-Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).called

a- Data Mining b-Data Analysis c-Data Extraction d-Information retrieval

e- Web Mining

- Is the process of analyzing dense volumes of data to find patterns, discover trends, and gain insight into how that data can be used.

a- Data Mining b-Data Analysis c-Data Extraction d-Information retrieval e-Web Mining

-Is the process of using data mining techniques and algorithms to extract information directly

a- Data Mining b-Data Analysis c-Data Extraction d-Information retrieval

e- Web Mining

-Sometimes called a "spider," is a standalone bot that systematically scans the Internet for indexing and searching for content.called

a- web crawler b- Crawling c- web scraper-

-Is a process of extracting specific data . called (a searches for specific information on specific websites or pages)

a- web crawler b- Crawling c-web scraper

-Collection: a set of documents

-Goal: retrieve documents with Information that is relevant to the user's information need and help the user to complete a task

-This process is commonly referred to as grepping through text, after the Unix command grep, which performs this process

1- The process that involves retrieval of data from various sources in order to process it further is

called:

a- Data Mining b-Data Analysis c-Data Extraction d-Information retrieval e-Web Mining

2- The automated retrieval of specific information related to a selected topic from bodies of text is called

a-Crawling b-Data Extraction c-Data Mining

d-Information Extraction e-Data Analysis

3- The Goldberg machine is a …………………across catalog entries stored on a roll of microfilm. Machine that searched for a pattern of dots or letters

a-Mechanical b-Electronic c-Laser d-Digital e-Magnetic Tape

4-…………. is the topic about which the user desires to know more

a- A query b-An information need c-A user task d- A misconception e. A misformulation

5- …………..is what the user conveys to the computer in an attempt to communicate the information need.

a-A query b- An information need c-A user task d- A misconception e. A misformulation

6- if the result is called with respect to his information need. that means the user perceives as containing information of value

a. valid b. complete c. reasonable d. relevant e. incomplete

7- The fraction of the relevant documents in the collection were returned by the IR system is called---

a-recall b. precision c. f-measure d. relevance e. soundness

8- The fraction of the returned results are relevant to the information need is called-

a. recall b.precision c. f-measure d. relevance e. soundness

9- Consider Grepping: It is NOT true that:

a. It is a very effective process b. grep is a UNIX command

c. Impractical for near queries

d. good for ranked retrieval e. allows useful possibilities for wildcard pattern matching

10- The Boolean Retrieval model is a --------

a. model for information retrieval b. model that views a document as a set of sentences

c. data model d. good model for ranked retrieval e. a model for ranked retrieval

A **forward index** (or just index) is the list of documents, and which words appear in them.

The **inverted index** is the list of words, and the documents in which they appear.

1-What is an inverted index in information retrieval?

a) A data structure that stores documents in their original order

b) A data structure that maps terms to the documents in which they appear

c) A data structure that arranges documents in alphabetical order

d) is a data structure used to quickly locate documents or records that contain specific keywords.

Which of the following steps is involved in constructing an inverted index?

a) Tokenization

b) Stemming

c) Stop word

d) All of the above

What is **Tokenization** in the context of constructing an inverted index?

a) The process of breaking a document into words or tokens

b) The process of sorting documents based on their content

c) The process of mapping terms to the documents in which they appear

d) The process of removing stop words from a document

What is **Normalization** in the context of constructing an inverted index?

a) The process of breaking a document into words or tokens

b) The process of sorting documents based on their content

c) The process of Map text and query term to the same form

d) The process of removing stop words from a document

What is **Stemming** in the context of constructing an inverted index?

a) The process of breaking a document into words or tokens

b) The process We may wish different forms of a root to match

c) The process of Map text and query term to the same form

d) The process of removing stop words from a document

What is **Stop word** in the context of constructing an inverted index?

a) The process of breaking a document into words or tokens

b) The process We may wish different forms of a root to match

c) The process of Map text and query term to the same form

d) The process of removing stop words from a document .or We may omit very common words (or not!)

