nothing includes software solutions (examples)

intro

Classic workflow and typical tasks.

- Data Ingestion
  - Collect data
  - Describe data
  - Move data
- Data Transformation
  - Data modeling
  - Data migration
  - Pipeline orchestration
- Data Optimization
  - Selection, export, assessment

Information Retrieval distinguishes itself from Data Retrieval in two central aspects:

- Not limited to exact matching (e.g. multiple words, synonyms)
- Is ordered by relevance, i.e. importance to the user's query (a central concept in IR).

"Ad hoc" search is the most common type of search task.

- "Ad hoc" refers to an independent search episode to retrieve information for an isolated information need. Contrast with a search.

Vertical search focuses on domain-specific information such as news, travel, music, sports, academic papers, etc.

Enterprise Search is designed to allow search across an enterprise's content.

Desktop Search is a single-user system focused on searching the contents stored in a computer.

Data collection

Data

- is a measurement of something on a scale;
- a fact known by direct observation. Metadata
- is "data about data";
- not the content of data but data providing information about one or more aspects of the data, such as description (date, time, author), structure (format, version), administrative (permissions), legal, etc. Information
- is data with a context / meaning, thus enabling decision making;
- is data that has been processed, organized and structured

Life cycle of information: - Occurence - Transmission - Processing and Management - Usage

Value in Data

- Indirect value — data provides value by influencing of supporting decisions, e.g. risk analysis in insurance, purchase decisions in retail
- Direct value — data provides value by feeding automated systems, e.g. search system, product recommendation system

Data Stages

- Raw - focus on data discovery
- Refined - focus on data preparation for further exploration
- Production - focus is on integrating it into production processes or products

ETL pattern (recent evolution) - Extract Transform Load (!= ELT)

OSMN

- Obtain
- Scrub
- Explore
- Model
- Interpret

Data Preparation (/wrangling)

ad hoc search task:

- standard retrieval task in which the user specifies his information need through a query which initiates a search (executed by the information system) for documents which are likely to be relevant to the user.
    - Google search
    - Desktop search
    - Email search

includes:

- Understand what data is available;
- Choose what data to use and at what level of detail;
- Understand how to combine multiple sources of data;
- Deciding how to distill the results to a size and shape that enables the follow-up steps.

Common tasks:

- Cleaning
- Transformation
    - Normalization
    - Scaling values to the same range
    - Non-linear transformations
    - Discretization/binning
- Synthesis (create new attributes from existing data)
- Integration

        - Combine data that originally exists in multiple sources
        - Linking the corresponding records is a central step of many tasks of this "area"
- Reduction or selection
    - Data filtering
        - Used to remove data from the dataset
        - Can be used just to test with more manageable data portion
        - Operations are deterministic in nature
    - Data Sampling
        - Takes a random subset of the data items of a requested size (important to make sure the resulting sample is representative)
            - May need to analyze data distribution before and after
        - Non deterministic
    - Data aggregation
        - Grouping data via the aggregation operator (mean, median, min, max, percentile)
        - May be used to reduce ecessive detail

Requires a good understanding of data properties --> data visualization also used

- can be done before, during or after exploring properties of data

Data pipelines should be: --> should (really) be treated/viewed as software //// are software

- Reliable
- Scalable
- Maintainable

Makefiles

- used to automate software build processes, by defining targets and rules to execute
- abstraction of a dependency graph

Data documentation

- key element in various steps
- distinguishes between ad-hoc processes and repeatable, inspectable, shareable processes

Data Flow Diagrams (DFD):

- can be used to represent the flow of data from external entities into the system, show how data moves from one process to another, and data's logical storage
- squares - external entities
- rounded rectangles - processes
- arrows - data flows
- open-ended rectangles - data stores

Data Processing

Document model vs Relational model vs graph model vs triple-store model

- Data models central in data processing

Document model advantages

- good when dealing with 1-1 or 1-N relations
- Schema flexibility
- locality (related data stored together) --> better performance
- data model might be closer to the application's data structures

