**Unit II. IR Models**

Textbook:

<https://srikarthiks.files.wordpress.com/2016/07/t2-modern-information-retrieval.pdf>

**Introduction:**

**Taxonomy of Information Retrieval Models:**

**Retrieval: Ad hoc and Filtering:**

**A Formal Characterization of IR models:**

**Refer Textbook for the above topics:**<https://srikarthiks.files.wordpress.com/2016/07/t2-modern-information-retrieval.pdf>

**Classic Information Retrieval Models:**

The classical IR systems are designed **based on mathematical concepts** and are the most widely used, simplest, and easy-to-implement systems for information retrieval models.

In this system, the retrieval of information depends on documents containing the **defined set of queries and there is no ranking or grading** of any kind.

Classic Information Retrieval models can be **easily implemented** and updated accordingly.

The different classical IR models are based on **mathematical knowledge** that was easily recognized and understood as well and take concepts like Document Representation, Query representation, and Retrieval / Matching function into account in their modeling.

The term classic in the name of classical IR systems denotes that they **use foundational techniques** for text documents without extra information about the structure or content of a document.

The different types of classical Information Retrieval models are: boolean models, Vector space models, and Probabilistic IR models.

**1. Boolean Model:**

The Boolean model in information retrieval is **based on the set theory and boolean algebra**. We can pose any query in the form of a Boolean expression of terms where the terms are logically combined using the **Boolean operators AND, OR, and NOT** in the Boolean retrieval model.

* Using the Boolean operators, the **terms** in the query and the concerned documents can be **combined** to form a whole new set of documents.
  + The Boolean **AND** of two logical statements x and y means that **both x AND y must be satisfied** and will be a set of documents that will **smaller or equal** to the document set.
  + The Boolean **OR** of these same two statements means that **at least one of these statements** must be satisfied and will fetch a set of documents that will be **greater or equal** to the document set otherwise.
  + **Any number** of logical statements can be **combined** using the three Boolean operators.
* The queries are designed as boolean expressions which have precise semantics and the retrieval strategy is based on **binary decision criterion**.
* The Boolean model can also be explained well by **mapping the terms in the query with a set of documents**.

The most famous web search engine in recent times Google also ranks the web page result set based on a two-stage system: In the **first** step, a **Simple Boolean Retrieval** model, returns matching documents in no particular order, and in the next step **ranking** is done according to some **estimator of relevance**.

### Aspects of Boolean Information Retrieval Model

**Indexing:** Indexing is one of the **core functionalities** of the information retrieval models and the **first step** in building an IR system assisting with the efficient retrieval of information.

* Indexing is majorly an **offline operation** that collects data about **which words occur in the text corpus** so that at search time we only have to access the **pre-compiled index** done beforehand.
* The boolean model builds the indices for the terms in the query considering that **index terms are present or absent** in a document.

**Term-Document Incidence matrix:** This is one of the basic **mathematical models to represent text data** and can be used to answer Boolean expression queries using the Boolean Retrieval Model. It can be **used to answer any query** as a Boolean expression.

* It views the document as the **set of terms and creates the indexing** required for the Boolean retrieval model.
* The text data is represented in the form of a **matrix** where **rows** of the matrix represent the sentences and the **columns** of the matrix represent the word for the data which needs to be analyzed and retrieved and the **values** of the matrix represent the number of occurrences of the words.
* This model has **good precision** as the documents are retrieved if the condition is matched but, it **doesn't scale well** with the size of the corpus, and an inverted index can be used as a good alternative method.

**Processing the data for Boolean retrieval model**

* We should **strip** unwanted characters/markup like HTML tags, punctuation marks, numbers, etc. before breaking the corpus into tokens/keywords on whitespace.
* **Stemming** needs to be done and then common stopwords are to be removed depending on the application need
* The term document incidence matrix or **inverted index** (with the keyword a list of docs containing it) is built.
* Then the common queries/phrases may be detected using a **domain-specific dictionary** if needed.

