Information Retrieval Systems (IRS)

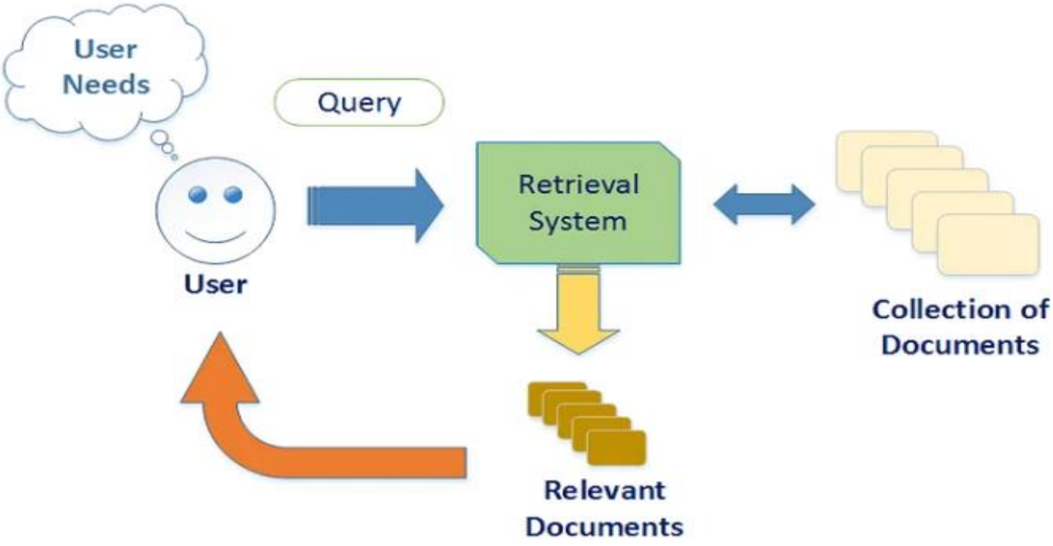
There is an urgent requirement of effective Information Retrieval Systems(IRS) has aroused with the quick growing development of the Internet, and plethora of availability of online text based information. The goal of an IR System is to retrieve relevant information regarding a user's query.

Information retrieval (IR) aims at defining systems able to provide a fast and effective content-based access to a large amount of stored information. The aim of an IR system is to estimate the relevance of documents to users' information needs, expressed by means of a query. This is a very difficult and complex task, since it is pervaded with imprecision and uncertainty. Most of the existing IR systems offer a very simple model of IR, which privileges efficiency at the expense of effectiveness. A promising direction to increase the effectiveness of IR is to model the concept of "partially intrinsic" in the IR process and to make the systems adaptive, i.e. able to "learn" the user's concept of relevance. To this aim, the application of soft computing techniques can be of help to obtain greater flexibility in IR systems.

Components of Information Retrieval Model

Here are the prerequisites for an IR model:

1. An automated or manually-operated indexing system used to index and search techniques and procedures.
2. A collection of documents in any one of the following formats: text, image or multimedia.
3. A set of queries that serve as the input to a system, via a human or machine.
4. An evaluation metric to measure or evaluate a system’s effectiveness (for instance, precision and recall). For instance, to ensure how useful the information displayed to the user is.



The **key idea behind question answering and search systems powered by Neural Networks is clustering in a high-dimensional space**. Typically, such clustering is coupled with a data structure to imbue elements of persistence, hierarchy, and organization into the data. Clustering also helps to some extent in dealing with real-world data which typically do not come from the same distribution. You may have noticed that this is very similar to algorithms like Kmeans, SVM, or GMM and that's exactly the idea. The only difference here is that Neural Network clustering allows for semantic understanding beyond just adjustable weights. However before we can do any clustering, we will need to **represent the data in a multi-dimensional space**.

A traditional approach is to represent a collection of texts as a bag of words, where each passage is a row of numbers and each number indicates the frequency of occurrence for a unique word. This works well in practice but has its own drawbacks such as vector sparsity, the curse of dimensionality, and a limited ability to capture the semantic meaning of a text.

