SEIR IA-2 (15)

1. Is PageRank a query dependent or a query independent measure? Why? **(1.5)**

   Ans: *It is a query independent measure.*

   The working principle of PageRank is independent of any query. It evaluates the importance/significance of web pages instead of analyzing the content of the query. PageRank algorithm calculates a score of each webpage that represents its importance. This is done even before considering any query for search. Initially it counts the number and quality of links to a page. This is a strong indicator of each webpages's importance.It is often seen that the important websites are more likely to have inlinks and outlinks to other popular webpages.

2. Why is an inverted index named "inverted"? **(1.5)**

   Ans: *Normally a document is stored as the list of words that appear in that document, however inverted indices are the inverse of normal storage, i.e., they store the list of documents in which a word/term appears.*

3. What are some common pieces of information stored in an inverted index? **(1.5)**

   Ans: An inverted index stores, for a term, a list of document-IDs that contain that terms. In addition, it may store the counts of the term in the document as well as positions of all occurrences.

4. Encode 1033 using V-Byte Encoding. Show your steps. **(1.5)**

   Ans: The binary representation of 1033 is 10000001001

   Here, the least significant seven bits are (0001001) and the remaining 4 bits are (1000)

   Since 0001001 are the least significant bits, we add a 1 to the beginning of it.

   Hence it becomes 1 0001001

   Now, we pad out (1000) to 7 bits by adding 000 at its beginning which gives us 0001000.

   Since there are no more bits remaining, we add a before (0001000) and it gives us the following

   0 0001000

   Now we combine these representations into one as follows:

   0 0001000 1 0001001

   This can be converted into hexadecimal as

5. Are skip lists more efficient for conjunctive queries (AND) or disjunctive queries (OR)? Why? **(1.5)**

   Ans: Skip lists are more efficient for conjunctive queries. All words need to be present to satisfy conjunctive queries. In contrast, a disjunctive query needs to satisfy only one of multiple conditions to be true. Inverted lists are well-suited for ordered retrieval, and jumping between multiple lists for conjunctive (AND) conditions. Skip lists provide facility of going through multiple lists parallelly by skipping a large portion of a list during the traversal. This results in efficient retrieval that satisfies all the conditions.

6. Describe the difference between a user information need and a user query? **(1.5)**

   Ans: A user information need is the requirement of specific information that a user is looking for. It can refer to the user's intention for retrieving information for a certain purpose such as seeking knowledge, solving a problem etc.

   On the other hand, a user query is the representation of the user's information need in terms of search keywords. It can be considered as the medium through which a user's information need is transferred from her brain to the search engine.

7. What is the difference between query based stemming and document based stemming? **(1.5)**

   Ans: In Query-based stemming, the decision about stemming is taken during query time instead of during indexing. In this approach, the documents are not stemmed but the query is expanded using word variants. For example, in the query “rock climbing”, the term “climbing” can be expanded with “climb”, a word variant of “climbing”. Query-based stemming improves flexibility and effectiveness.

   In contrast, document-based stemming follows the approach of stemming the terms during indexing.

8. A boolean query retrieves a set of documents, using appropriate mathematical representations for the retrieved and the relevant set of documents, define Recall and Precision for your search results. **(1.5)**

   Ans: Recall measures the ability of the search system to retrieve all the relevant documents out of the total relevant documents available. It is defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents.

   Therefore,

   Recall = $\frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Number\ of\ Relevant\ Documents}$

On the other hand, Precision measures the relevancy of the retrieved documents. It is defined as the ratio of the number of relevant documents retrieved to the total number of documents retrieved

Therefore,

Recall = $\frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Number\ of\ Documents\ Retrieved}$

Mathematically, we can represent the sets as follows:

TP (True Positives): Number of documents that are both retrieved and relevant.

FP (False Positives): Number of documents that are retrieved but not relevant.

FN (False Negatives): Number of documents that are relevant but not retrieved.

With these definitions, we can express Recall and Precision as:

Recall = $\frac{TP}{TP+FN}$

Precision = $\frac{TP}{TP+FP}$

These measures are often used together to evaluate the performance of an information retrieval system. High recall means that the system retrieves most of the relevant documents, while high precision indicates that the retrieved documents are mostly relevant. Balancing these two metrics is crucial in designing effective search algorithms.

9. Assuming your ranking retrieves documents in the following sequence for a query that has four relevant document:

   N, R, R, N, N, R, N, N, N, R

   Here R is a relevant document and N is a non-relevant document. Compute average precision without and with interpolation (11-point average 0, 0.1, 0.2, … 1.0). **(3)**

   Ans: We will first calculate precision (P) without interpolation.

Precision when first relevant document (of 4) is retrieved (Recall 25%) = ½ = 0.5

Precision when second relevant document (of 4) is retrieved (Recall 50%) = ⅔ = 0.67

Precision when third relevant document (of 4) is retrieved (Recall 75%) = 3/6 = 0.5

Precision when fourth relevant document (of 4) is retrieved (Recall 100%) = 4/10 = 0.4

Average precision = (0.5 + 0.67 + 0.5 + 0.4) / 4

= 2.07/4

≈ 0.52

Highest Precision at any recall point is at R=0.5 and is 0.67. So precision at all R <= 0.5 will be 0.67:

P@0 = 0.67, P@0.1 = 0.67, P@0.2 = 0.67, P@0.3 = 0.67, P@0.4 = 0.67, P@0.5 = 0.67

Highest precision after recall 0.5 is at recall 0.75 which is 0.5. So

P@0.6 = 0.5, P@0.7 = 0.5

Highest precision after that (at recall 1) is 0.4. So

P@0.8 = 0.4, P@0.9 = 0.4, and P@1 = 0.4

Interpolated average precision = (0.67*6 + 0.5*2 + 0.4*3) / 11 = 0.5654