retrieval) and DR (Data Retrieval)

Information retrieval is the process of finding and returning relevant documents or unstructured data based on a user’s query, like using a search engine.

Data retrieval involves fetching specific, structured data from a database, such as querying a database for customer records.

Aspect	Information Retrieval	Data Retrieval
Type of data	Typically unstructured	Typically Structured or Semi-structured
Model it uses	Boolean or probabilistic models	Deterministic models
Methodology	Typically uses a bag of words model of the source text.	Typically based on some form of semantic analysis of the source text.
User interaction language	Any natural languages	Need to learn languages like, SQL,MongoDB
Searching Methods	Check for relevant document for the given query or information need. (partial matching)	Exact matching needs to be done.
Display of Outputs	Links or set of relevant document	Table or record (depending on the query) Return facts out of documents
Goal	The goal is to find documents that are relevant to the user's information need	The goal is to extract pre-specified features from documents or display information.

Example of IR problem.

- To understand the problem of IR in general,
  - Consider a YouTube playlist for data compression.
  - In the playlist there are videos on various topics.
  - Now, let's assume the user want to retrieve the following information:

> “ Which videos in the playlist contains Terms: Huffman and Tree but NOT Dangling Suffix”

Applications of IR

Web Search

The most common application, powering search engines like Google, Bing, and Yahoo.

Enables users to find information, products, services, and people online.

Enterprise Search

Helps employees locate relevant information within an organization's internal systems, such as documents, emails, databases, and intranets.

Digital Libraries

Facilitates searching and retrieving books, articles, research papers, and other digital content.

E-commerce

Enables users to find products based on keywords, categories, and filters.

Recommender systems often utilize IR techniques to suggest products based on user preferences and purchase history.

Information Retrieval in Other Fields

Bioinformatics: Searching and analyzing biological data, such as DNA sequences and protein structures.

Legal: Finding relevant case law, statutes, and legal documents.

contd...

Intelligence: Analyzing text data for information extraction and intelligence gathering.

Customer Service: Providing efficient and accurate answers to customer queries.

Social Media Analysis: Analyzing social media content for sentiment analysis, trend identification, and market research.

Digital Humanities: Studying historical texts and documents using IR techniques.

History and Evolution of IR

Here we cover the history when there are no computer for information retrieval.

Early Developments (Before Computers)

- Mechanical and Electro-mechanical Devices (1891 - 1950s):
- Early attempts at information organization and retrieval using physical systems.
- Examples: Dewey Decimal Classification, card catalogs.

Ancient Libraries: The concept of organizing information for retrieval dates back to ancient civilizations like Egypt and Greece.

Manual Indexing: Early librarians developed systems for classifying and cataloging books, such as the Dewey Decimal System.

Mechanical Aids: In the 19th century, devices like punched cards were used for data processing, laying the groundwork for later computational methods.

contd...

Early use of computers for IR (1948 - 1950s):

- First explorations of using computers for information storage and search.
- Pioneering systems like IBM’s Selective Dissemination of Information (SDI).

...

The Computer Age (Mid-20th Century)

Pioneering Systems: Early computers were applied to information retrieval tasks, with systems like IBM's Selective Dissemination of Information (SDI) emerging.

Boolean Retrieval: This model, based on logical operators (AND, OR, NOT), became the foundation for early search systems.

Relevance Feedback: Techniques for refining search results based on user feedback were explored.

Vector Space Model: Introduced in the 1960s, this model represented documents and queries as vectors in a high-dimensional space, enabling similarity calculations.

...

Growth and Commercialization (1970s-1980s)

Online Databases: Commercial systems like DIALOG and LexisNexis offered access to large collections of information.

Interactive Search: Users gained the ability to refine queries in real-time.

Evaluation Measures: Metrics like precision, recall, and F-measure were developed to assess IR system performance.

Research on User Interfaces: Efforts focused on improving the user experience of search systems.

...

The Internet Era and Beyond (1990s-Present)

Web Search: The rise of the World Wide Web presented new challenges and opportunities for IR.

Information Overload: The vast amount of online information increased the need for effective search and filtering techniques.

Search Engines: Companies like Google developed sophisticated algorithms to rank search results.

Multimedia Retrieval: IR expanded to include images, audio, and video.

...

Natural Language Processing (NLP): Integration of NLP techniques improved understanding of user queries and document content.

Semantic Search: Efforts to understand the meaning of information beyond keywords.

Personalization: Tailoring search results to individual users based on their preferences and behavior.

Mobile Search: Optimization of IR for mobile devices.

Social Media Search: Integration of social media data into search results.

Information Retrieval (IR)???

Task: To find a few among many

It is probably motivated by the situation of information overload and acts as a remedy to it

When defining IR, we need to be aware that there is a broad sense and a narrow sense

Broad Sense of IR

- It is a discipline that finds information that people want
- The motivation behind would include
  - Humans' desire to understand the world and to gain knowledge
  - Acquire sufficient and accurate information/answer to accomplish a task
- Because finding information can be done in so many different ways, IR would involve:
  - Classification Clustering
  - Recommendation
  - Social network
  - Interpreting natural languages
  - Question answering
  - Knowledge bases
  - Human-computer interaction
  - Psychology, Cognitive Science
  - Any topic that listed on IR conferences such as SIGIR/ICTIR/CHIIR/CIKM/WWW/WSDM...