### Advantages of Boolean Model

* It is **easy** to implement and it is computationally **efficient**. Hence, it is the **standard** model for the current large-scale, operational retrieval systems and many of the major online information services use it.
* It enables users to **express structural and conceptual constraints** to describe important linguistic features. Users find that synonym specifications (reflected by OR-clauses) and phrases (represented by proximity relations) are useful in the formulation of queries
* The Boolean approach possesses a **great expressive power** and clarity. Boolean retrieval is **very effective** if a query requires an exhaustive and unambiguous selection.
* The Boolean method offers a **multitude of techniques to broaden or narrow down a query**.
* The Boolean approach can be especially **effective** in the **later stages of the search process**, because of the clarity and exactness with which relationships between concepts can be represented.

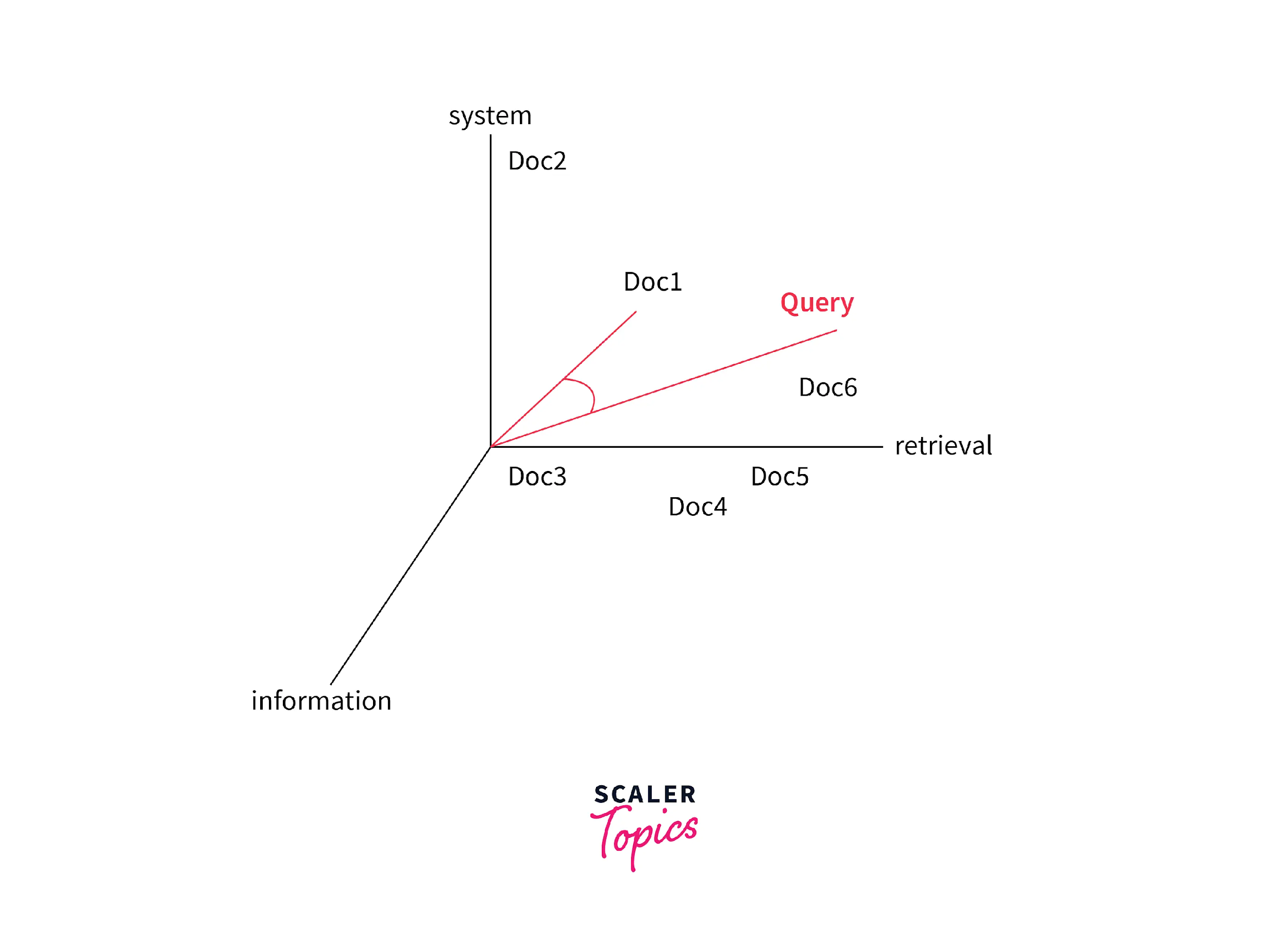
### Shortcomings of Standard Boolean Approach