Graph model advantages

- good when dealing with many-to-many relation
- Not limited to homogeneous data

Relational model advantages

- locality (related data stored together) --> better performance
- data model might be closer to the application's data structures

Triple-Store model

- Information stored in 3-part statements - (subject, predicate, object)

Interaction types

- Online systems - services:
  - Waits for requests from a client. When it does, handle it ASAP and send response. Performance = response time
- Offline systems - batch processing:
  - Takes a large input of data, runs job and produces output. Best for long jobs or async processes
- Stream processing systems:
  - Operate on inputs and produce output. Result of event happening

Data visualization can be divided into exploration and explanation ends

Applied NLP for Information Retrieval

key NLP tasks for IR

- Tokenization
- Stemming and Lemmatization
- POS tagging
- NER
- Relation Extraction
- Sentiment analysis

IR

Table 1.1 Comparison of database systems and information retrieval, based on [40]

	Database systems	Information retrieval
Data type	Numbers, short strings	Text
Foundation	Algebraic/logic based	Probabilistic/statistics based
Search paradigm	Boolean retrieval	Ranked retrieval
Queries	Structured query languages	Free text queries
Evaluation criteria	Efficiency	Effectiveness (user satisfaction)
User	Programmer	Nontechnical person

SOLR

Textual fields go through a pipeline of analyzers, tokenizers, and filters.

- Analyzers, receive a a textual field as input and generates a token stream.
- Tokenizers, receive a character stream and produce a sequence of token objects.
- Filters, examine tokens and transform them (keep, discard, create, modify).

Analyzers can include one tokenizer and multiple filters.

ir evaluation

In order to evaluate relevance the information need must be clear.

Manifestations of Relevance

- ➔ **System relevance**: relation between a query and information objects (texts) in the file of a system as retrieved.
- ➔ **Topical relevance (aboutness)**: relation between the subject or topic expressed in a query, and topic or subject covered by retrieved texts
- ➔ **Cognitive relevance (pertinence)**: relation between the state of knowledge and cognitive information need of a user, and texts retrieved (e.g., cognitive correspondence, informativeness, novelty, information quality).
- ➔ **Situational relevance (utility)**: relation between the situation, task, or problem at hand, and texts retrieved (e.g., usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty).
- ➔ **Affective relevance (satisfaction)**: relation between the intents, goals, and motivations of a user, and texts retrieved (e.g., satisfaction, success, accomplishment)

Evaluation of Unranked Retrieval

- ➔ The two most frequent and basic measures for information retrieval effectiveness are Precision and Recall.
- ➔ Precision is the fraction of retrieved documents that are relevant.
  - ➔ Precision (P) = #(relevant items retrieved) / #(retrieved items)
- ➔ Recall is the fraction of relevant documents that are retrieved.
  - ➔ Recall (R) = #(relevant items retrieved) / #(relevant items)
- ➔ Precision and Recall are set-based measures.

P(q1) = 4 relevant documents retrieved / 6 documents retrieved = 0.67 R(q1) = 4 relevant docs retrieved / 8 existing relevant docs = 0.5

Precision = true positives / (true positives + false positives) Recall = true positives / (true positives + false negatives)

recall -> percentagem dos relevantes para conceito que estao presentes na lista precisao -> percentagem dos da lista que sao relevantes

Boolean model provides no ranking

Modeling in IR

- ➔ Modeling in IR aims at **producing a raking function**,
  - ➔ i.e. assign scores to documents with regard to a given query.
- ➔ The process consists of **two main tasks**:
  - ➔ Conception of a logical framework for representing documents and queries.
  - ➔ Definition of a ranking function that quantifies the similarities between documents and queries.

