MODULE 5

Q.9 (a) State and explain Zipf's law

Ans * Zipf's law made an important observation on the distribution of words in Natural program,

* This observation has been named Zipf's law.

* Simply stated, Zipf's law says that frequency of words ~~manipulated~~ multiplied by their ranks in a large corpus is more or less constant.

* Frequent x rank ≃ constant

* This means that if we compute the frequencies of the words in a corpus, and arrange them in descending order of frequency, then the product of the frequency of a word and its rank is approximately equal to the product of the frequency and rank of another word.

* This indicates that the frequency of a word is inversely proportional to its rank.

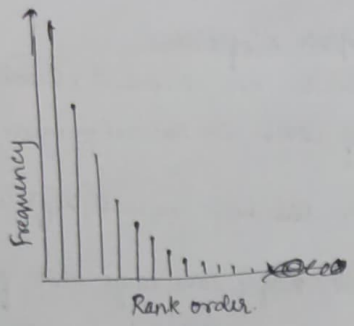

Fig: Relationship between the frequency of words and their rank order.

* Empirical investigation of Zipf's law on large corpus suggest that human language contain a small number of words that occur with high frequency; large number of words the occur with low frequency, In between middling number of medium frequency terms

* This distribution has important significance in IR.

* High frequency words have less discriminating power, and not useful for indexing.

* Low frequency word are less likely to be included in the query, and are also not useful for indexing.

* As there are a large number of rare words, dropping them considerably reduces the size of a list of index terms.

* The remaining medium frequency words are content-bearing terms and can be used for indexing.

* This can be implemented by defining thresholds for high & low frequency, and dropping words that have frequencies above or below these thresholds.

* Stop word elimination can be thought of as an implementation of Zipf's law, where high frequency terms are dropped from a set of index terms.

Q9(b) Define the following wrt Information Retrieval

a] Vector Space Model :

→ The vector space model is one of most well-studied retrieval models.

→ Important contribution to its development was made by Luhn (1959), Salton (1968), Salton and McGill (1983), and van Rijsbergen.

→ The vector space model represents documents and queries as vector of features representing items that occur within them.

→ Each document is characterized by a Boolean or numerical vector

→ These vectors are represented in a multi-dimensional space, in which each dimension corresponds to a distinct term in the corpus of documents.

→ In the simplest form each features takes a value of either zero or one, indicating the absence or presence of that term in a document or query.

→ Ranking algorithm compute the similarity between document and query vectors, to yield a retrieval score to each document.

→ Given a finite set of n documents

$$D = \{d_1, d_2, \ldots d_p \ldots d_n\}$$

and a finite set of m terms

$$T = \{t_1, t_2, \ldots t_i, \ldots, t_m\}$$

each document is represented by a column vector of weights as follows

$$(w_{1j}, w_{2j}, w_{3j}, \ldots w_{ij}, \ldots w_{mj})^t$$

where w_{ij} is the weight of the term t_j in document d_j. The document collection as a whole is represented by an mxn term document matrix as

$$\begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1j} & \ldots & w_{1n} \\ w_{21} & w_{22} & \ldots & w_{2j} & \ldots & w_{2n} \\ w_{i1} & w_{i2} & \ldots & w_{ij} & \ldots & w_{in} \\ w_{m1} & w_{m2} & \ldots & w_{mj} & \ldots & w_{mn} \end{bmatrix}$$

b] Term Frequency :

n $tf = tf_{ij}$

b $tf = 0$ or 1 (binary weight)

} Raw term frequency

A

a $tf = 0.5 + 0.5\left(\frac{tf_{ij}}{\max tf \text{ in } D_j}\right)$ → Augmented term frequency

l $tf = \ln(tf_{ij}) + 1.0$ → Logarithmic term frequency

2 $tf = \frac{\ln(tf_{ij} + 1.0)}{1.0 + \ln[\text{mean}(tf \text{ in } D_j)]}$ → Average term frequency - based normalization

→ There are many ways to compute each compute. The simplest is to use either binary weight or raw term frequency.

→ The first occurance of a term is more important than successive repeating occurance

→ Thus, tf can be computed as $0.5 + 0.5(tf_{ij}/\max tf \text{ in } d_j)$ in which normalization is achieved by dividing tf by maximum tf value for any term in the document, or as $\ln(tf_{ij}) + 1.0$, which is known as logarithm term frequency.

- The former computation is called augmented normalized term frequency. It cause tf to vary b/w 0.5 and 1.
- The problem with maximum normalization and argumented normalization of the tf component is that a single term in a document, with an unusually high frequency may degrade the weights of the others terms significantly.
- This effect is not too pronounced with the argumented tf, because the highest frequency term cannot degrade the frequency of other terms below 0.5.
- The logarithmic term frequency reduces the effect unusually frequent terms within a document.

