UNIT II – INFORMATION RETRIEVAL

Part A - Questions

1. **What do you mean information retrieval models?**

   A retrieval model can be a description of either the computational process or the human process of retrieval: The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

2. **What is cosine similarity?**

   This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities.

3. **What is language model based IR?**

   A language model is a probabilistic mechanism for generating text. Language models estimate the probability distribution of various natural language phenomena.

4. **Define unigram language.**

   A unigram (1-gram) language model makes the strong independence assumption that words are generated independently from a multinomial distribution θ

5. **What are the characteristics of relevance feedback?**

   - It shields the user from the details of the query reformulation process.
   - It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
   - Provide a controlled process designed to emphasize some terms and de-emphasize others.

6. **What are the assumptions of vector space model?**

   Assumption of vector space model:

   - The degree of matching can be used to rank-order documents;
   - This rank-ordering corresponds to how well a document satisfying a users information needs.

7. What are the disadvantages of Boolean model?

- It is not simple to translate an information need into a Boolean expression
- Exact matching may lead to retrieval of too many documents.
- The retrieved documents are not ranked.
- The model does not use term weights.

8. Define term frequency.

Term frequency: Frequency of occurrence of query keyword in document.

9. Explain Luhn's ideas

Luhn's basic idea to use various properties of texts, including statistical ones, was critical in opening handling of input by computers for IR. Automatic input joined the already automated output.

10. Define stemming.

Conflation algorithms are used in information retrieval systems for matching the morphological variants of terms for efficient indexing and faster retrieval operations. The Conflation process can be done either manually or automatically. The automatic conflation operation is also called stemming.

11. What is Recall?

Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents retrieved.

12. What is precision?

Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

13. Explain Latent semantic Indexing.

Latent Semantic Indexing is a technique that projects queries and documents into a space with "latent" Semantic dimensions. It is statistical method for automatic indexing and retrieval that attempts to solve the major problems of the current technology. It is intended to uncover latent semantic structure in the data that is hidden. It creates a semantic space where in terms and documents that are associated are placed near one another.

Part –B

1. **Explain about term weighting.**

Term frequency

Suppose we have a set of English text documents and wish to determine which document is most relevant to the query "the brown cow". A simple way to start out is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its *term frequency*. However, in the case where the length of documents vary greatly, adjustments are often made (see definition below).

The first form of term weighting is due to Hans Peter Luhn (1957) and is based on the Luhn Assumption:

- The weight of a term that occurs in a document is simply proportional to the term frequency.

Inverse document frequency

Because the term "the" is so common, term frequency will tend to incorrectly emphasize documents which happen to use the word "the" more frequently, without giving enough weight to the more meaningful terms "brown" and "cow". The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less common words "brown" and "cow". Hence an *inverse document frequency* factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

2. Explain role of AI in IR

Knowledge representation. IR's representation of entities and relations is very weak. "Concept names are not normalised, and descriptions are mere sets of independent terms without structure … Concepts and topics, term and description meanings are left implicit… The relation between terms is only association based on co-presence..." While, the representation in AI is strong. There already exist various full-fledged methods and techniques to model the knowledge. Ontology can be considered as the generic term for generalising these representation ideas.

Reasoning. The weak reasoning in IR is "looking at what is in common between descriptions and preferring one item over another because more in shared (whether as different words or, via weighting, occurrences of the same word)… The probabilistic network approach, that allows for more varied forms of search statement and matching condition, does not alter the basic style of reasoning." While development in knowledge representation of AI, especially ontology provides the backbone for reasoning and also guarantees the reasoning.

Learning. Loosely speaking, the relevance feedback of IR can be considered as forms of learning. This again is very weak in IR. In this part, machine learning will link the IR and AI together to improve both sides

3. Explain about boolean model

Model is an idealization or abstraction of an actual process

• Mathematical models are used to study the properties of the process, draw conclusions, make predictions

• Conclusions derived from a model depend on whether the model is a good approximation of the actual situation

• Statistical models represent repetitive processes, make predictions about frequencies of interesting events

• Retrieval models can describe the computational process – e.g. how documents are ranked – Note that how documents or indexes are stored is implementation

• Retrieval models can attempt to describe the human process – e.g. the information need, interaction – Few do so meaningfully

• Retrieval models have an explicit or implicit definition of relevance

4. Explain about vector and cosine similarity model

Vector space model or **term vector model** is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. The name derives from the term "direction cosine": in this case, note that unit vectors are maximally "similar" if they're parallel and maximally "dissimilar" if they're orthogonal (perpendicular). This is analogous to the cosine, which is unity (maximum value) when the segments subtend a zero angle and zero (uncorrelated) when the segments are perpendicular.