

A Review on recent research in Information Retrieval

Group - 12

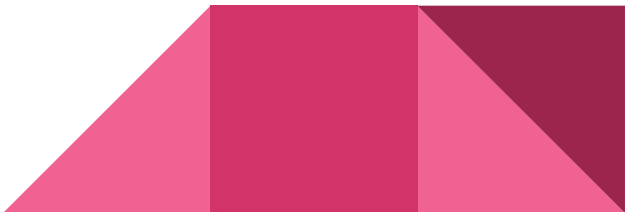
Presentation by :

Group Members:
66: Diya theryani
57: Anjali punsi
31: Ronak karia





Table of Content

- **Introduction of Information Retrieval (IR)**
 - **IR Notion**
 - **IR Components**
 - **IR Challenges**
 - **IR applications**
 - **Processing Text in IR**
 - **Literature Survey**
 - **Conclusion**
- 

Introduction

The expansion of data has been seen to have expanded dramatically in recent years, which typically contains different types of data such as

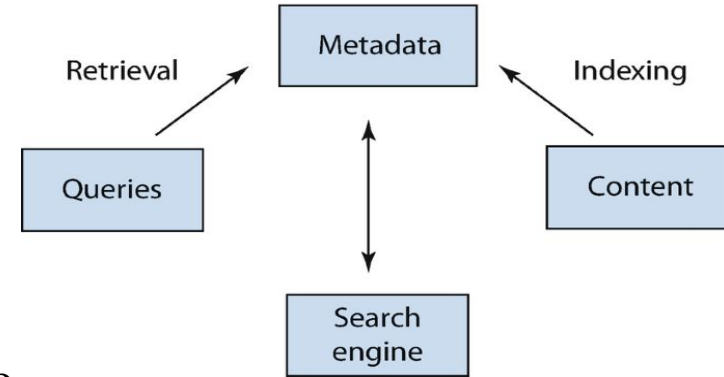
- structured
- semi-structured
- unstructured
- heterogeneous data

Therefore, an **Information Retrieval System** is used to search for information in documents, search for documents themselves and also searching for the metadata that describes data, and for databases of texts, images or sounds.