The neural network approach is similar where vectors of continuous random digits are used to represent text. When fed to a model like Bert, the self-attention mechanisms allow for the model to capture the semantic meaning between the words. Specifically, **the neural network will map texts into a common vector space wherein similarity measures such as dot product is then used to measure similarity between the vectors**. To produce learned vectors that can be clustered, such neural networks are typically trained to maximize the similarity between relevant pairs.

Hence, we want to **learn a relevance function (denoted by h) that outputs high similarity values for passages (denoted as p) that are relevant to a query (q) and low otherwise**. This will be achieved via a pair of neural networks denoted as E_q and E_p, defined as the multiplication of the transposed output of E_q with that of E_p:

$$h(q,p) = E_Q(q)^T E_P(p)$$

Additionally, as with every neural network, we need to optimize a function. We do so by minimizing the subsequent loss and finding a set of weights defined as theta:

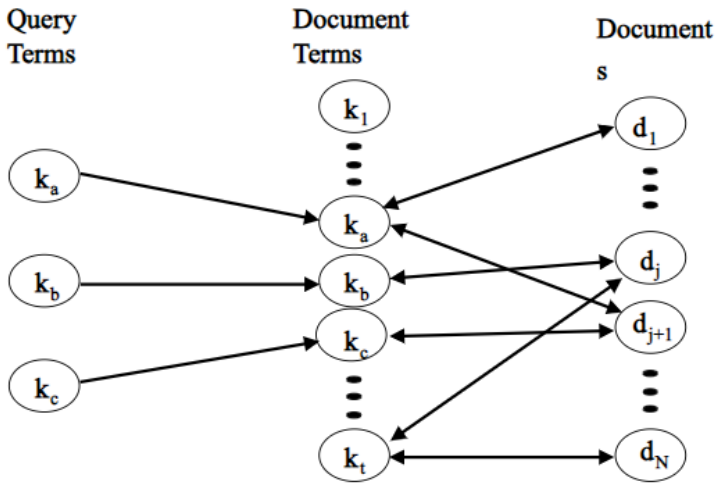
$$\mathcal{L}_\theta = -\log \frac{e^{h(q_i, c_i^+)}}{e^{h(q_i, c_i^+)} + \sum_{j=1}^{n} e^{h(q_i, c_{i,j}^-)}}$$

This loss function is Cross-Entropy and it will cause the network to bring texts that are relevant (higher similarity value) nearer together when visualized in a 3-dimensional space. It does so by maximizing the positive part which is seen in the numerator. While those incorrect pairs captured by the summed term in the denominator will cause there to be at least some distance further than the existing distance difference between a positive pair. The following images illustrate that:

Other Vector Model: Neural Network

- Basic idea:
  - 3 layer neural net: query terms, document terms, documents
  - Signal propagation based on classic similarity computation
  - Tune weights.

Neural Network Diagram



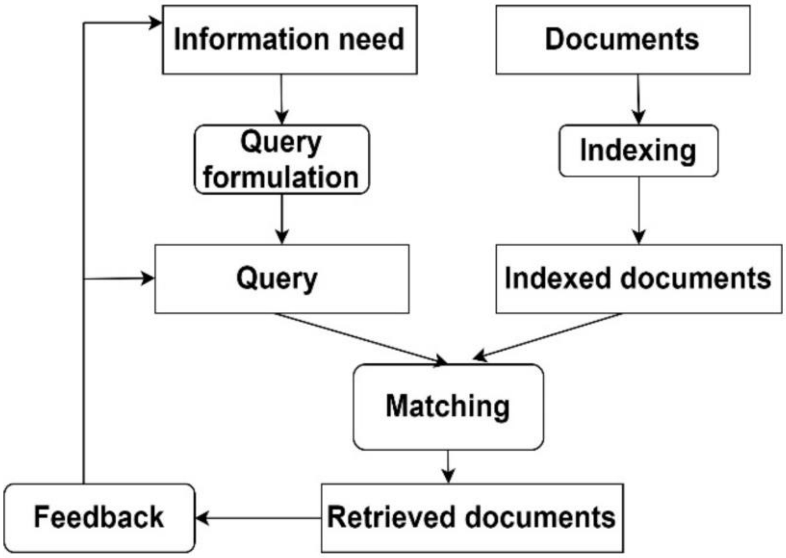
Computing Document Rank

- Weight from query to document term
  - Wiq = $\frac{\text{wiq}}{\text{sqrt} (\Sigma_i \ \text{wiq} \ )}$
- Weight from document term to document
  - Wij = $\frac{\text{wij}}{\text{sqrt} (\Sigma_i \ \text{wij} \ )}$