Which of the following is a benefit of using an inverted index?

a) Reduced storage requirements

b) Improved query performance

c) Efficient ranking of documents

d) All of the above

What is a posting list in the context of an inverted index?

a) A list of documents that contain a specific term

b) A list of stop words in a document

c) A list of terms present in a document

d) A list of documents sorted by their relevance to a query

Which query type is suitable for retrieving documents containing multiple terms?

a) AND query

b) OR query

c) Phrase query

d) Proximity query

What is term frequency-inverse document frequency (TF-IDF)?

a) A measure of the importance of a term in a document

b) A measure of the importance of a document in a collection

c) A measure of the relevance of a document to a query

d) A measure of the similarity between two documents

Which algorithm is commonly used for ranking documents in information retrieval systems?

a) PageRank

b) Cosine similarity

c) Levenshtein distance

d) Breadth-first search

Which of the following is NOT a technique for improving query performance in inverted indexes?

a) Index compression

b) Caching query results

c) Parallel processing

d) Stemming

Which data structure is typically used to store inverted indexes?

a) Hash table

b) Linked list

c) B-tree

d) Trie

1. The goal of the Extended Boolean model?

   is to overcome the drawbacks of the Boolean model that has been used in information retrieval.

-The Boolean model doesn't consider term weights in queries and the result set of Boolean query ,why?

a) is often either too small or too big.

2. In the Boolean Retrieval Model, the basic operations include:

   a) AND, OR, NOT

   b) AND, OR, XOR

   c) AND, OR, NEAR

   d) AND, OR, WITHIN

3. The Boolean Retrieval Model retrieves documents that:

   a) Exactly match the query terms

b) Partially match the query terms

c) Have similar meaning to the query terms

d) Are ranked based on relevance to the query terms

4. The Extended Boolean Model expands the basic Boolean model by:

a) Allowing the use of phrase queries

b) Introducing fuzzy matching

c) Incorporating relevance ranking

d) Enabling synonym matching

5. The Extended Boolean Model introduces which operator for proximity searching?

a) AND

b) OR

c) NEAR

d) NOT

6. The key advantage of the Extended Boolean Model over the basic Boolean Model is:

a) Greater precision in retrieval results

b) Faster query processing

c) Simpler query formulation

d) Improved scalability

-Query optimization is a process that aims to:

a) Improve the performance of the database system

b) Increase the security of the database system

c) Enhance the scalability of the database system

d) Ensure data integrity in the database system

Which of the following is NOT a common technique used in query optimization?

a) Indexing

b) Caching

c) Parallel processing

d) Data encryption

The primary goal of query optimization is to:

a) Minimize the execution time of the query

b) Maximize the number of rows returned by the query

c) Minimize the storage space used by the query result

d) Maximize the number of columns returned by the query

The query optimizer relies on:

a) Cost-based optimization techniques

b) Rule-based optimization techniques

c) Both cost-based and rule-based optimization techniques

d) None of the above

Which of the following factors can influence query optimization decisions?

a) Indexing strategies

b) Join order

c) Selection predicates

d) All of the above

In information retrieval, a phrase query refers to:

a) A query that consists of a single word

b) A query that retrieves documents based on their relevance score

c) A query that retrieves documents containing an exact sequence of words

d) A query that retrieves documents based on their term frequency

The positional index is used to:

a) Store the documents in a retrieval system

b) Rank the documents based on relevance scores

c) Store the positions of terms within documents

d) Process query terms to improve retrieval speed

Which of the following best describes the purpose of the positional index?

a) To provide a mapping between terms and documents

b) To store the term frequencies in the collection

c) To enable efficient processing of phrase queries

d) To calculate the relevance scores of documents

When performing a phrase query search using a positional index, the query terms must:

a) Appear in any order within the document

b) Appear in the same order and adjacent to each other within the document

c) Appear in the same order within the document, but not necessarily adjacent

d) Appear in any order and at any distance within the document

The inverted index is a type of index structure that:

a) Organizes documents based on their content

b) Stores the positions of query terms within documents

c) Ranks documents based on their relevance to a query

d) Represents the occurrence of terms across the entire collection

Which of the following is an advantage of using a positional index for phrase queries?