Narrow Sense of IR

- It is ‘search’
  - Mostly searching for documents
- It is a computer science discipline that designs and implements algorithms and tools to help people find information that they want
  - from one or multiple large collections of materials (text or multimedia, structured or unstructured, with or without hyperlinks, with or without metadata, in a foreign language or not
  - where people can be a single user or a group
  - who initiate the search process by an information need
  - and, the resulting information should be relevant to the information need (based on the judgement by the person who starts the search)

Narrowest Sense of IR

- It helps people find relevant documents
  - from one large collection of material (which is the Web or a TREC collection),
  - where there is a single user,
  - who initiates the search process by a query driven by an information need,
  - and, the resulting documents should be ranked (from the most relevant to the least) and returned in a list

Components of IR

Basic IR System Architecture

(Stefan Buettcher)

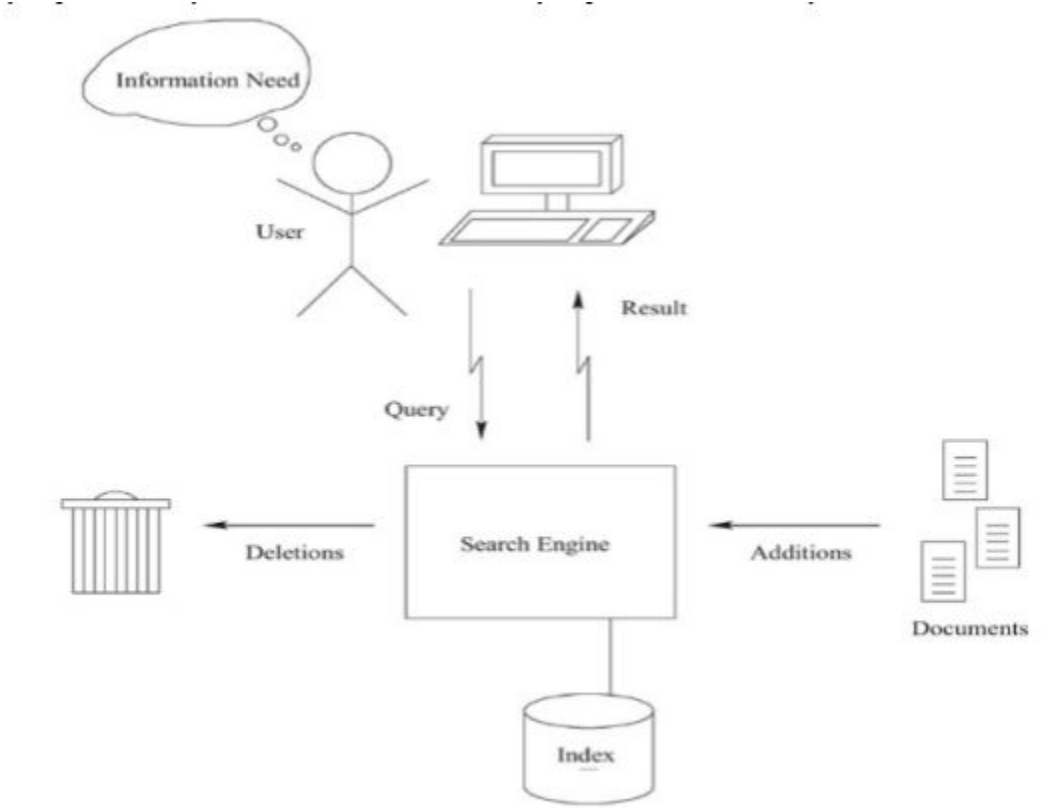


Figure 1.1 Components of an IR system.

Information Need:

- The search process begins with the user's information need, which motivates the query.
- This need is often referred to as a “topic” in IR evaluations.

Query Construction:

- Users express their information need as a query, typically consisting of a few terms (e.g., 2-3 terms for a web search).
- Queries can include diverse elements like dates, numbers, phrases, and wildcard operators (e.g., “inform*” matching “inform”, “informative”, etc.).

Query Syntax:

- While most users use simple keyword queries, IR systems can support complex Boolean and pattern-matching queries to refine searches, restrict to specific fields, or apply filters.

User Interface:

- The interface mediates between the user and the IR system, simplifying query creation, especially when advanced query syntax is used.

Search Engine Components:

- The search engine, whether local or remote, processes the user's query. It relies heavily on an inverted index, which maps terms to their occurrences in documents.
- The engine maintains collection statistics (e.g., term frequencies, document lengths) and accesses original content to provide meaningful search results.

