JUNAID GIRKAR
6000419005-7

29/3/2023

WI ASSIGNMENT - 1

BE COMPS A2

Q3 Explain vector space model clearly with an example

ANS

- Information retrieval is study of helping users to find information that match their information needs. It is about acquisition, organization, storage, retrieval and distribution of information.
- An IR model governs how a document and query are represented and how relevance of document to user query is defined.
- there are 4 main IR models :-
  - → Boolean
  - → Vector space
  - → Language
  - → Probabilistic

- Important terms in IR Model :-

1. Each document and query treated as a bag of words.
2. Each term associated with a weight
3. Given collection of documents D, V (vocabulary) $= \{t_1, t_2, \ldots t_{|V|}\}$ set of distinct terms in collection D.
4. weight $w_{ij} > 0$ associated with term t_i of document $d_j \in D$; if term does not appear in d_j then $w_{ij} = 0$
5. $d_j = (w_{1j}, w_{2j}, \ldots, w_{|V|j})$, collection of d_j represented as matrix, in different models w_j is computed differently.

__/__/____

⇒ VECTOR SPACE MODEL

- Best known and widely used IR model

• Document Representation.

A document in vector space model is represented as a weight vector, in which weight of each component is computed based on some variation of TF or TF-IDF scheme. Thus w_{ij} can be any number

(i) **TERM FREQUENCY (TF) scheme:**

The weight of term t_i in d_j is number of times t_i appears in document d_j denoted by f_{ij}. Normalization may also be applied.

$$tf_{ij} = \frac{d_{ij}}{\max\{f_{ij}, f_{2j_3}, \cdots f_{|V|j}\}}$$

The shortcoming of TF scheme is that it doesn't consider a situation where a term appears in many documents of the collection. Such a term may not be discriminative.

(ii) **TF - IDF SCHEME:**

Most well known weighting scheme, TF stands for term frequency & IDF stands for inverse document frequency.

Let N be total number of documents in collection. df_i be number of documents

in which term t_i appears atleast once. f_{ij} be raw count of t_i in d_j, then the normalized term frequency of t_i in d_j is given by

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \cdots f_{ivj}\}}$$

where maximum is computed over all terms that appears in document d_j.

The inverse document frequency of term t_i is given by, $idf_i = \log\left(\frac{N}{df_i}\right)$

The intuition is that if term appears in large number of documents, it is probably not important or that discriminative. The final TF-IDF weight is given by: $w_{ij} = tf_{ij} \times idf_i$

- **QUERIES**

A query q is represented in exactly the same way as document in collection. The term weight w_{iq} of each term t_i in q can also be computed in same way as normal document.

- **DOCUMENT RETRIEVAL & RELEVANCE RANKING**

The documents are ranked according to their degrees of relevance to the query. One way to compute degree of relevance is to calculate similarity of query q to each document d_j in D. There are many similarity measures,

well known one is cosine similarity, which is cosine of angle between q and d_j

$$\text{cosine}(d_j, q) = \frac{\langle d_j \cdot q \rangle}{\|d_j\| \times \|q\|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$

Ranking of documents is done using similarity values. Some other methods of computing relevancy scores of document d_j with query q are: simple dot product, okapi method, pivoted normalization weighting.

EXAMPLE

Documents : N = 3

d1 – New York times
d2 – New York Post
d3 – los Angeles times

Query q = new new times

Documents :

Term frequency matrix $\frac{d_{ij}}{\max\{f_{1j}, f_{2j}, \ldots f_{|V|j}\}}$

	d1	d2	d3
new	1	1	0
york	1	1	0
times	1	0	1
post	0	1	0
los	0	0	1
Angeles	0	0	1

Inverse Document Frequency of terms = $\log\left(\frac{N}{df_i}\right)$

new	log (3/2)	= 0.176
york	log (3/2)	= 0.176
times	log (3/2)	= 0.176
post	log (3/1)	= 0.477
los	log (3/1)	= 0.477
angeles	log (3/1)	= 0.477

TF-IDF Score Matrix ($tf_{ij} \times idf_i$)

	d1	d2	d3
new	0.176	0.176	0
york	0.176	0.176	0
times	0.176	0	0.176
post	0	0.477	0
los	0	0	0.477
angeles	0	0	0.477

Query : TF-IDF score

	q	
new	2/2 x 0.176	= 0.176
york	0	0
times	1/2 x 0.176	= 0.088
post	0	0
los	0	0
angeles	0	0

__/__/__

Computing values of

$$||d1|| = \sqrt{0.176^2 + 0.176^2 + 0.176^2} = 0.305$$

$$||d2|| = \sqrt{0.176^2 + 0.176^2 + 0.477^2} = 0.538$$

$$||d3|| = \sqrt{0.176^2 + 0.477^2 + 0.477^2} = 0.697$$

$$||q|| = \sqrt{0.176^2 + 0.088^2} = 0.196$$

Finding cosine similarities (d_j, q).

$$\cos(d1, q) = \frac{\langle d1, q \rangle}{||d1|| \times ||q||} = \frac{0.031 + 0.015}{0.305 \times 0.196} = 0.7695$$

$$\cos(d2, q) = \frac{\langle d2, q \rangle}{||d2|| \times ||q||} = \frac{0.031}{0.538 \times 0.196} = 0.2940$$

$$\cos(d3, q) = \frac{\langle d3, q \rangle}{||d3|| \times ||q||} = \frac{0.015}{0.697 \times 0.196} = 0.1098$$