learning to rank

Ranking in Information Retrieval

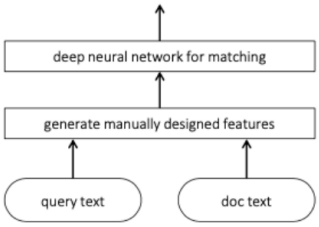
- ➔ Conventional document ranking models:
  - ➔ Query-dependent models:
    - ➔ Boolean model, set based model (no degree of relevance)
    - ➔ Vector space model, using term weighting such as TF-IDF.
    - ➔ Probabilistic models, such as language models and BM25.
  - ➔ Query-independent models:
    - ➔ Link-based models for the web, such as PageRank.

neural ir

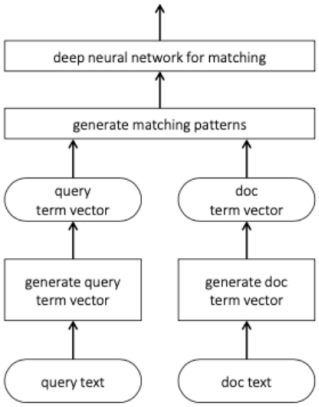
Concepts and Terminology (2)

- ➔ **Transformers**, neural network architecture with high impact in NLP problems; efficiently captures relationships between words.
  - ➔ **Encoder**: component in neural networks that processes input data to an encoded representation (e.g., dense vector).
  - ➔ **Decoder**: component in neural networks that generates output based on the encoded representation.
- ➔ **BERT** (Bidirectional Encoder Representations from Transformers), pre-trained language model built on transformer architecture.
- ➔ **Neural IR**, application of neural networks in the context of information retrieval tasks.
- ➔ **Semantic Search**, application of search techniques that leverage on understanding the meaning (semantics) of user queries and documents to improve relevance estimation.
- ➔ **Dense Retrieval**, use of dense vector representations for document retrieval and ranking.
- ➔ **Language Model**, learned computational model representing text; used in many NLP tasks.
- ➔ **Large Language Model**, language models that use a large number of parameters (i.e., millions or billions).

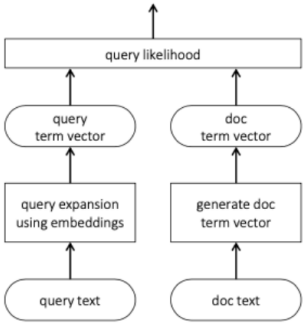
Document Ranking



(a) Learning to rank using manually designed features (*e.g.*, Liu (2009))



(b) Estimating relevance from patterns of exact matches (*e.g.*, (Guo *et al.*, 2016a; Mitra *et al.*, 2017a))



(d) Query expansion using neural embeddings (*e.g.*, (Roy *et al.*, 2016; Diaz *et al.*, 2016))

sparse vectors: 0 or 1 (words present)

word embeddings / dense vectors --> numerical value (categories)

- The vector of a word does not change with the other words used in a sentence around it.

transformers caputre context of text (relations between words), an can be used in word embeddings to further tune the understanding of them --> now each worth is influenced by the rest

From “exact match” to “soft match”

➔ **Key point**: neural methods replace exact matching with soft matching.

➔ With traditional methods, soft matching is possible:

- ➔ Stemming handles singular/plural and different tenses.
- ➔ Synonyms expand soft match possibilities.
- ➔ Query expansion techniques can add context to the query by adding new terms.

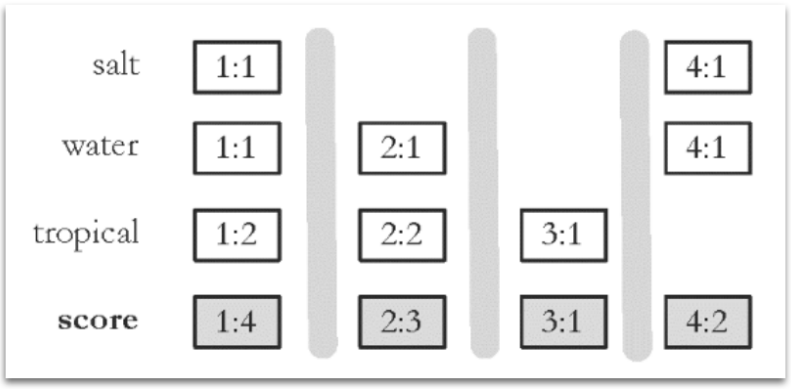
➔ With neural methods, soft match is central:

- ➔ Word embeddings capture semantic similarities, allowing nuanced understanding of related terms.