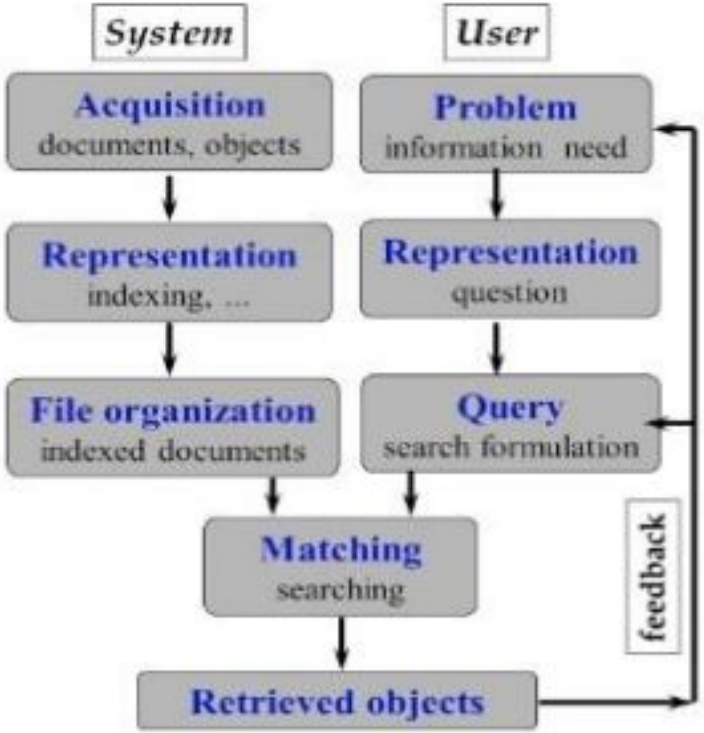
Relevance Ranking and Result Processing:

- The engine calculates a relevance score, often termed as the Retrieval Status Value (RSV), to rank documents.
- The ranked list undergoes further processing, such as removing duplicate or redundant results (e.g., showing only one result per domain in web searches).

Traditional IRS

Three major components of Traditional IRS

1. **Document subsystem**
    a. Acquisition
    b. Representation
    c. File organization
2. **User sub system**
    a. Problem
    b. Representation
    c. Query
3. **Searching /Retrieval subsystem**
    a. a) Matching
    b. b) Retrieved objects



Acquisition (Document subsystem)

- Selection of documents & other objects from various web resources.
- Mostly text based documents
  - full texts, titles, abstracts...
  - but also other objects:
    - data, statistics, images, maps, trade marks, sounds ...
- The data are collected by web crawler and stored in databases.

Representation of documents, objects(document subsystem)

- Indexing – many ways :
  - free text terms (even in full texts)
  - controlled vocabulary - thesaurus
  - manual& automatic techniques.
- Abstracting; summarizing
- Bibliographic description:
  - author, title, sources, date...
  - metadata
- Classifying, clustering
- Organizing in fields & limits
  - Basic Index, Additional Index. Limits

File organization (Document subsystem)

- Sequential