• It actually decreases the effect of all sorts of variations in tf, because for any two terms frequencies tf_1 and $tf_2 > 0$ such that $tf_1 > tf_2$, the ratio of the logarithmic term frequencies will always less than the ratio of the raw frequencies

$$\frac{\log(tf_1) + 1}{\log(tf_2) + 1} < \frac{tf_1}{tf_2}$$

c] Inverse Document Frequency

→ n wt = tf → no conversion

B

t $wt = tf \cdot \ln\left(\frac{n}{ni}\right)$ → multiply tf with idf.

Q10(a) Explain the architecture of an Information Retrieval system with a neat diagram.

ANS:

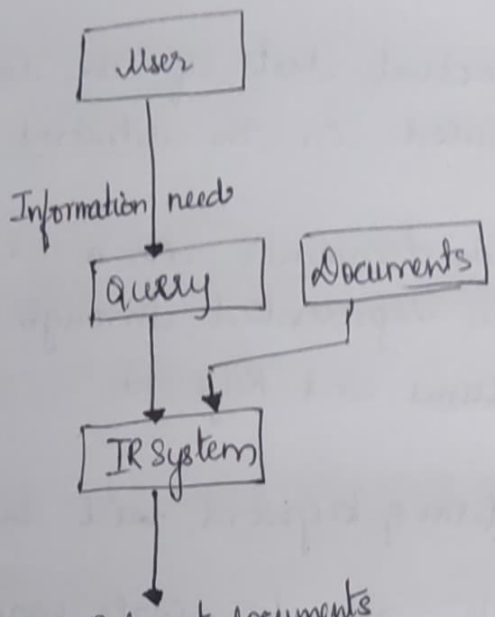

Fig: Basic information retrival process

→ It begins with the user's information need. Based on this need, he/she formulates a query. The IR system returns documents that seem relavent to the query. This is an engg account of the IR system.

→ The basic question involved is, 'what constitutes the information in the document and the queries'. This in turn is related to the problem of representation of documents and queries.

→ The retrieval is performed by matching the query representation with document representation.

→ The actual text of the document is not used in the retrieval process.

→ Instead, documents in a collection are frequently represented through a set of index terms and keyword.

→ The term & keyword will be used independently.

→ Keywords can be single word or multiword phrases. They might be extracted automatically or manually. Such a representation provides a logical view of the document

→ The process of transforming document text to some representation of it is known as indexing. There are different types of index structure.

→ One used data structure, commonly by the IR system, is the inverted index; An inverted index is simply a list of keywords, with each keyword carrying pointers to the document containing that keyword.

→ The computational cost involved in adopting a full text logical view using a full set of words to represent a document is high.

→ Some text operations are usually performed to reduce the set of representative keywords.

→ The two most commonly used text operations are stop word elimination and stemming.

→ Stop word elimination removes grammatical or functional words, while stemming reduces words to their common grammatical roots.

→ Zipf's law can be applied to further reduce the size of index set. Not all the terms in a document are equally relevant.

→ Some might be more important in conveying a document's content.

→ Attempts have been made to quantify the significance of index terms to a document by assigning them numerical values, called weights.

→ The choice of index terms and weights is a difficult theoretical and practical problem and several technique are used to cope with it.

→ A number of term-weighting schemes have been proposed in the literature over the years.

Q10(b) write the hyponym chain for "RIVER" extracted from the wordnet 2.0

ANS: 1 sense of 'river'

Sense 1

RIVER – (a large natural stream of water (larger than a creek); the river was navigable for 50 miles')

⇒ stream, watercourse – (a natural body of running water flowing on or under the earth)

⇒ body of water, water – (the part of the earth's surface covered with water (such as a river or lake or ocean); they invaded our territorial waters; they were sitting by the water edge')

⇒ thing - (a separate and self-contained entity)

⇒ Intity - (that which is preceived or known or inferred to have its own distinct existence (living or nonliving))

Q.10 (c) How stemming affects the performance of IR system?

Ans:→ stemming normalizes morphological variants, though in a crude manner, by removing affixes from the words to reduce them to their stem.

→ Eg: The words compute, computing, computes and computer, are all reduced to same word stem, comput.

→ Thus, the keywords or terms used to represent text are stems, not the actual words.

→ Note that stop words have been removed in this representation and the remaining terms are in lower case.

→ one of the problem associated with stemming is that it may throw away useful distinctions. In some cases, it may be useful to help conflate similar terms, resulting in increased recall.

→ It may be harmful, resulting in reduced precision.

→ Recall and precision are the two most commonly used measures of the effective of an information retrieval system.