# Statistical Methods for Information Retrieval

## I. Objectives

- To study the implementation of Statistical models in Information Retrieval

## II. Learning Outcome

After reading this module:

- The reader will understand the functions of statistical and mathematical models and how they are used to calculate the relevance in retrieval.
- The reader will understand the requirements of a good IR system.
- The reader will gain the knowledge of various advantages and limitations of Vector Space Model (VSM) and Binary Independence Model (BIM).
- The reader will be enriched with the knowledge of searching process and ranking mechanism of documents.

## III. Structure



## 1. Introduction

Any Information Retrieval framework works on documents collected and indexed by some tools and different methods can be applied to get information from the set of documents or the corpus. The way of retrieving relevant information from users' query are of different types. Researchers and scientists are continuously researching on this field as how to formulate a better searching algorithms for better felicitations. Now, the heterogeneity of information formats demand to invent more sophistication in order to address users' demands and business needs. The performance of any information retrieval is dependent on its scale of indexing and heterogeneity of resources. Web resources are different types and its identification and indexing strategy also differs. Whatever may be the types, the user needs to search and find the relevant resources with ease. IR adopts several statistical and mathematical models to calculate relevancy, rank them with relevance score with pre-determined formulae. The diversity of retrieval approaches is obvious due to change of technology and continuous research in this field. In this chapter, some basic statistical models would be covered keeping in mind its intended audience.

Searching is performed on the set of documents and each document is consisted of single-valued or multi-valued fields. On the contrary, the document may be unstructured, i.e., the content is not categorised under any structure following some rule or pattern. For example, if all the sentences of a book are collected and then just written in text file omitting all the punctuation or chapter delimiting formats etc., then it is an unstructured information/ content. Both the structured and unstructured content are being indexed by any IR software or any search engine. The proliferation of information generation embarked by different organizations is managed by different search engines which crawls web pages from WWW network.

Traditionally, IR system uses index terms to index and retrieve indexed documents. Each term in the document is not selected, only the terms which are semantically rich and add sense to the sentence are only selected and other non-semantic terms are omitted. These omitted words are called stop words. Usually stop words are articles, prepositions, conjunctions etc. Every indexer has predefined set of such words. The index terms(selected from the document) are stemmed and posting list is prepared. Posting list is nothing but a list which represents index terms occurring in which documents.

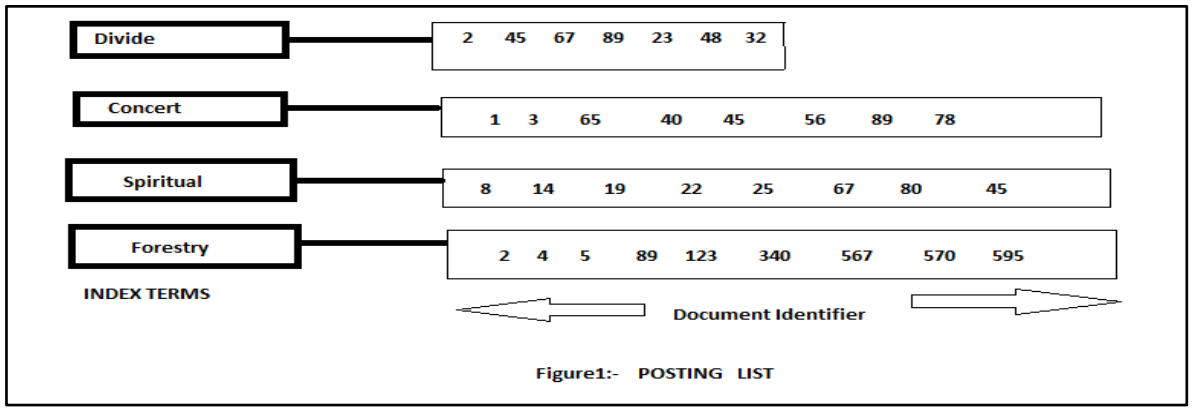


Figure1:- POSTING LIST

From the posting list, it's very easy to calculate the frequency of any term and thus an inverted index is produced which is the heart of any search engine. Inverted index is schematically represented in the following:

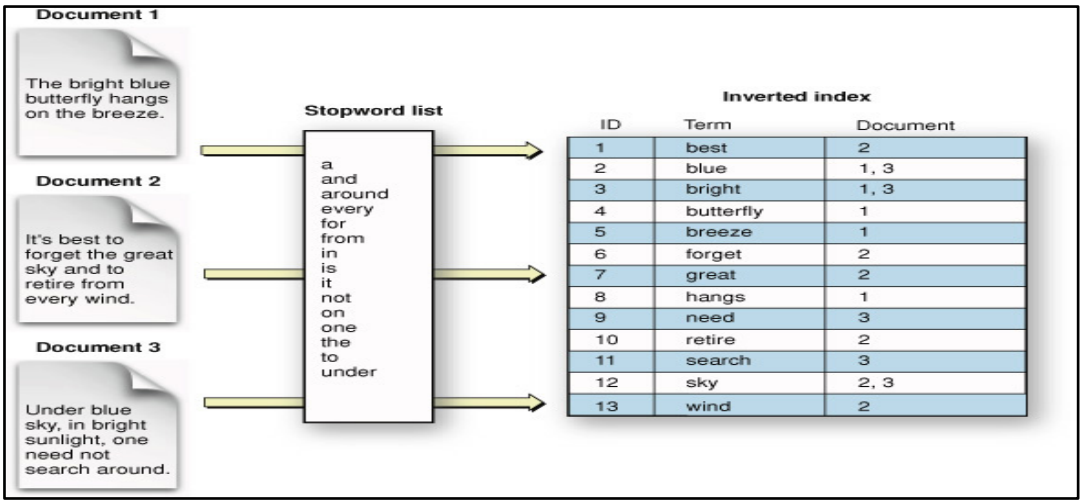


**Fig.2**: **Schematic view of Inverted Index**
**(source:-https://developer.apple.com/library/)**

To make an IR system efficient, strategy of searching must be robust and satisfy users' query in a systematic manner. So, the central point is all about relevance of documents. Classical models of IR are known as Boolean, Vector space and Probabilistic retrieval method. Boolean model represent documents and queries as set of index terms and the approach of calculating is based on set theory. Vector space model is a t-dimensional model which is based on algebra and the last method is based on probability. This module will only discuss the three classical models.

## 2. Boolean Search

Boolean retrieval is the first and most simple method in Information Retrieval system as they mainly adopted this approach to find the resources which undoubtedly indicates one obvious fact, that is, Boolean search by nature acts more or less like a locator type mechanism. The search mechanism is based on Boolean algebra. While formulating query, each term is appended with OR, AND, NOT like Boolean operators and they are acting as form of combined Boolean expression of terms. The main operation what it executes basically give the answer whether the term is present or not in the whole document set indexed by the IR system. Given its simplicity and formalism, this model attained a grand attention from the commercial search engines. Let's take an example for understanding.

Suppose a term-document matrix is:

| Doc-Id<br>Term | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 |
|---|---|---|---|---|---|---|---|---|---|
| T1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| T2 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0` | 1 |
| T3 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| T4 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| T5 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| T6 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| T7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| T8 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| T9 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

Here, the term-document matrix represents term occurring in which documents. 1 denotes the term occurring in that document otherwise the value is 0 in that particular cell of the matrix. Suppose the Query term is "Find those documents where term T1 and T2 are present but T8 is absent ". This query can be translated into Boolean notation-- **(T1) AND (T2)AND NOT (T8)** $\approx$ $[(T1 \wedge T2) \wedge \neg T8]$- that means we need to find the documents which carry the:terms T1 and T2 but not T8. So, the query is simplified by the Boolean operators present in the query form. Boolean operation here is bitwise. NOT is complement of the term expression.

(011001010) AND (011101101) AND NOT(001001000) which transformed into
(011001010) AND (011101101) AND (110110111)= (010000000).

That means document D2 is the right answer in which T1 and T2 are present but T8 is absent. So, logically Boolean expression of terms represents "Disjunction of conjunctive vectors" also known as Disjunctive Normal Form (DNF).

Disadvantages:

i. A Boolean search criterion is based on binary decision criteria executed with bitwise binary operation which eradicates the possibility of including relevance factor. It also can't segregate the searched result with its degree of relevance.

ii. Although Boolean model has precise pragmatic formalism and formulation, but it can't understand the natural semantic present among the terms. So finding related concepts or resources from search space is out of the scope. Natural language based query from the user sometimes are very difficult to transform into Boolean type structured query.

### 3. Vector Space Model (VSM)

Vector space model or term vector model is an algebraic model which represents text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. As the Boolean method is practically hit or miss type, the scope of partial matching is not possible in this approach as binary weight is limiting and straight-forward. Here, documents and queries are represented as t-dimensional vectors and consisted of words.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

Each expression for dimension here indicates to a separate term. If a term occurs in the document, its value in the vector is non-zero. The definition of term depends on the application. The query terms may be single words, keywords, or string of words (also known as phrases).The dimensionality of the vector is the number of words in the vocabulary (distinct words in the corpus) and it's dependant on the term chosen for the query. Several Vector operations can be used to evaluate relevance for documents with queries.