  - record (document) by record
- Inverted
  - term by term; list of records under each term
- Combination
- indexes inverted, documents sequential
- When citation retrieved only, need for document files
- Large file approaches
- for efficient retrieval by computer

Problem (user subsystem)

- Related to users' task, situation
  - vary in specificity, clarity
- Produces information need
  - ultimate criterion for effectiveness of retrieval
    - how well was the need met?
- Information need for the same problem may change, evolve, shift during the IR process adjustment in searching
  - often more than one search for same problem over time

Representation (user subsystem)

- Converting a concept to query.
- What we search for.
- These are stemmed and corrected using dictionary.
- Focus toward a good result
- Subject to feedback changes

Query - search statement (user & system)

- Translation into systems requirements & limits
  - start of human-computer interaction
    - query is the thing that goes into the computer
- Selection of files, resources
- Search strategy - selection of:
  - search terms & logic
  - possible fields, delimiters
  - controlled & uncontrolled vocabulary
  - variations in effectiveness tactics
- Reiterations from feedback
  - several feedback types: relevance feedback, magnitude feedback..
  - query expansion & modification

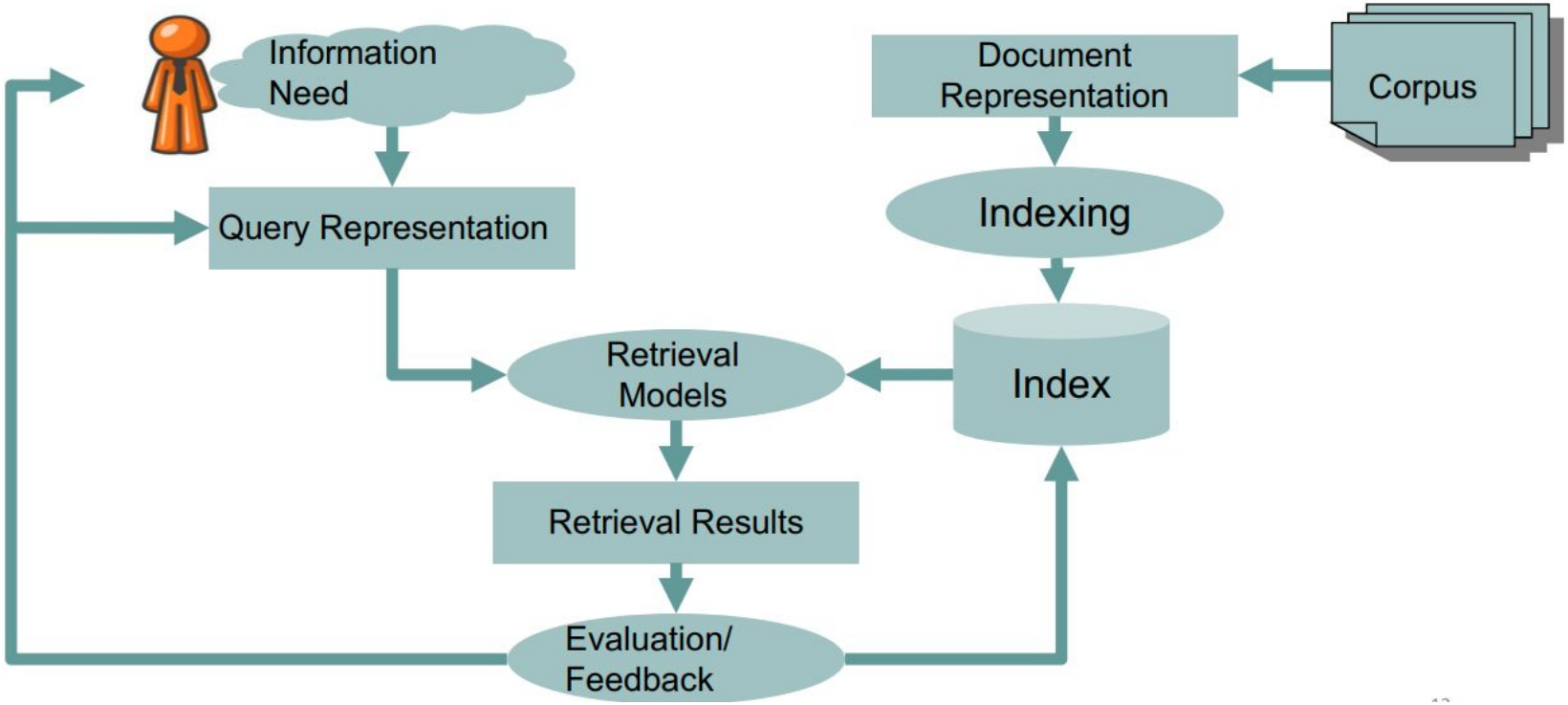
Matching - searching (Searching subsystem)

