Vector processing information retrieval model

Vector space model or **term vector model** or **Vector processing model** is a classical, structural or system-centered information retrieval model which represents both documents and queries by term sets and compares global similarities between queries and documents. It is based on mathematical principles. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

Definition

In this model *term vectors* are assigned for the keywords of the documents and weights are provided according to relevance. It is helpful to compare different texts and retrieve the relevant records that are found similar to the queries. The definition of *term* depends on the application. Typically terms are single words, keywords, or longer phrases. If the words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary (the number of distinct words occurring in the corpus).
Vector operations can be used to compare documents with queries.

Applications

Consider two documents - Doc_i and Doc_j. Let terms $Term_{ik}$ represent the weight of the term *k* assigned to the document Doc_i and similarly $Term_{jk}$ for document $Doc_{j.}$
Now the two document vectors may be represented as

$Doc_i = (Term_{i1}, Term_{i2}, Term_{i3},\ldots Term_{it})$
$Doc_j = (Term_{j1}, Term_{j2}, Term_{j3}, \ldots Term_{jt})$

Where *t* terms are assigned as keywords for each document.
To compute the similarity between these two vectors, the following vector functions are used.

$$\sum_{k=1}^{T} TERM_{ik}$$

This denotes the sum of weights of all terms of a document.

$$\sum_{k=1}^{T} TERM_{ik}.TERM_{jk}$$

This denotes the sum of products of the corresponding term weights of the two documents.

$$\sum_{k=1}^{T} min(TERM_{ik}.TERM_{jk})$$

This denotes the sum of the minimum of the weights of the two documents.

Salton and McGill have formulated various coefficients for measuring the similarity of the documents, i.e. $Sim(Doc_i,Doc_j)$.

The dice coefficient

$$SIM(DOC_i, DOC_j) = \frac{2[\sum_{k=1}^{T}(TERM_{ik}.TERM_{jk})]}{\sum_{k=1}^{T} TERM_{ik} + \sum_{k=1}^{T} TERM_{jk}}$$

The Jaccard coefficient

$$SIM(DOC_i, DOC_j) = \frac{\sum_{k=1}^{T} (TERM_{ik}.TERM_{jk})}{\sum_{k=1}^{T} TERM_{ik} + \sum_{k=1}^{T} TERM_{jk} - \sum_{k=1}^{T} (TERM_{ik}.TERM_{jk})}$$

Advantages

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

1. Simple model based on linear algebra
2. Term weights are not binary
3. Allows computing a continuous degree of similarity between queries and documents
4. Allows ranking documents according to their possible relevance
5. Allows partial matching

Limitations

The vector processing model has the following limitations:

1. Long documents are poorly represented because they have poor similarity values
2. Search keywords must precisely match document terms; word substrings might result in a "false positive match"
3. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".
4. The order in which the terms appear in the document is lost in the vector space representation.
5. It assumes that terms are statistically independent.
6. Weighting is intuitive but not very formal.

Many of these difficulties can, however, be overcome by the integration of various tools, including mathematical techniques such as singular value decomposition and lexical databases such as WordNet .

Reference

1. Introduction to modern information retrieval (2nd ed.), G.G.Chowdhury
2. https://en.wikipedia.org/wiki/Vector_space_model