### 3.1 Relevance Rankings

"Relevance rankings of documents in a keyword search can be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents[4].Vector space model proposes to evaluate the degree of similarity of the document $d_j$ with respect to the query q as the correlation between the vector $\vec{d_j}$ and $\vec{q}$ . This correlation can be quantified by the cosine angle between the two vectors. The similarity function can expressed as –$Sim(d_j,q)= \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| X |\vec{q}|}$

In practice cosine of the angle between the vectors is calculated with the following equation: $\cos\theta = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| X |\vec{q}|}$.Where $\vec{d_j} \cdot \vec{q}$ is the dot product of the document and the query q vectors, $\|\vec{d_j}\|$ is the norm of vector $\vec{d_j}$, and norm of the vector q can be calculated as-

$$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2}$$

As all vectors under consideration by this model are element wise nonnegative, the cosine value may range from 0 to +1. If the value of cosine is zero that indicates the query and document vector are orthogonal to each other and no match as such is found (i.e. the query term does not exist in the document being considered). $\left|\underset{q}{\rightarrow}\right|$ does not affect ranking rather the factor $\left|\underset{dj}{\rightarrow}\right|$ provides normalization in the document cluster.

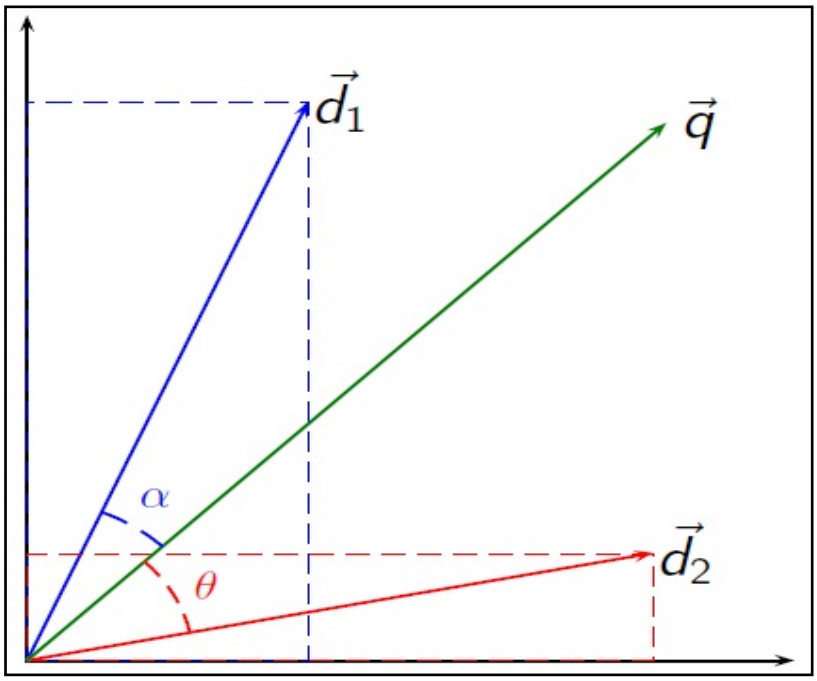


**Fig.3: Cosine Similarity among documents with Query (Source: - Wikipedia)**

In the classic vector space model (proposed by Salton, Wong and Yang) the term-specific weights in the document vectors are products of local and global parameters. The model is known as term frequency-inverse document frequency model. Instead of attempting to calculate whether a document is relevant or not, vector space model ranks the documents according to the degree of similarity to the query.VSM can rank documents matched partially according to cosine values. Index term weights are necessary for ranking. Let's explore this idea here.

The weight vector for document d is $\mathbf{v}_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T$, where the term weight is defined as -

$$w_{t,d} = \mathrm{tf}_{t,d} \cdot \log \frac{|D|}{|\{d' \in D \mid t \in d'\}|}$$

- $\mathrm{tf}_{t,d}$ is term frequency of term t in document d (a local parameter). Term frequency gives an idea how well the term describes the document or set of documents. Inter-cluster dissimilarity is quantified by measuring the inverse of the frequency of a term $k_i$ among the documents in the collection. If the term $k_i$does not appear in document $d_i$ ,then the corresponding $tf_{t,d}$= 0.