a) Faster retrieval speed compared to other index structures

b) Better handling of partial matches in query terms

c) Ability to rank documents based on their relevance

d) More efficient storage of the document collection

Which of the following operations is commonly used in a positional index to support phrase queries?

a) Intersection

b) Union

c) Complement

d) Difference

The size of the positional index can be affected by:

a) The number of terms in the collection

b) The number of documents in the collection

c) The average length of the documents

d) All of the above

In information retrieval, ranked retrieval refers to:

a) Sorting documents alphabetically

b) Ordering documents based on relevance to a query

c) Filtering out irrelevant documents

d) Categorizing documents into predefined topics

The most common ranking algorithm used in information retrieval is:

a) Boolean retrieval

b) Vector space model

c) PageRank

d) TF-IDF

The primary goal of ranked retrieval is to:

a) Retrieve all documents containing a query term

b) Retrieve the most recent documents

c) Retrieve the most relevant documents

d) Retrieve documents with the highest word count

Which of the following factors are commonly considered in ranking documents?

a) Term frequency

b) Inverse document frequency

c) Document length

d) All of the above

The cosine similarity measure is often used in:

a) Boolean retrieval

b) Probabilistic retrieval

c) Vector space model

d) PageRank

The term frequency-inverse document frequency (TF-IDF) weighting scheme assigns higher weights to terms that appear:

a) Frequently in a document and frequently in the collection

b) Frequently in a document and rarely in the collection

c) Rarely in a document and frequently in the collection

d) Rarely in a document and rarely in the collection

Precision at k and recall at k are commonly used evaluation measures in ranked retrieval. Which of the following statements is true?

a) Precision at k measures the proportion of relevant documents retrieved at rank k.

b) Recall at k measures the proportion of relevant documents retrieved at rank k.

c) Precision at k measures the proportion of retrieved documents that are relevant at rank k.

d) Recall at k measures the proportion of retrieved documents that are relevant at rank k.

The Discounted Cumulative Gain (DCG) metric is used to evaluate:

a) The relevance of retrieved documents

b) The ranking of retrieved documents

c) The recall of retrieved documents

d) The precision of retrieved documents

The Mean Average Precision (MAP) metric is commonly used to evaluate:

a) The relevance of retrieved documents

b) The ranking of retrieved documents

c) The recall of retrieved documents

d) The precision of retrieved documents

Relevance feedback is a technique used to:

a) Improve the precision of ranked retrieval

b) Improve the recall of ranked retrieval

c) Improve the ranking of retrieved documents

d) Improve the scalability of ranked retrieval

Free text queries: Rather than a query language of operator and expressions, the user's query is just one or more words in a human language.

The Jaccard coefficient is a measure of:

a) Document relevance

b) Document length

c) Term frequency

d) Set similarity

The Jaccard coefficient is calculated as the:

a) Intersection of two sets divided by the union of two sets

b) Union of two sets divided by the intersection of two sets

c) Intersection of two sets minus the union of two sets

d) Union of two sets minus the intersection of two sets

The Jaccard coefficient is often used to measure the similarity between:

a) Documents and queries

b) Terms and documents

c) Queries and search results

d) Relevance judgments and search results

The Jaccard coefficient ranges from:

a) -1 to 1

b) 0 to 1

c) 1 to infinity

d) 0 to infinity

Which of the following statements about the Jaccard coefficient is true?

a) It takes into account term frequency information.

b) It is sensitive to document length.

c) It is suitable for measuring similarity between ordered sequences.

d) It is unaffected by the order of elements in the sets.