- Process of matching, comparing
  - search: what documents in the file match the query as stated?
- Various search algorithms:
  - exact match - Boolean
    - still available in most, if not all systems
  - best match - ranking by relevance
    - increasingly used e.g. on the web
  - hybrids incorporating both
    - e.g. Target, Rank in DIALOG
- Each has strengths, weaknesses
  - No ‘perfect’ method exists and probably never will

Retrieved documents -from system to user (IR Subsystem)

- Various order of output:
  - Last In First Out (LIFO); sorted
  - ranked by relevance
  - ranked by other characteristics
- Various forms of output
- When citations only: possible links to document delivery
- Base for relevance, utility evaluation by users
- Relevance feedback

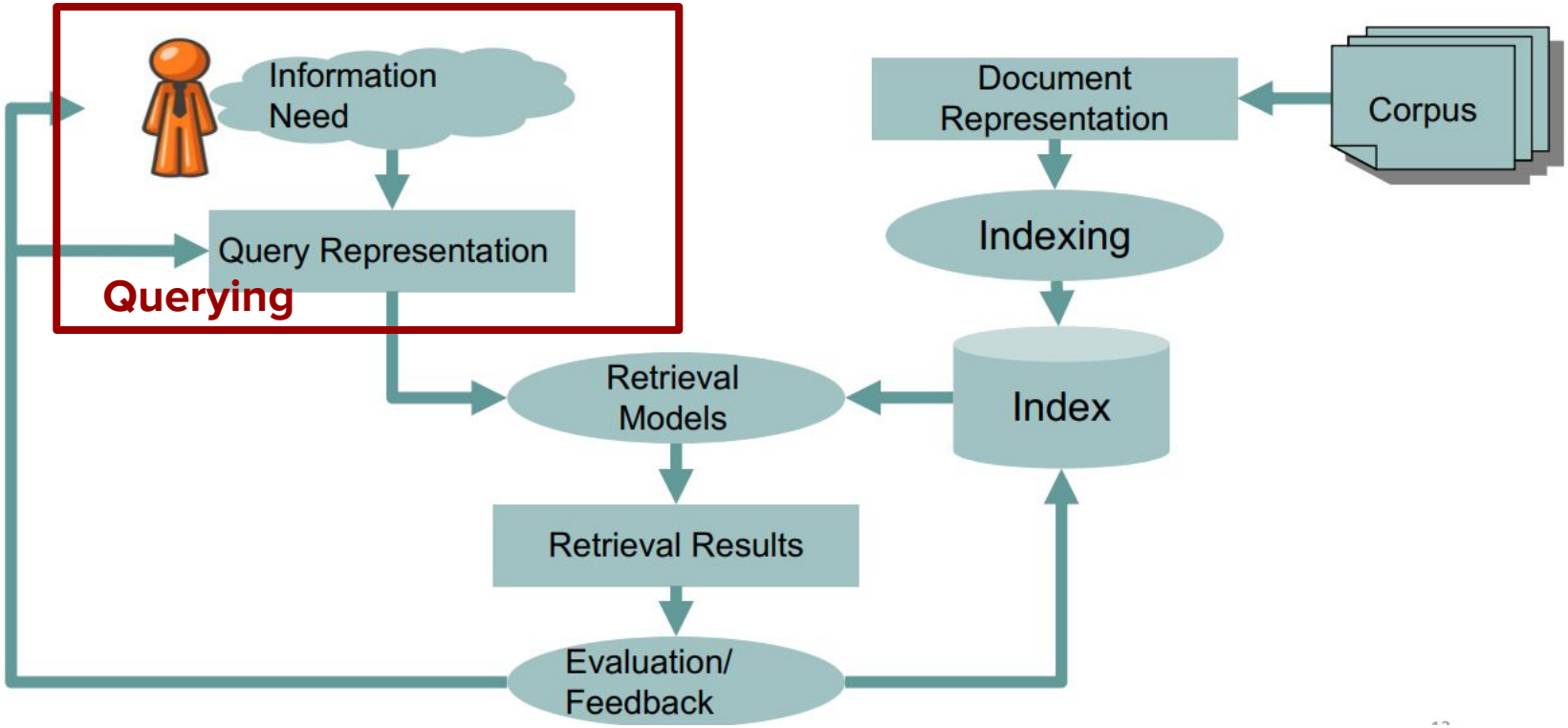
Process of Information Retrieval



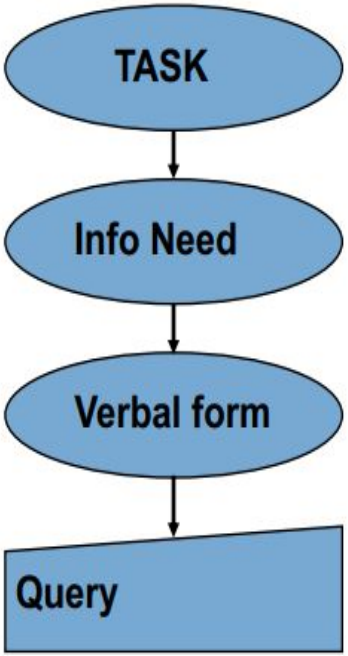
Basic Terminology

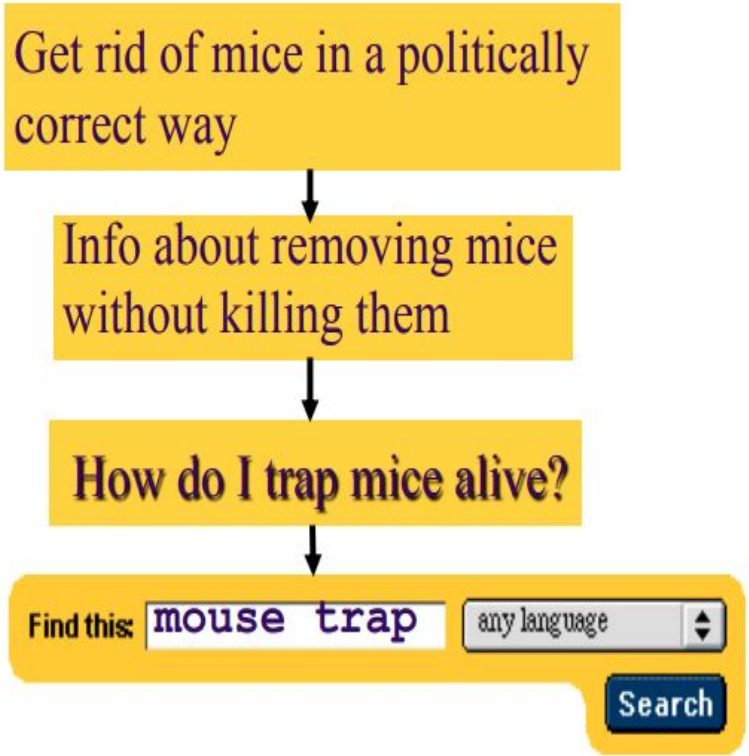
Query: text to represent an information need
Document: a returned item in the index
Term/token: a word, a phrase, an index unit
Vocabulary: set of the unique tokens
Corpus/Text collection
Index/database: index built for a corpus
Relevance feedback: judgment from human
Evaluation Metrics: how good is a search system?
    Eg.: Precision, Recall, F1