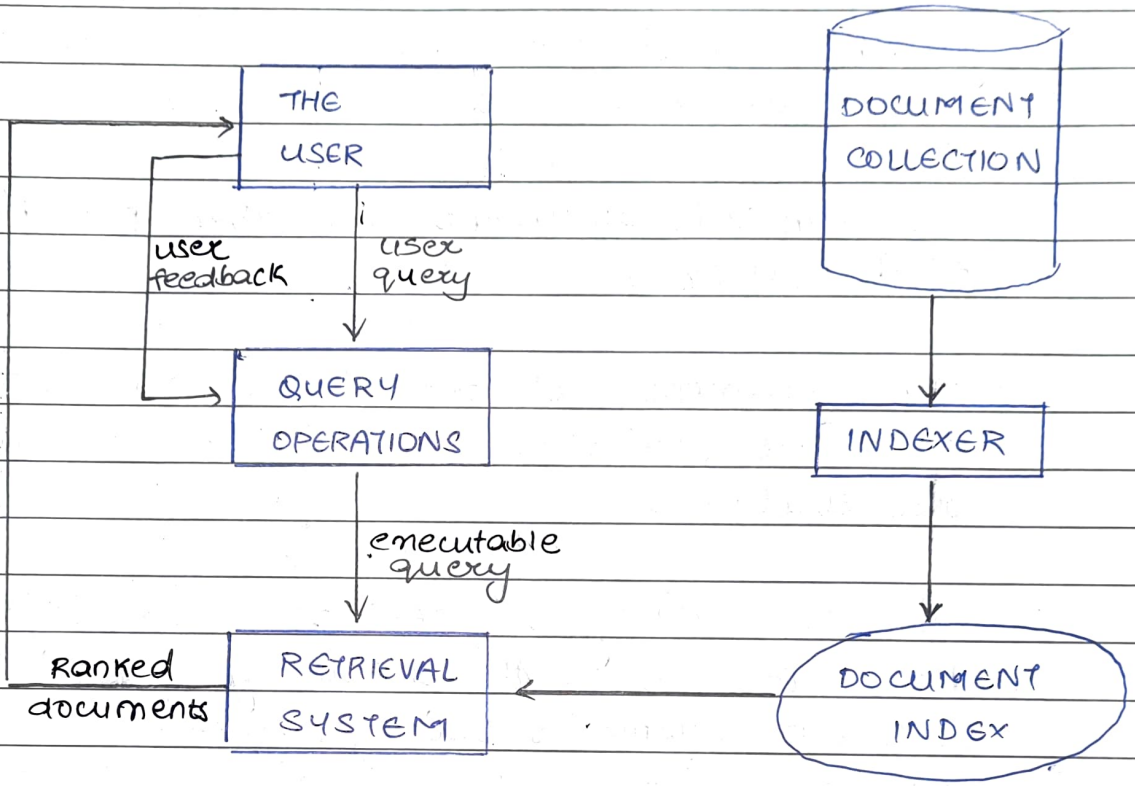
$\therefore$ $0.7695 > 0.2940 > 0.1098$

$\therefore$ Based on query, documents will be presented in the order $d_1 > d_2 > d_3$

__/__/____

Q4 Draw and explain general IR system Architecture.

ANS



GENERAL IR SYSTEM

- Information retrieval is the study of helping users find information that matches their information needs. It studies the acquisition, storage, retrieval and distribution of information
- IR is about document retrieval, emphasizing document as basic unit.
- the user with information neds issues a query (user query) to the retrieval system through query operations module.

__/__/___

- The retrieval module uses document index to retrieve those document that contain some query terms, compute relevancy scores for them and then rank the retrieved documents according to the scores.

- The ranked documents are then presented to the user

- The document collection is called the text database, indexed by indexer for efficient retrieval.

- A user query represents users information needs in one of the following forms:
  (i) KEYWORD QUERIES : List of keywords
  (ii) BOOLEAN QUERIES : Boolean operators to construct complex queries
  (iii) PHRASE QUERIES : sequence of words
  (iv) PROXIMITY QUERIES : Relaxed version of phrase query and can be combination of terms and phrases.
  (v) FULL DOCUMENT QUERIES : Full document
  (vi) NATURAL LANGUAGE QUERIES : human languages. Also known as answering questions.

- Query operation module in simplest case does nothing and pass query to retrieval module after pre-processing (removing stopwords). In complex cases, it transforms natural language

queries to executable queries. It may also accept user feedback and use it to expand & refine original queries called as relevance feedback.

- The indexer indexes raw documents for efficient retrieval, result is document index, most popular indexing scheme is inverted index.

- The retrieval system doesn't compare query with every document in collection which is inefficient. Instead, a small subset of documents containing atleast one query term is first found from index and then relevance scores are computed for this subset of documents with user query.

- An IR model governs how a document & query are represented and how relevance of document to user query is defined.

- There are 4 main IR models :-
  (i) Boolean Model (simplest)
  (ii) Vector space Model (widely used)
  (iii) Language Model (NLP)
  (iv) Probabilistic Model (Stochastic).