For Fuzzy Logic visit link: https://pchats.tripod.com/istebhaskar.pdf

Genetic Algorithm:

Retrieval of relevant documents from a collection is a tedious task. As Genetic Algorithms (GA) are robust and efficient search and optimization techniques, they can be used to search the huge document search space. In this paper, a general frame work of information retrieval system is discussed. The applicability of Genetic algorithms in the field of information retrieval is also discussed. A review on how GA is applied to different problem domains in information retrieval is presented.

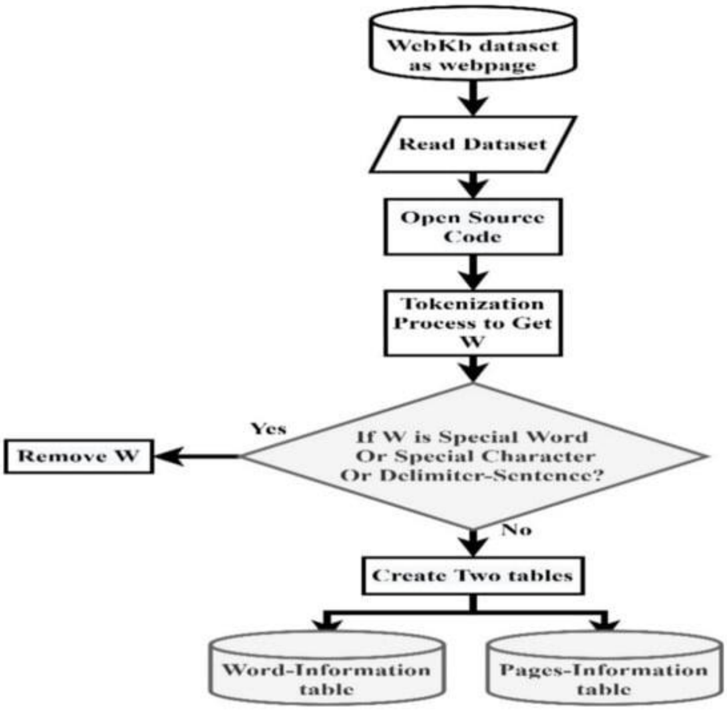


The user and system communicate with each other using respective queries, retrieving the set of documents. The most natural form of communication is used to communicate with each other for the information needed; such a natural communication method is called a request. In the automatic query, it takes the input as a request and gives the output as the initial query. Based on the initial query, some or all words in the request are converted to query terms by a trivial algorithm. Relevance feedback inputs the initial query to some retrieved relevant or irrelevant documents to output a successive query. The next subsections describe the IRS' techniques and methods.

The proposed IRS includes three stages: stage 1 represents an advanced document indexing method (ADIM) used to prepare the WebKb dataset to maintain the high performance of IRS (the dataset requires preprocessing in order for the algorithm to work efficiently on accurate and reliable data), the second stage is query search processing, and the final stage is the evolutionary algorithms (i.e., genetic algorithm and culture algorithm) as an integration of two of the ML techniques. GAs are used to solve an optimization problem and are applied to reduce the overhead during the classification. CAs are also used for classification by applying the evaluation stage (i.e., precision, recall, and accuracy). The next subsections explain these stages.

Advanced Document Indexing Method (ADIM) Stage

This stage is composed of two main steps: WebKb dataset reading and ADIM. The main steps of ADMI are shown in the flowchart in **Figure**


WebKb dataset as webpage
Read Dataset
Open Source Code
Tokenization Process to Get W
If W is Special Word Or Special Character Or Delimiter-Sentence?
Yes
Remove W
No
Create Two tables
Word-Information table
Pages-Information table