* Users find it **difficult to construct effective Boolean queries** for several reasons. Users are using the natural language terms AND, OR, or NOT that have different meanings when used in a query.
* Hence the users will **make errors** when they form a Boolean query because they resort to their own knowledge of English.

**2. Vector Model:**

Also called term vector models, the vector space model is an **algebraic model for representing text document**s (or also many kinds of multimedia objects in general) as vectors of identifiers such as index terms.

The vector space model is based on the **notion of similarity between the search document** and the representative query prepared by the user which should be similar to the documents needed for information retrieval.



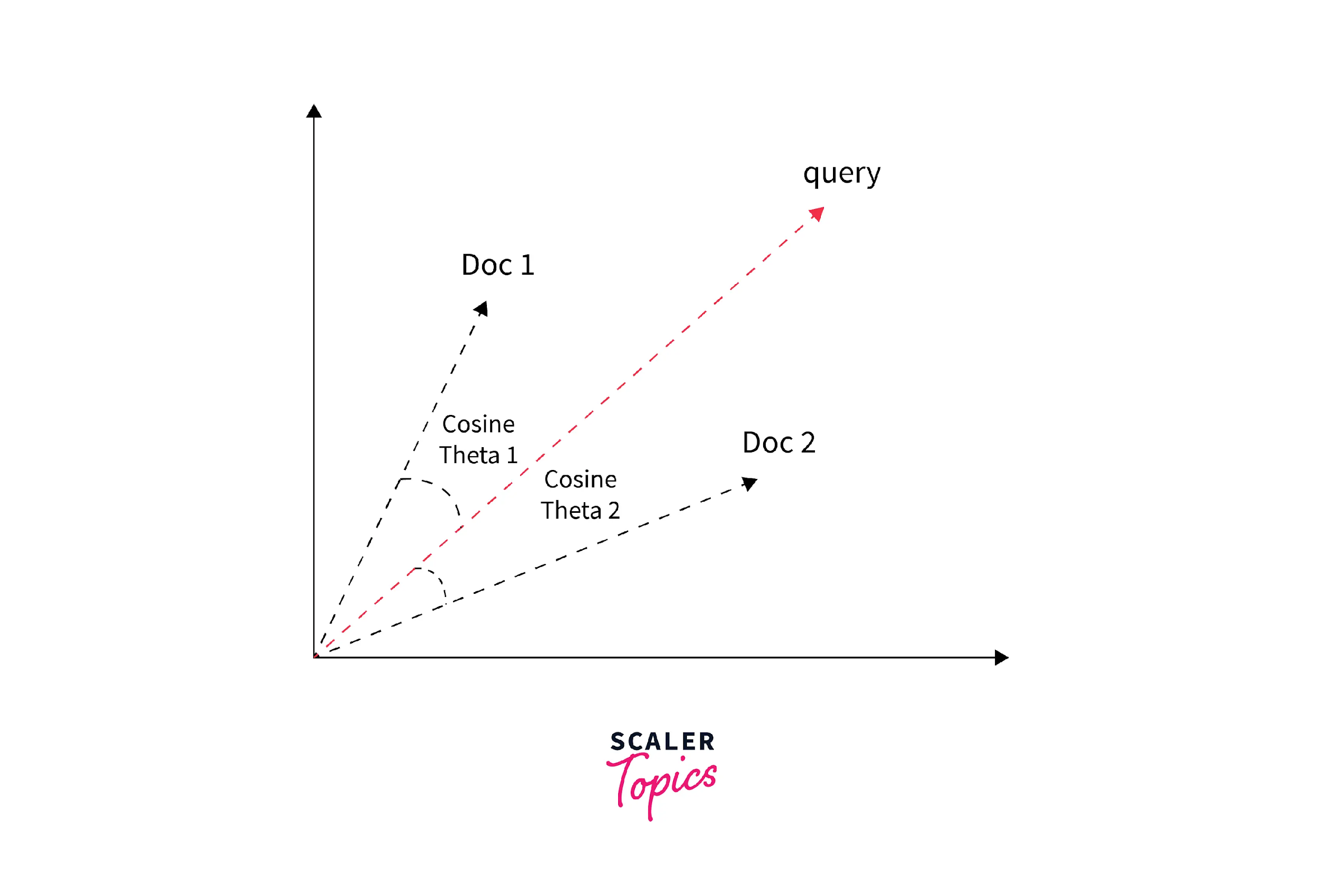
We can represent both documents and queries with vectors with a t-dimensional vector representation:

* dj = (w1,j,w2,j,...,wt,j) for a document with
* q = (w1,q,w2,q,...,wt,q)
* **Definition of Terms**: The **components of each dimension** in the document and query representation correspond to a separate term whose **terms can be** single words, keywords, or longer phrases also.
  + - If the **words** of the documents are chosen to be the terms in the above representation, the **dimensionality** of the vectors is the number of **distinct words** occurring in the vocabulary of the corpus.
* **Computing the values of the Terms**: The value of the vector is **non-zero** if a term or word occurs in the document. Many **different schemes can be taken** for example count of occurrences, normalized counts, etc. but the most **popular** and efficient representation is **tf-idf term weighing**.
* The vector space model represents the documents and queries as vectors in a **multidimensional space whose dimensions are the terms** further used to build an index to represent the documents.

### Notion of Similarity in Vector Space Model

The assumptions of the **document similarities theory** are used to compute the relevancy rankings of documents and the keywords in the search in vector space models.

* **Angle of deviation between query and document:** One way is to compare the **deviation of angles between each document vector and the original query vector** where the query is represented as some kind of vector as the documents.
* **Cosine distance as similarity metric:** The most popular and easier method in practice is to **calculate the cosine of the angle between the vectors** - A cosine value of zero means that the query and document vector are orthogonal and have no match at all.
  + A **zero cosine similarity** value implies that the **terms** in the query **do not exist** in the document we are considering.
* **Ranking the results using a similarity metric:** The degree of **similarity** between the representation of the prepared **document** and the representations of the documents in the **collection** is used to rank the search results.



One other way of looking at the similarity criterion in the vector space model is that - the more the two representations of search documents and the user-prepared query **agree** in the **given elements and their distribution**, the **higher** would the **probability of their representing** similar information.

### Index Creation for Terms in the Vector Space Model

The **creation** of the indices for the vector space model **involves** lexical scanning, morphological analysis, and term value computation.

* **Lexical scanning** is the creation of individual word documents to identify the **significant terms and morphological analysis reduces** to reduce different word forms to common stems and then compute the values of terms on the basis of stemmed words.
* The terms of the query are also **weighted** to take into account their **importance**, and they are computed by using the **statistical distributions of the terms** in the collection and in the documents.
* The vector space model assigns a **high ranking score** to a document that contains only a few of the query terms if these terms occur **infrequently** in the collection of the original corpus but frequently in the document.

### Assumptions of the Vector Space Model

* The **more similar** a document vector is to a query vector, the more likely it is that the document is **relevant** to that query.
* The **words** used to define the dimensions of the space are **orthogonal or independent**.
* The **similarity** assumption is an approximation and **realistic** whereas the assumption that words are pairwise **independent doesn't hold true** in realistic scenarios.