Document-at-a-Time

➔ query = [salt water tropical]

inverted lists



score is the sum of words in each document

each column represents a different document

1st step: compute all counts for the 1st document

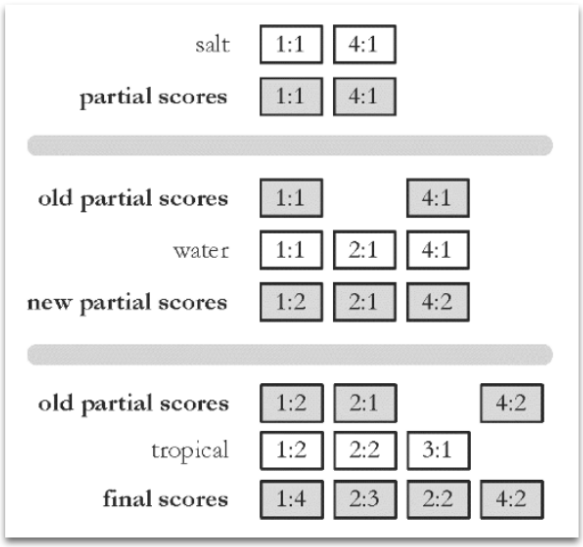
one document in each step

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

9

Term-at-a-Time

➔ query = [salt water tropical]



1st step: compute all counts for the 1st term [salt]

accumulators keep track of partial scores

one term is processed in each step

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

10

Pseudo Relevance Feedback

- ➔ Pseudo relevance feedback provides a method for automatic local analysis.
- ➔ It automates the manual part so that a relevance feedback algorithm is applied without extended user interaction.
- ➔ The method is applied to normal retrieval — assume that the top k ranked documents are relevant, and apply a relevance feedback algorithm under this assumption.

Implicit Relevance Feedback

- ➔ We can use indirect sources of evidence rather than explicit feedback on relevance.
- ➔ This is called implicit (relevance) feedback.
- ➔ Implicit feedback is less reliable than explicit feedback, but is more useful than pseudo relevance feedback, which contains no evidence of user judgements.
- ➔ Clickstreams are one of the main examples of indirect relevance information — clicks on links are assumed to indicate that the page was likely relevant for the query.

semantic search --> meaning and intent behind a user query

syntactic search --> structure and syntax of query

Entity-Oriented Search

- Entity-Oriented Search is the search paradigm of organizing and accessing information centered around entities, and their attributed and relationships.
- From a user perspective, entities are natural units for organizing information. Allowing users to interact with specific entities offers a richer and more effective user experience than what is provided by conventional document-based retrieval systems.
- From a machine perspective, entities allow for a better understanding of search queries, of document content, and even of users. Entities enable search engines to be more effective.

Knowledge Bases

- A knowledge base is comprised of a large set of assertions about the world.
- When the emphasis is on the relationships between the entities, the term knowledge graph is commonly used.
- Resource Description Framework (RDF) is a language designed to describe "things", which are referred to as resources.
- Each resource is assigned a Uniform Resource Identifier (URI), making it uniquely and globally identifiable.
- Each RDF statement is a triple, consisting of subject, predicate, and object components
  - Subject, always a URI, denoting a resource;
  - Predicate, always a URI, corresponding to a relationship or property of the subject resource;
  - Object, either a URI (referring to another resource) or a literal.