Page No.
Date : / /

Name : Dhiraj Bodake
Roll No : 18141216

TUTORIAL NO : 02

Q. 1 Write Short note on
i) Indexing ii) Index term weighting

→ i) Indexing -

An index language is used to describe doc & request. the element of index language are index term which may be derived from text document to be described or may be arrived of independently. Index language may be describ as pre-coordinate or post co-ordinate. First indicates that term are coordinated at time of indexing & latter at time of indexing & latter searching move specifically, in pre-coordinate indexing a logical combination of any index term may be used as a label to identify a class of docs. whereas in past coordinate indexing the some class would be identified at search time by combining the classes of docs.

keen & Digger further define indexing specificity is a ability of index language to describe topics.

Rajdhani
DATE / /

Q.2. Explain various measures of association in details.

→ There are 5 commonly used measures of association in IR. The simplest of all association measures $|X \cap Y|$ Simple matching coefficient. which is no. of shared index terms. The coefficient does not take into account the sized X & Y. following coefficient have been used in doc.

2 $\frac{|X \cap Y|}{|X|+|Y|}$ Dice's Coefficient

$\frac{|X \cap Y|}{|X \cup Y|}$ Jaccard's Coefficient

$\frac{|X \cap Y|}{|X|^{1/2} \times |Y|^{1/2}}$ Cosine Coefficient

$\frac{|X \cap Y|}{\min(|X|, |Y|)}$ Overlap Coefficient

A measure of association increases as the no. of proportion of shared attribute state increases. Numerous weefficients of association have been described in literature. It follows that a cluster method depending only on rank-ordering of the association values would give identical clusterings for all those measures

Rajdhani
DATE / /

Q.3. Explain concept of probabilistic indexing in details.

→ In model they consider the difference in distributional behaviour of words as a guide to whether a word should be assigned index term. Found that funct^n words were closely modelled by poisson distribution over all docs. whereas specially words did not follow poisson distribution.

Specifically, if one is looking at the distribution of a function word over a set of texts than the probability, f(n), that a text will have n occurrences of function word w is given by

$$f(n) = \frac{e^{-x} x^n}{n!}$$

In general, the parameter x will vary from word to word, & for given word should be proportional to length of text. We also interpret x as mean no. of occurrences of w in set of texts.

The resulting technique called probabilistic indexing allows a computing machine given a request for information to make a statical interference & derive a no. for each doc., which is a measure of probability that the doc. will satisfy given request.