The Jaccard coefficient is often used in which type of retrieval model?

a) Boolean retrieval

b) Probabilistic retrieval

c) Vector space model

d) PageRank

In Jaccard coefficient-based scoring, a higher value indicates:

a) Higher document relevance

b) Lower document relevance

c) Longer document length

d) Higher term frequency

The Jaccard coefficient is particularly useful for measuring similarity when dealing with:

a) Sparse documents

b) Long documents

c) Structured data

d) Categorical data

The Jaccard coefficient can be used to calculate the similarity between:

a) Two sets of terms

b) Two vectors of term weights

c) Two sequences of terms

d) All of the above

The Jaccard coefficient is often used in which stage of the information retrieval process?

a) Indexing

b) Query formulation

c) Retrieval scoring

d) Relevance feedback

-Issues with Jaccard for scoring

Rare terms in a collection are more information than frequent terms

- Jaccard doesn't consider this information

We need or sophisticate way of normalizing length

Later we will use= A intersection B/ root A U B

Bag of word model ?

Vector representation does not consider the ordring of words

Term frequency tfs?

d is defined as the number of times that T occurs in D.

wtd= {1+log10(tf)   tf>0

score= sum(1+ logtf)

What does IDF represent in the context of information retrieval?

a) Inverse Document Frequency

b) Important Document Frequency

c) Indexed Document Frequency

d) Informational Document Frequency

Does idf have an effect on ranking for one term queries?

No ,idf affect the ranking of documents for queries with at least two terms

What is the formula for calculating IDF?

a) IDF = log10(N / DF)

b) IDF = log(DF / N)

c) IDF = N / log(DF)

d) IDF = DF / log(N)

In tf-idf weighting, what does "tf" stand for?

a) Text Frequency

b) Term Frequency

c) Total Frequency

d) Token Frequency

What is the purpose of tf-idf weighting?

a) To determine the relevance of a document to a query

b) To count the total number of terms in a document

c) To calculate the number of unique terms in a document

d) To calculate the average term frequency across all documents

Which component of the Vector Space Model represents the importance of a term in a document?

a) Term Frequency

b) Document Frequency

c) Inverse Document Frequency

d) Term Weighting

Documents as Vectors:

VI-dimensional vector space

What does the Vector Space Model aim to achieve in information retrieval?

a) To determine the most relevant document for a given query

b) To measure the size of a document collection

c) To calculate the average length of documents in a collection

d) To determine the term frequency of individual terms

Which of the following best describes the term "term frequency"?

a) The number of occurrences of a term in a document

b) The number of documents containing a specific term

c) The total number of terms in a document collection

d) The number of unique terms in a document

Which weighting scheme assigns higher weights to rare terms and lower weights to common terms?

a) Inverse Document Frequency (IDF)

b) Term Frequency (TF)

c) Normalization

d) Binary Weighting

What does the IDF value indicate for a term?

a) The relevance of a term to a query

b) The uniqueness of a term across documents

c) The frequency of a term in a document

d) The length of a document containing the term

Which mathematical model is commonly used to represent documents and queries in the Vector Space Model?

a) Euclidean Space Model

b) Boolean Space Model

c) Vector Space Model

d) Weighted Space Model

- Formalizing vector space proximity

distance between two points (Euclidean distance)

-Euclidean distance is a bad idea...

.. Because Euclidean distance is large for vectors of different lengths

((المسافة عالية لو اطوال الفكتور مختلفة))

- Use angel instead of distance. (Key idea: Rank documents according to angel with query)

Which metric is commonly used to evaluate the effectiveness of ranked results in information retrieval?

a) Precision

b) Recall

c) F1 score

d) Mean Average Precision

Which of the following is true regarding precision and recall?

a) Precision is the ratio of relevant documents retrieved to the total number of documents in the collection.

b) Recall is the ratio of relevant documents retrieved to the total number of relevant documents in the collection.

c) Precision is the ratio of relevant documents retrieved to the total number of relevant documents in the collection.

d) Recall is the ratio of relevant documents retrieved to the total number of documents in the collection.