- $\log \frac{|D|}{|\{d' \in D \mid t \in d'\}|}$ is inverse document frequency (a global parameter). $|D|$ is the total number of documents in the document set; $|\{d' \in D \mid t \in d'\}|$ is the number of documents containing the term t. d' is a member document in the document set D.
Alternatively, it can be written as -idf = log $N/n_i$

Using the cosine the similarity between document $d_j$ and query q can be calculated as:

$$\text{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

Let's take an example to understand the scheme. Consider a small collection of documents:

**Doc1:- "New Delhi News"**
**Doc2:- "New Delhi Post"**
**Doc3:- "Mumbai News"**

So, total number of documents is N=3. Inverse document frequency of each term is:

**New = $\log_2$(3/2)= 0.584**
**Delhi= $\log_2$(3/2)= 0.584**
**News = $\log_2$(3/2)= 0.584**
**Post = $\log_2$(3/1)= 1.584**
**Mumbai = $\log_2$(3/1)= 1.584**

Now, for term frequency calculation, term-document matrix should be constructed.

| | **New** | **Delhi** | **News** | **Post** | **Mumbai** |
|---|---|---|---|---|---|
| **doc1** | **1** | **1** | **1** | **0** | **0** |
| **doc2** | **1** | **1** | **0** | **1** | **0** |
| **doc3** | **0** | **0** | **1** | **0** | **1** |

Now, tf value should be multiplied withidf values and to be put in term-document matrix.
So, thematrix becomes:

| | **New** | **Delhi** | **News** | **Post** | **Mumbai** |
|---|---|---|---|---|---|
| **doc1** | **0.584** | **0.584** | **0.584** | **0** | **0** |
| **doc2** | **0.584** | **0.584** | **0** | **1.584** | **0** |
| **doc3** | **0** | **0** | **0.584** | **0** | **1.584** |

Now , suppose the query(Q) is "**New New News**", tf-idf calculation for the query term is:

| | | | | | |
|---|---|---|---|---|---|
| Q → | **(2/2)* 0.584** | 0 | **(1/2)* 0.584** | 0 | 0 |

The length of each document and of the query are:

**Length of doc1: square root of (0.584^2+0.584^2+0.584^2)=1.0115**
**Length of doc2: square root of (0.584^2+0.584^2+1.584^2)=1.78638**
**Length of doc3:square root of (0.584^2+1.584^2)=1.6882**
**Length of Q:square root of (0.584^2+0.292^2)= 0.6529**

Then the similarity values are:

| | | | | | |
|---|---|---|---|---|---|
| **Doc1** | **0.584** | **0.584** | **0.584** | **0** | **0** |
| **Query** | **0.584** | **0** | **0.292** | **0** | **0** |

**Cosine Similarity($d_1$,q)=(0.584*0.584+0+0.584*0.292+0+0)/(1.0115*0.6529)= 0.7746**

| | | | | | |
|---|---|---|---|---|---|
| **Doc2** | **0.584** | **0.584** | **0** | **1.584** | **0** |
| **Query** | **(2/2)* 0.584** | **0** | **0.292** | **0** | **0** |

**Cosine Similarity ($d_2$,q)= (0.584*0.584+0+0+0+0)/(1.78638*0.6529)= 0.2924**

| | | | | | |
|---|---|---|---|---|---|
| **Doc3** | **0** | **0** | **0.584** | **0** | **1.584** |
| **Query** | **0.584** | **0** | **0.292** | **0** | **0** |

**Cosine Similarity ($d_3$,q)=(0+0+0.584*0.292+0+0)/(1.6882*0.652)=0.1549**

According to the similarity values calculated, documents can be ordered as – doc1, doc2 and doc3. Doc1 is the most relevant document corresponding to the query made.

The vector space model has the following advantages over the Standard Boolean model:

a. Simple model based on linear algebra
b. Term weights are not binary
c. Allows computing a continuous degree of similarity between queries and documents
d. Allows ranking documents according to their possible relevance
e. Allows partial matching

The major limitations of vector space model are [4]:

i. Long documents are poorly represented because they have poor similarity values (a small scalar product and a large dimensionality)

ii. Search keywords must precisely match document terms; word substrings might result in a "false positive match"

iii. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".

iv. The order in which the terms appear in the document is lost in the vector space representation.

v. Theoretically assume terms are statistically independent.

vi. Weighting is intuitive but not very formal.