Process of Information Retrieval

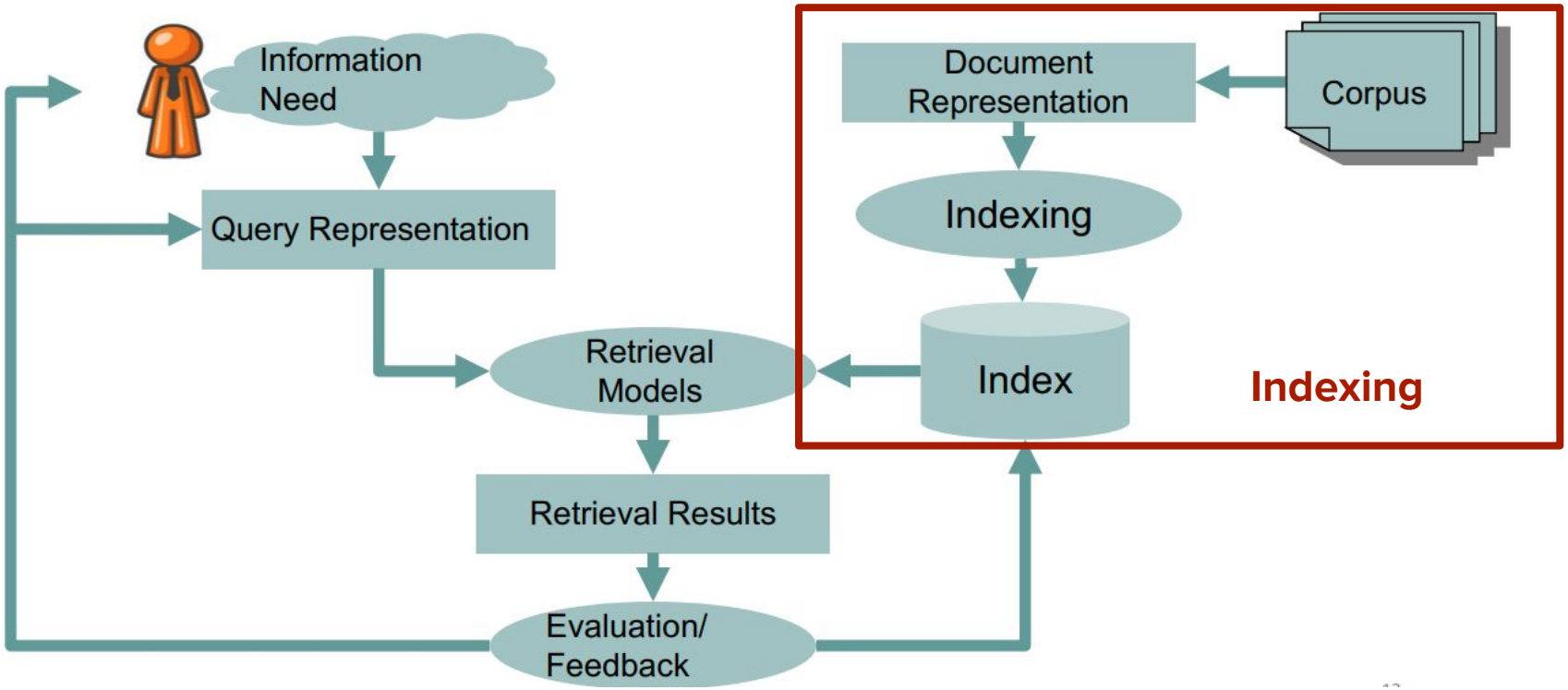


From Information Need to Query

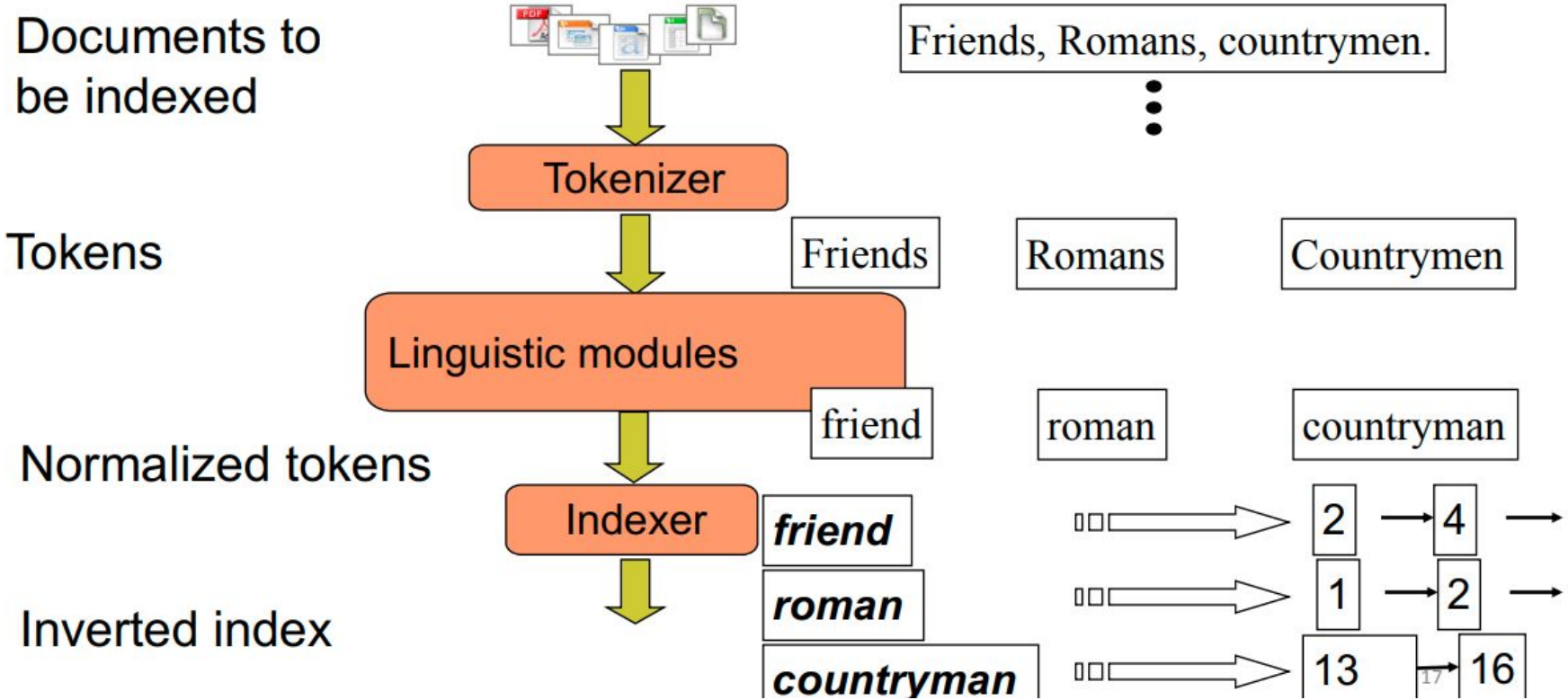




Process of Information Retrieval



Inverted index construction



An Index

- Sequence of (Normalized token, Document ID) pairs.