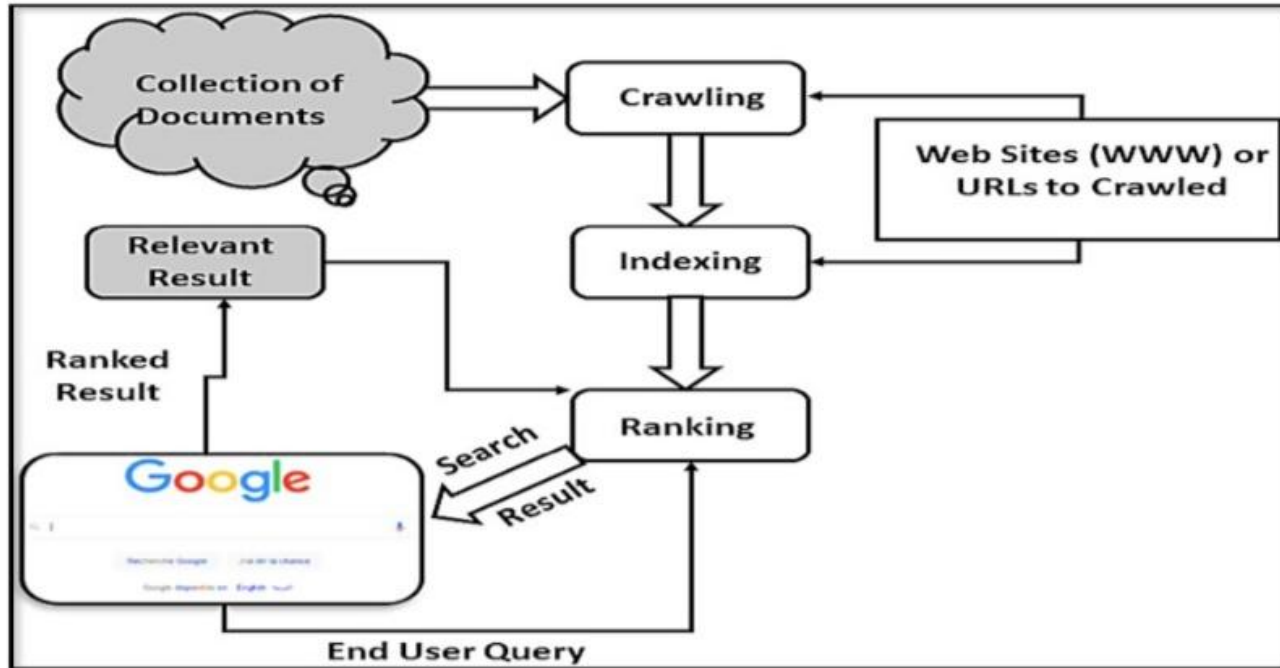


IR Notion

- In today's world, the situation of a user having a need for information, translating that need into a search phrase, and then executing that search to find the information has become commonplace.
- Google is the most well-known example of an information retrieval system that everyone has used it.
- Information retrieval (IR) is the procedure of representing, storing, and searching a collection of big data for the goal of extracting knowledge and access to find relevant results that satisfy the user's needs as a reaction to a user's query.



IR Components



- **Crawler:** The crawler component is in charge of seeking and retrieving documents for the search engine. Crawlers come in a variety of shapes and sizes.
- **Indexing:** The technique of representing documents is referred to as indexing. Basically, it means that the system creates a document index. In this phase, the user writes a query in order to retrieve relevant information.
- **Ranking:** After that, the system searches the index for documents that are relevant and pertinent to the query and presents them to the user, and that's what we call ranking.

The last step is where the users can provide the search engine with relevant feedback.



IR Challenges

- The mismatch between how a user conveys the information they are seeking for and how the author of the item expresses the information he is delivering is the main challenge in information retrieval. In other words, the problem is a mismatch between the user's vocabulary (language) and the author's vocabulary (language).
- Besides, there are barriers to specifying the information a user requires due to limitations in the user's capability to explain what information is required.
- Uncertainties and ambiguities in languages are also one of the challenges that a user can face.



IR Applications

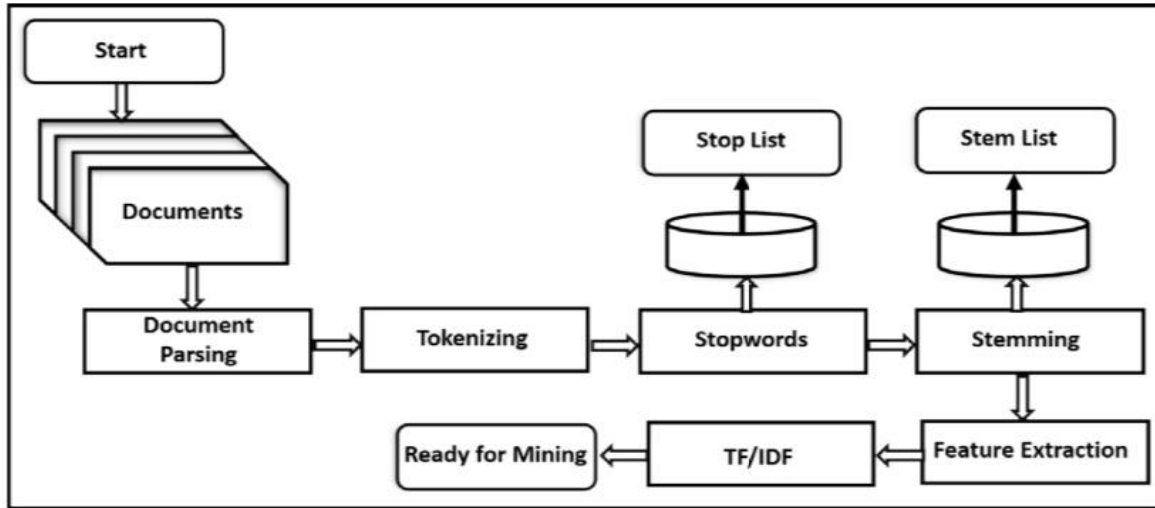
Information retrieval (IR) systems were first created to aid in the management of massive amounts of data. Today, information retrieval is employed in a variety of applications. The following are some examples of IR system applications:

- Media Search
- Search Engines
- Digital Library
- Businesses



Processing Text in IR

The preprocessing procedure is extremely significant and crucial. It's the initial phase in the process of text mining.



1. **Documents Parsing:** It is the process of identifying the content and form of text documents that are written and presented in a variety of languages, character sets, and forms.
2. **Tokenizing:** Tokenization or lexical analysis is the operation of creating words from a series of letters (characters) in a document.
3. **Stopwords:** The aim of this phase is to remove all the useless and insignificant common terms from the tokens streams such as articles, prepositions etc.
4. **Stemming:** The goal of this stemming is to eliminate multiple suffixes, minimize the number of words, ensure that stems are precisely matched, and save memory space and time.
5. **Feature Extraction:** The technique of removing extraneous and superfluous characteristics from a dataset is known as feature extraction. It is used for document classification. One method for extracting features is to use the Term Frequency Inverse Document Frequency (TF-IDF) weighted scheme.

6. Information Retrieval models: The classical models which are the boolean and vector space model (VSM). The Boolean offers the precise match, use the logic operators. VSM is applied to compute the similarity distance among document and query that are represented as vectors, by using different measures like "Cosine similarity", "Jaccard measure", "scalar product" and "Measurement Dice". We can also find the probabilistic model which depends on the probability of the pertinent of a given query in order to rank them.

7. Evaluation: The ability of an IR system to return pertinent documents, as well as the accuracy and precision of these retrieved documents, is commonly measured. There are two different criteria for judging the quality of IR system. **Precision** =

$$\frac{|relevantdocuments \cap retrieveddocuments|}{|Retrieveddocuments|}$$

$$\text{Recall} = \frac{|relevantdocuments \cap retrieveddocuments|}{|Relevantdocuments|}$$

$$F - measure = \frac{2 * precision * recall}{precision + recall}$$

Literature Survey


Approach	Author	Feature Extraction	Corpus	Validation measure	Limitation
Rocchio Algorithm	B.J. Sowmya et al [31]	TF-IDF	Wikipedia	F1-Macro	Only retrieves a few pertinent documents from hierarchical data sets
SVM and KNN	K. Chen et al. [32]	TF-IDF	20 Newsgroups and Reuters-21578	F1-Macro	Polysemy is not captured, and semantics and sentatics are still unsolved
Naïve Bayes	Kim, S.B et al. [33]	Weights words	Reuters-21578	F1-Macro	This strategy relies on a strong presumption about the data distribution's form

Conclusion

- Therefore, complete information about Information Retrieval was studied.
- We discussed the different phases of pre-processing techniques that a document goes through which helps to retrieve efficient results.
- We also compared some of the existed approaches such as KNN, SVM, Naïve Bayes, and Rocchio Algorithm.
- In future work, we will study and compare more techniques used in information retrieval.



References

- [1] W. Bruce Croft, Donald Metzler and Trevor Strohman.(2015) "Search Engines Information Retrieval in Practice", Pearson Education, Inc.
 - [2] G. Kowalski.(2011) "Information Retrieval Architecture and Algorithms", Springer (eds) , Boston, MA.
 - [3] Lal, N., Qamar, S., and Shiwani, S.(2016) "Information retrieval system and challenges with dataspace." International Journal of Computer Applications 147 (8).
 - [4] Roshdi, A., and Roohparvar, A.(2015) "Information retrieval techniques and applications." International Journal of Computer Networks and Communications Security 3 (9): 373–377
- 

Thank you !!

