Kingdom of Saudi Arabia

Ministry of Education

Taibah University

Collage of Computer Science and Engineering

ASSIGNMENT 1:

Introduction To Information Retrieval PURPOSE

Course:

IS476 – Introduction to Information Retrieval

Student Name: Faisal Mohammed D Setaih

Student Identification Number: 4102546

Instructor: Dr. Mohammed Al-Sarem

Submission Date: 11/8/2023

Course Number: IS476

Section: M4

Table of contents

Introduction

Definition and goals of Information Retrieval

Information retrieval is a process in which people can deal and communicate with automated information retrieval systems such as spam filtering, or specialized computer servers such as search engines (Google, Bing, and Yahoo). The goal is to enable users to find relevant information from an organized collection of documents, it all depends on the user's query. So, the process starts as follows, firstly the user visits a search engine and enters a query in a text bar, then the search engine searches for documents that matches the user's query, after that it returns a large collection of documents that are relevant to the user need.

Components of an information retrieval system

There are three main components of information retrieval system / model:

1. User subsystem (query)

A user's subsystem/query is related to the user's desire and need.

2. Document subsystem (collection of documents)

A document subsystem is a selection of documents and other objects from different web resources, most of the documents are text based.

3. Retrieval subsystem (search engine)

A retrieval subsystem is the process of matching and comparing user's query with a collection of documents in a search engine's server using algorithms.

Most common Information retrieval models

There are 3 major information retrieval models (Boolean, vector space, and probabilistic)

1. Statistical model: following are two types of statistical retrieval approach. Both use statistical information of term frequencies to match and determine the relevance between query and documents.
   - Vector Space model: it represents the documents and queries as vectors in a multidimensional space, its dimensions are the terms that help to build an index to represent the documents. It has some disadvantages such as 'bag of words' that ignores the order of words and context.

- Probabilistic retrieval: it depends on the probability ranking principle, the IR system's job is to compare and then rank each document with their probability of relevance to the query. Its disadvantage is that it ignores frequency of words.

2. Boolean retrieval model: there are two types of Boolean retrieval, standard Boolean, and smart Boolean.
    - Standard Boolean: standard Boolean retrieval uses logical operators such as (AND / OR) it is very effective if a query requires a difficult and unambiguous selection among documents, in general, it is efficient and easy to implement. On the other hand, Boolean retrieval is not good for the majority of users because users are required to use natural language terms 'AND', 'OR', and 'NOT' these words have a different meaning when used in a query which will most likely show errors when they form a Boolean query. Another disadvantage is that the query either matches the documents or not "feast or famine".
    - Smart Boolean: smart Boolean has overcome some of the disadvantages of the standard Boolean. It became more user friendly and effective by asking users targeted questions to automatically modify the query. It does not require Boolean operators, instead it converts operator-free statement into ANDs of ORs automatically.
3. Linguistic and knowledge-based models: in any simple form of automatic text retrieval users enter a string of keywords (queries) and then the search engine will retrieve documents based on the presence or absence of that query. Obviously, this approach will skip many relevant documents because it does not consider the context and deep meaning of the user's query. Linguistic and knowledge-based models have been developed to solve this problem by performing a morphological, syntactic, and semantic analysis. In morphological analysis, roots and affixes are analyzed to determine the part of speech. And then complete phrases have to be transferred using syntactic analysis. Finally, the linguistic method resolves word ambiguities and generates relevant synonyms based on the semantic relationships between words.

Evaluation measures in information retrieval

Evaluation measures of IR system are algorithms used to evaluate and estimate how well an index, search engine or database retrieve results from a collection of documents that matches the user’s desire. There are two types of evaluation measures (Online-metrics, and Offline-metrics)

1. Online Metrics: this type of evaluation measure is actually created from search logs. It is used to determine the success of an A/B test which is a user experience research methodology that consists of a randomized experiment that involves two variants (A and B).
2. Offline Metrics: this one is based on relevance score; each document will be compared with the user’s query and then all documents will get a scale, for example from 0 to 5. There might be ambiguity in the query, for example homographs words (words that have same spelling but different meaning) like “close (near)” or “close (to shut)”.

Here are some equations used to measure the relevance score:

- Precision: precision is the result of dividing the documents retrieved that are relevant to the user’s information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

- Recall: recall is the result of dividing documents that are relevant to the retrieved query.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

- Fall-out: dividing non-relevant documents that are retrieved over all non-relevant documents.

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

- F-score: it is the result of multiplying 2 by precision and recall, divided by precision plus recall.

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

Other measures: precision at k, average precision, R-precision, etc.