Doc 1

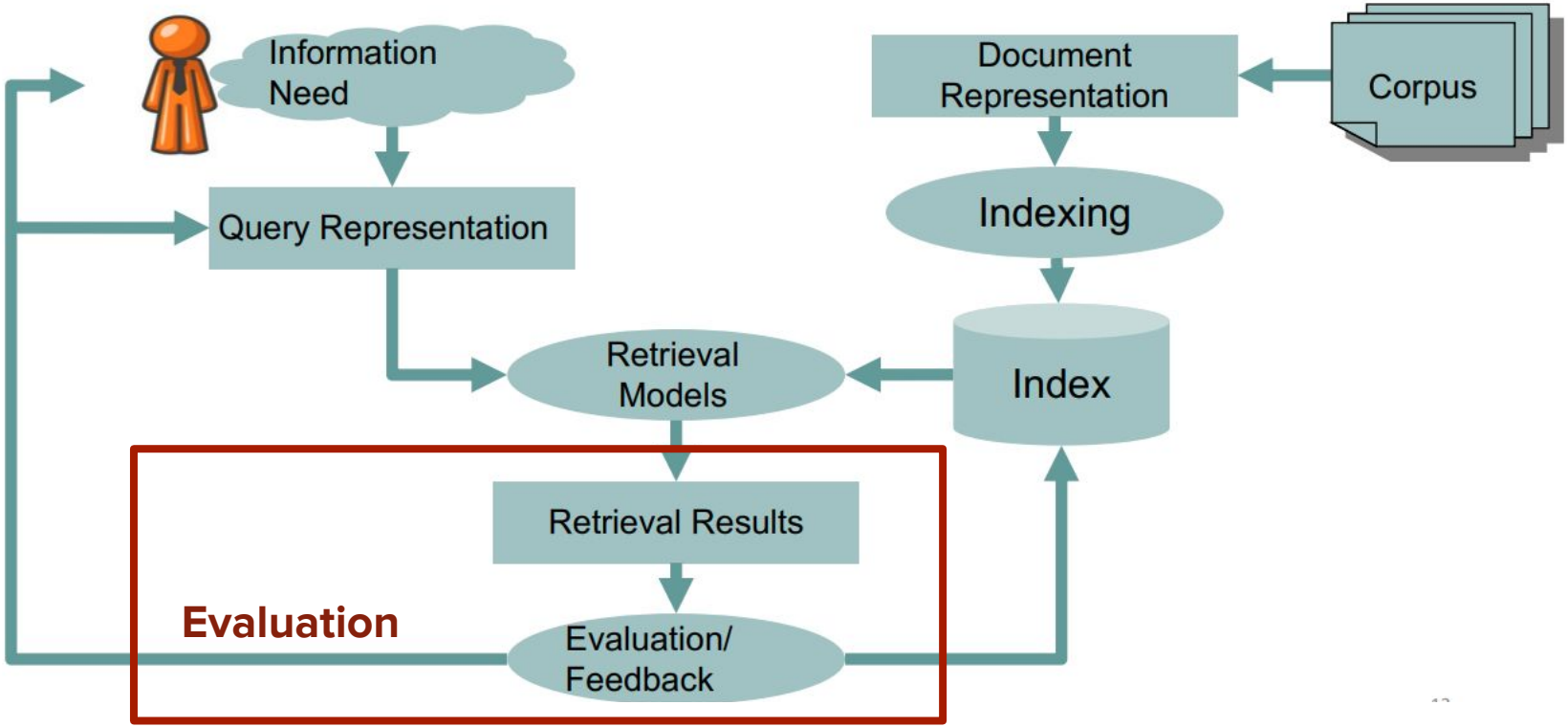
I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

