Introduction to Information Retrieval

William Scott
Feb 3, 2019 · 3 min read

This is a series on Information Retrieval techniques with implementation basic concepts and easily understandable examples.

For those who are highly interested, i suggest the book “Introduction to Information Retrieval” book by Manning

> ***Click here*** *to checkout the git repo.*

Information Retrieval Series:

- 1. Introduction
- 2. Unigram Indexing & Positional Indexing
- 3. TF-IDF
- More to come…

. . .

What is Information Retrieval?

Information Retrieval, just as the name suggests is retrieval of information. What we basically do in this is refine the retrieval of information just so that we can satisfy an information need.

so, we can sum up information retrieval as

> Finding relevant materials to satisfy an information need.

Few points to be noted here is that when we say we want to find **materials** we basically mean **documents**, specifically **text documents**. And the text in those documents are highly unstructured. If you are still not sure what a text document could be, just think of it as a website, for the time being.

> *So what an IR system does is, it takes the query from user, understands it, searches it in its corpus and sends the results of the relevant documents.*

Why cant we do ctrl+f?

we are saying that we want to find and find, so why not just build a program to search for a query, if it exists in document or not? we could do that but it need not work. the reasons are as follows

- **Synonyms:** There are many words which have alternative words. for example, when a user is trying to get a haircut, his search query could be "salon" or "barber". we cannot just show him the documents which have barber and which doesn't have salon. because they both mean the same thing. More general examples are Mom — Mother, Hat — Cap.

- **Homographs:** These are the words which have the same spelling but have different meaning in different sentences. we do not basically deal with the pronunciation of words here. Lie — can be lying on bed, or lying to another person. tear — could be tearing a paper, or having tears (as in crying). Apple — Could be a company or a fruit.

so due to these above problems, we need to build an intelligent IR model which can understand the query of the user and give the relevant documents. do not worry about the above problems, we will basically deal with them later, just as a gist, we deal with this by going through a important stage called, preprocessing, where the information is turned into a more general form which can help us relate the words much better.

Intelligent IR

When we are trying to retrieve relevant documents, we need to first define relevance. are we going to retrieve the latest documents? or are we going to retrieve the documents which match the subject?

An Intelligent IR model do not just depend on one factor to find out relevance, **metadata**, **authoritativeness**, type of **information need**, **meaning** of the query, meaning of the sentence in the document and many such factors are considered.

Basic Terminology:

- **Collection / Corpus:** collection of documents
- **Query:** Information need
- **Token:** An individual entity as word / set of characters
- **Rank:** Relevance of a document in some measure.

Information Retrieval Series:

- 1. Introduction
- 2. Unigram Indexing & Positional Indexing
- 3. TF-IDF
- More to come…

Resources:

Introduction to Information Retrieval — Manning

Search Information Retrieval Artificial Intelligence NLP Introduction