Challenges of information retrieval systems

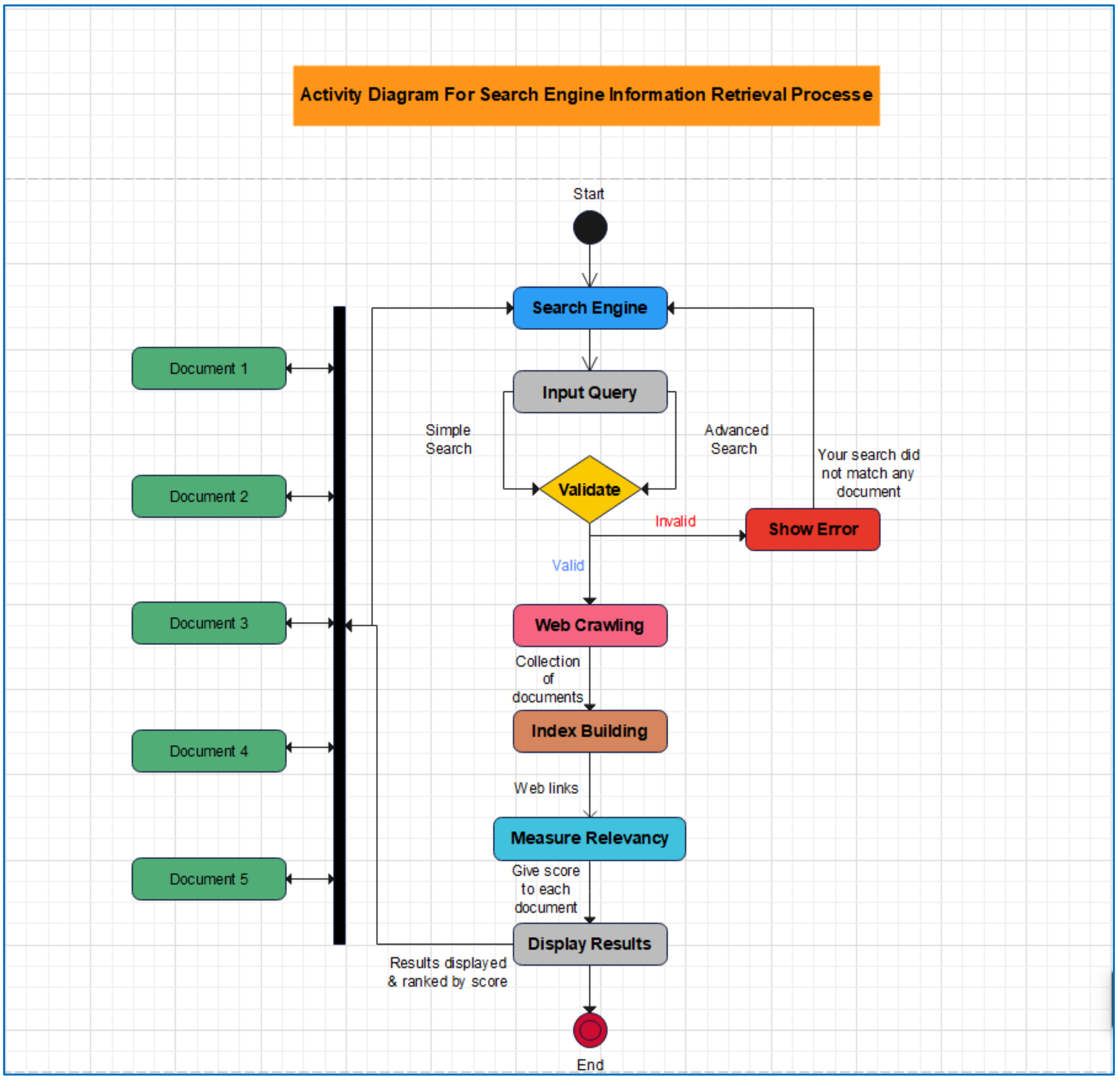
Information retrieval systems are still facing problems, following are some of them:

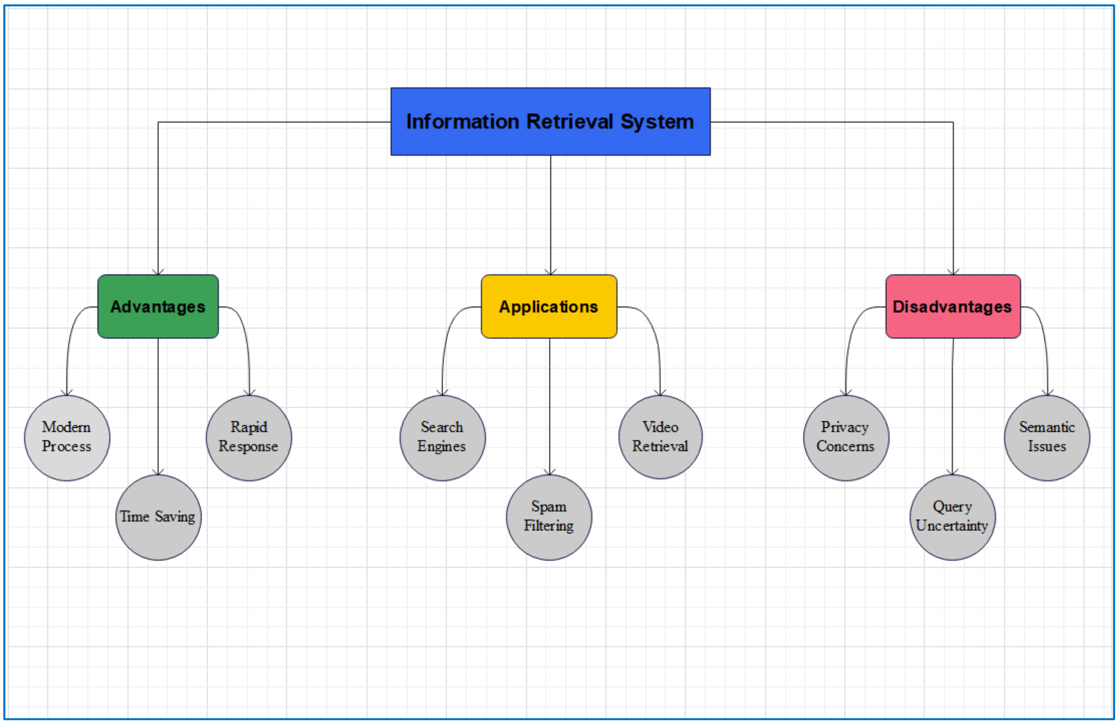
- Multimedia dimensionality: high dimensionality of multimedia characteristics (text, video, audio, image, animation, etc.) which makes the space very sparse, and the solution to that problem is using more training data to have accurate results.
- Semantic weak point: how to close the semantic gap between low-level features (color, texture, shape, object motion etc.) and high-level user’s information need.
- Search speed: efficiency and scalability are another challenge, we have huge data, how can we search between those data as fast as possible. A possible solution to this problem is enhancing the 3-tier architecture.

Future trends of information retrieval systems:

Information retrieval systems are very important to make sense of data. So, there are some future trends that can enhance the majority of IR systems.

- Privacy: privacy and intellectual property rights that prevent sharing data required for designing effective benchmarking systems.
- Feedback: relevance feedback allows the user to rank the retrieved results, and then the system (search engine) can use the feedback to improve the searching process.
- Conversational search: it is a collaborative technique that requires participants to engage in a conversation and perform a social search activity, and then the IR system will process the conversation and learn more with interactions and feedback from participants. This trend uses the semantic retrieval model plus naturel language processing.
- Overcoming the semantic gap: dealing with that gap between low-level and high-level features can enhance the Multimedia Information Retrieval Systems (MMIR) or search engines, question answering systems, etc. We can fill the semantic gap by analyzing the relativity between objects in images in case we are dealing with (MMIR).


Activity Diagram For Search Engine Information Retrieval Processe
Start
Search Engine
Input Query
Simple Search
Advanced Search
Validate
Your search did not match any document
Invalid
Show Error
Valid
Document 1
Document 2
Document 3
Document 4
Document 5
Web Crawling
Collection of documents
Index Building
Web links
Measure Relevancy
Give score to each document
Display Results
Results displayed & ranked by score
End


Information Retrieval System
Advantages
Applications
Disadvantages
Modern Process
Rapid Response
Time Saving
Search Engines
Video Retrieval
Spam Filtering
Privacy Concerns
Semantic Issues
Query Uncertainty

Recourses

1) Website Topic: Information Retrieval System Explained: Types, Comparison & Components
   Url: https://www.upgrad.com/blog/information-retrieval-system-explained/

2) Website Topic: Evaluation measures (information retrieval)
   Url:https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#:~:text=Precision,-Main%20article%3A%20Precision&text=Precision%20is%20the%20fraction%20of,all%20retrieved%20documents%20into%20account.

3) Website Topic: Information Retrieval Models
   Url: https://aspoerri.comminfo.rutgers.edu/InfoCrystal/Ch_2.html

4) Website Topic: Evaluation Measures in Information Retrieval
   Url: https://www.pinecone.io/learn/offline-evaluation/

5) Website Topic: Multimodal Information Retrieval: Challenges and Future Trends
   Url:https://www.researchgate.net/publication/255686190_Multimodal_Information_Retrieval_Challenges_and_Future_Trends

6) Website Topic: History of IR
   Url: https://www.stannescet.ac.in/cms/staff/qbank/CSE/Notes/CS6007-INFORMATION%20RETRIEVAL-1428610647-UNIT%20I%20IR%20Final.pdf

7) eBook:
   Url:
   https://books.google.com.sa/books?hl=en&lr=&id=cN4qDgAAQBAJ&oi=fnd&pg=PR11&dq=introduction+to+modern+information+retrieval+baeza+yates&ots=rEo6rq40ip&sig=hFVq8cxNLny7xOcbuFew8HFh00s&redir_esc=y#v=onepage&q&f=false