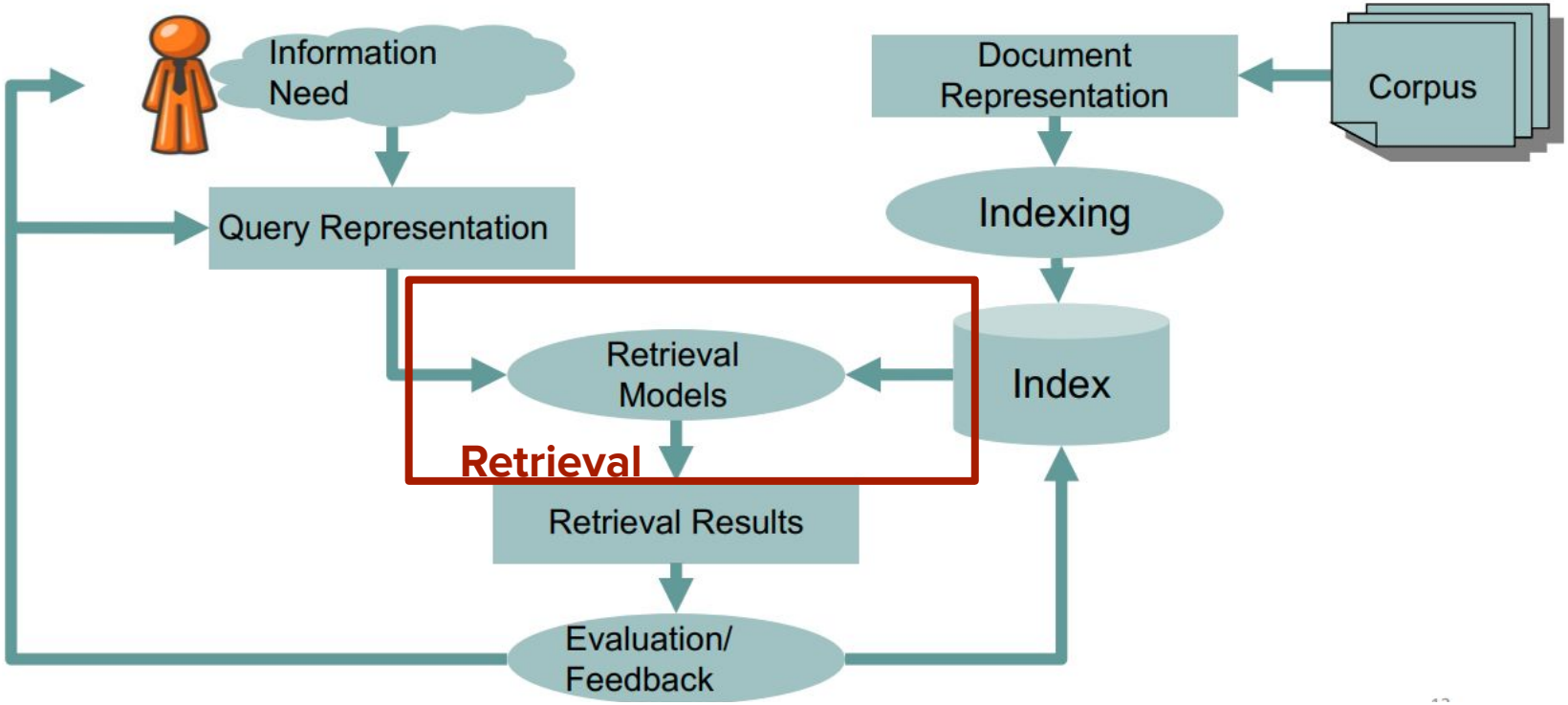
Process of Information Retrieval



Evaluation

- Implicit (clicks, time spent) vs. Explicit (yes/no, grades)
- Done by the same user or by a third party (TREC-style)
- Judgments can be binary (Yes/No) or graded
- Assuming ranked or not
- Dimensions under consideration
  - Relevance (Precision, nDCG)
  - Novelty/diversity
  - Usefulness
  - Effort/cost
  - Completeness/coverage (Recall)
  - Combinations of some of the above (F1), and many more
- Relevance is the main consideration. It means
  - If a document (a result) can satisfy the information need
  - If a document contains the answer to my query

Process of Information Retrieval



How to find relevant documents for a query?

- By keyword matching
  - boolean model
- By similarity
  - vector space model
- By imaging how to write out a query
  - how likely a query is written with this document in mind
  - generate with some randomness
  - query generation language model
- By trusting how other web pages think about the web page
  - pagerank, hits
- By trusting how other people find relevant documents for the same/similar query
  - Learning to rank

Ethical Considerations in IR

FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval

https://sci-hub.st/https://doi.org/10.1145/3458553.3458556

Three Big Issues in IR

1.Relevance

- It is the fundamental concept in IR.
- A relevant document contains the information that a person was looking for when she submitted a query to the search engine.
- There are many factors that go into a person’s decision as to whether a document is relevant.
- These factors must be taken into account when designing algorithms for comparing text and ranking documents.
- Simply comparing the text of a query with the text of a document and looking for an exact match, as might be done in a database system produces very poor results in terms of relevance.

To address the issue of relevance, retrieval models are used.

- A retrieval model is a formal representation of the process of matching a query and a document. It is the basis of the ranking algorithm that is used in a search engine to produce the ranked list of documents.
- A good retrieval model will find documents that are likely to be considered relevant by the person who submitted the query.
- The retrieval models used in IR typically model the statistical properties of text rather than the linguistic structure.
    - For example, the ranking algorithms are concerned with the counts of word occurrences than whether the word is a noun or an adjective.

2.Evaluation

Two of the evaluation measures are precision and recall.

Precision is the proportion of retrieved documents that are relevant.

Recall is the proportion of relevant documents that are retrieved.

Precision = (Relevant documents ∩ Retrieved documents) / Retrieved documents
Recall = (Relevant documents ∩ Retrieved documents) / Relevant documents

- When the recall measure is used, there is an assumption that all the relevant documents for a given query are known.
- Such an assumption is clearly problematic in a web search environment, but with smaller test collection of documents, this measure can be useful.
- It is not suitable for large volumes of log data.

3.Emphasis on users and their information needs

- The users of a search engine are the ultimate judges of quality.
  - This has led to numerous studies on how people interact with search engines and in particular, to the development of techniques to help people express their information needs.
- Text queries are often poor descriptions of what the user actually wants compared to the request to a database system, such as for the balance of a bank account.
- Despite their lack of specificity, one-word queries are very common in web search.
  - A one-word query such as “cats” could be a request for information on where to buy cats or for a description of the Cats.
- Techniques such as query suggestion, query expansion and relevance feedback use interaction and context to refine the initial query in order to produce better ranked results.

Main problems

- Document and query indexing
  - How to represent their contents?
  - Query evaluation
- To what extent does a document correspond to a query?
  - System evaluation
  - How good is a system?
  - Are the retrieved documents relevant? (precision)
  - Are all the relevant documents retrieved? (recall)

Queries???