## 4. Probabilistic Relevance Model

It was developed by Robertson and Jones as a framework for probabilistic approach. It estimates the probability of finding if a document $d_j$ is relevant to a query q. This model assumes that this probability of relevance depends on the query and document representations. Furthermore, it assumes that there is a portion of all documents that is preferred by the user as the answer set for query q. Such an ideal answer set is called R and should maximize the overall probability of relevance to that user. The prediction is that documents in this set R are relevant to the query, while documents not present in the set are non-relevant and represented as $\overline{R}$.

$$sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\bar{R}|\vec{d_j})}$$

## 5. Probability Rank Principle

For a query q and a document d ,$R_{d,q}$ be the random variable to say whether a document retrieved given to a query is relevant or not. It can be defined as P(R=1|d,q) which is the basis of PRP. Rijsbergen(1979) stated that , 'If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.' So, the binary situation is incurred upon the ranking of documents in decreasing order after probability calculation.

**Document d is relevant iff P(R=1|d,q) , if irrelevant P(R=0|d,q).**

This is also known as 1/0 loss case of PRP. Here no cost is imposed either for retrieval or ranking.

The second case of PRP is with retrieval cost. It's assumed to incur cost upon the judgement of relevancy of the retrieved document. Let us assume that, the cost $c_1$is to be given for not retrieving a document upon query q and cost $c_0$ is given to retrieval of non-relevant document. According to PRP, for a specified document d and for all not retrieved document d' not yet retrieved-

$$\mathbf{c_0 . P(R=0|d) - c_1 . P(R=1|d) \leq c_0 . P(R=0|d') - c_1 . P(R=1|d')}$$

Here, d is the next document to be retrieved. The main advantage of this method is , it can model differential costs of false positive and false negative and related system performance issues along with validation of fault tolerance level.

## 6. Binary Independence Model

BIM is a probabilistic Information Retrieval technique based on Probability Rank Principle (PRP). Probabilistic IR model is based on basic probability computation from the available document corpus and application of bayes theorem to judge existence of likelihood [Posterior Probability- P(A|B)]. Here “Binary” has the same meaning with Boolean. Documents and queries are represented as binary term incidence vectors. A document d is represented as a vector form, $\vec{x}$= $\{x_1,x_2,x_3,x_4....x_t\}$ where $x_t$= 1 the term t is present in the document d and $x_t$=0 if the term t is not present in the document d. “Independent”means the terms modelled

here are occurring independently within the documents. So with this notation, many documents might have the same vector representation. The probability function of PRP i.e. P(R|d,q) is relevant here but with a modified version. In BIM, q is represented as an incidence vector $\vec{q}$. So, probability functionP(R|d,q) is tranformed into P(R $|\vec{x},\vec{q}$). Using baysian formula it can be written as:

$P(R = 1 \vee \vec{x},\vec{q}) = \frac{P(\vec{x} \vee R=1,\vec{q})P(R=1 \vee \vec{q})}{P(\vec{x} \vee \vec{q})}$ and $P(R = 0 \vee \vec{x},\vec{q}) = \frac{P(\vec{x} \vee R=0,\vec{q})P(R=0 \vee \vec{q})}{P(\vec{x} \vee \vec{q})}$

Here $P(R=1|\vec{x},\vec{q})$ and $P(R=0|\vec{x},\vec{q})$ are the probability of relevant and non-relevant documents respectivily. We must remember that $P(R=1|\vec{x},\vec{q})$+ $P(R=0|\vec{x},\vec{q})$=1 which is obvious and implied.

**Advantages**

i. Calculation is simple and binary decision is easy to understand.
ii. Inclusion of Probability refines decision and increases the chance to retrieve true positive documents.

**Disadvantages**

i. The probability of relevant and non-relevant documents is not known beforehand, so, the calculation of $P(R=1|\vec{x},\vec{q})$ and $P(R=0|\vec{x},\vec{q})$ are basically an estimation process.
ii. As the vector representations of many documents are same, it also increases the chance of retrieving false-negative documents as well as near duplicated documents also.

## 7. Summary

In this chapter, only the classical models are presented. Modern search engines incorporate very complex and indigenous algorithms at the back-end. Boolean retrieval is very straightforward but the most weak form of retrieval method. Vector space model has the edge over Boolean method as it can give fractional similarity measurement. Probabilistic method is relatively new and has created a route to incorporate advanced and critical statistical methods such as inference network model, Bayesian network model etc. Main beneficiaries of the IR models are library systems, specialised Information Retrieval system and the Web. There are several commercial and free software available which can be used for creating personalized search engine or creating searching interface of the documents stored at one's disposal. The models described here is the delineation of the inline methods involved in searching processes and ranking mechanism of the documents based on query made by the users.

## 8. References


1. Baeza-Yates, R., &Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

2. Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval.

3. Chowdhury, G. (2010). Introduction to modern information retrieval. Facet publishing.

4. G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975

5. Retrieved from http://nlp.stanford.edu/IR-book/html/htmledition/the-binary-independence-model-1.html