Mean Average Precision (MAP) is a metric commonly used in evaluating ranked results. What does MAP measure?

a) The average rank of relevant documents in the result set.

b) The average relevance score of documents in the result set.

c) The average precision at each rank position in the result set.

d) The average recall at each rank position in the result set.

In information retrieval evaluation, what does Discounted Cumulative Gain (DCG) measure?

a) The relevance of documents in the result set.

b) The diversity of documents in the result set.

c) The novelty of documents in the result set.

d) The gain of documents based on their position in the result set.

Which of the following metrics is commonly used to evaluate the top-k results in information retrieval?

a) Precision at k

b) Recall at k

c) Normalized Discounted Cumulative Gain (NDCG) at k

d) Mean Average Precision (MAP) at k

MAP : If a relevant document never gets retrieved?

we assume the precision corresponding to that relevant doc to be zero.

MAP is macro-averaging: each query counts equally.

Now perhaps most commonly used measure in research papers.

MAP :Good for web search?Why?

MAP assumes user is interested in finding many relevant documents for each query

MAP requires many relevance judgments in text collection

What does BERT stand for in the context of language modeling?

a) Bidirectional Encoder Representations from Transformers

b) Basic Encoder Representation Toolkit

c) Binary Encoding and Retrieval Technique

d) Biologically Enhanced Response Time

Which architecture is used in BERT to model the contextual relationships between words?

a) Recurrent Neural Network (RNN)

b) Convolutional Neural Network (CNN)

c) Long Short-Term Memory (LSTM)

d) Transformer

BERT was pre-trained on a large corpus of text from which source?

a) Wikipedia

b) Twitter

c) Reddit

d) GitHub

BERT is trained to predict missing words in a sentence using which task?

a) Named Entity Recognition (NER)

b) Part-of-Speech Tagging (POS)

c) Masked Language Modeling (MLM)

d) Sentiment Analysis

BERT's pre-training involves two steps: pre-training and ___________.

a) supervision

b) post-training

c) fine-tuning

d) validation

What is the advantage of BERT's bidirectional training compared to previous models like ELMo?

a) BERT can generate more accurate predictions.

b) BERT is faster in processing long sequences.

c) BERT captures contextual information from both directions.

d) BERT requires less computational resources.

How does BERT handle out-of-vocabulary (OOV) words during training?

a) It replaces them with a special token.

b) It ignores them and continues training.

c) It assigns them a random embedding.

d) It uses a character-based representation.

BERT's embeddings can be used as input to downstream tasks such as:

a) Sentiment analysis, question answering, and text classification, Abstract summarization, Sentence prediction,Conversational response generation

b) Image recognition, object detection, and speech synthesis.

c) Graph neural networks, reinforcement learning, and anomaly detection.

d) Compiler optimization, network security, and database management.

Which of the following best describes Neural Ranking in information retrieval?

a) A technique that uses artificial neural networks to rank search results.(We'll sort the results by decreasing score)

b) A method that uses natural language processing to organize information.

c) A process of indexing documents using neural network models.

d) A system that retrieves information based on user preferences.

What is the main advantage of using Neural Ranking over traditional retrieval methods?

a) Higher accuracy in ranking search results.

b) Faster processing speed.

c) Ability to handle larger datasets.

d) Improved user interface design.

Which of the following is an essential component of Neural Ranking systems?

a) Term frequency-inverse document frequency (TF-IDF) weighting.

b) Query expansion techniques.

c) Recurrent neural networks (RNNs).

d) Latent semantic indexing (LSI).

How does Neural Ranking contribute to personalized search experiences?

a) By tailoring search results to individual preferences and interests.

b) By enhancing the relevance of ads displayed during search.

c) By providing a broader range of search options.

d) By improving the user interface for mobile devices.

Which type of neural network architecture is commonly used in Neural Ranking models?

a) Convolutional Neural Networks (CNNs).

b) Long Short-Term Memory (LSTM) networks.

c) Transformer networks.

d) Multilayer Perceptrons (MLPs)