### Disadvantages of Vector Space Model

* Long documents are **poorly represented** because they have poor similarity values due to a small scalar product and a **large dimensionality of the terms** in the model.
* Search keywords must be **precisely designed** to match document terms and the word substrings might result in a **false positive match**.
* **Semantic sensitivity:** Documents with similar context but different term vocabulary won't be associated resulting in **false negative matches**.
* The **order in which the terms appear** in the document is lost in the vector space representation.
* **Weighting** is intuitive but not represented **formally** in the model.
* **Issues with implementation:** Due to the need for the similarity metric calculation and in turn storage of all the values of all vector components, it is **problematic** in case of **incremental updates** of the index
  + Adding a **single new document** changes the document frequencies of terms that occur in the document, which **changes the vector lengths of every document** that contains one or more of these terms.

**3. Probabilistic Model:**

Probabilistic models provide the **foundation for reasoning under uncertainty** in the realm of information retrieval.

Let us understand why there is **uncertainty** while retrieving documents and the basis for probability models in information retrieval.

**Uncertainty in retrieval models:** The probabilistic models in information retrieval are built on the idea that the process of retrieval is inherently uncertain from multiple standpoints:

* There is uncertainty in the **understanding of user’s information needs** - We can not sure that the user mapped their needs into the query they have presented.
* Even if the query represents the need well, there is **uncertainty in the estimation of document relevance** for the query which stems from either the uncertainty from the selection of the document representation or the **uncertainty from matching the query and documents**.

**Basis of probabilistic retrieval model:** Probabilistic model is based on the **Probability Ranking Principle** which states that an information retrieval system is supposed to rank the documents based on their **probability of relevance to the query given** all the other pieces of evidence available.

* Probabilistic information retrieval models **estimate how likely it is that a document** is relevant for a query.
* There may be a variety of sources of evidence that are used by the probabilistic retrieval methods and the most common one is the **statistical distribution of the terms in both** the relevant and non-relevant documents.
* Probabilistic information models are also among the **oldest and best performing** and most widely used IR models.

**Alternative Set Theoretic models:**

**1. Fuzzy Set Model:**

The fuzzy information retrieval model is fuzzy generalizations of the boolean model. The fuzzy information retrieval model defines the fuzzy relationship between query language and the retrieved documents. The fuzzy information retrieval system assumes that a set of fuzzy documents is associated with each word in the query language. That is to say, each word in the query language defines a fuzzy set, and the elements in the sets are retrieved documents. Correspondingly, each document in the set has a degree of membership to correspond to each word in the query language. As a retrieval result, the fuzzy set reflects how well each document matches the query.

Indexing is the preliminary operation in the creation of the documents’ representation. In the procedure of defining an indexing, we should ensure that the indexing can present textual information not only accurately but also overall. An indexing function, as the membership function of the fuzzy set, is used to calculate the correlation between words. In other words, the results of the membership function are the weights between index item and words in retrieval documents. People can employ the function to achieve the goal that presents textual information properly.

**2. Extended Boolean Model:**

In the three models discussed below, a document has a weight associated with each index term. This document weight is a measure of the degree to which the document is characterized by that term. Without loss of generality, we assume that document weights for all index terms lie in the range [0, 1]. This is less restrictive than in the standard Boolean model, which limits the values to the extremes of the range, namely 0 and 1.

To retrieve documents relevant to a given query, we need to calculate the query-document similarity for documents in the collection. The query-document similarity is an attempt to predict the relevance of a document to the query. In the following subsections, we consider each model and its method for calculating similarity.

The Extended Boolean Model is really just a general term that refer to all IR model that improve upon the Standard Boolean Model. There had been proposed many such model, such as the MMM model, the Paice model and P-norm model.

**Refer Textbook (Ch 02 Modeling) for the following topics:** <https://srikarthiks.files.wordpress.com/2016/07/t2-modern-information-retrieval.pdf>

**Probabilistic Models:**

**1. Inference Network Model:**

**2. Belief Network Model:**

**Structured text retrieval Models:**

**1. Models based on Non-overlapping lists:**

**2. Models based on Proximal Nodes:**

**Models for Browsing:**

**1. Flat Browsing**

**2. Structure Guided Browsing**

**3.The hypertext model**

**Questions:**

**1. Explain taxonomy of information retrieval model with classification diagram.**

**2. Structured text retrieval Models.**

**3. Explain Classic Information retrieval model in detail.**