UNIT 1

INTRODUCTION TO INFORMATION RETRIEVAL SYSTEMS AND INFORMATION RETRIEVAL SYSTEM CAPABILITIES

Short Questions with Answers

Q1. Write a short note on Information Retrieval System?

Answer :

An information retrieval system is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text, images, audio, video and other multi-media objects. Although the form of an object in an Information Retrieval System is diverse, the text aspects has been the only data type that text it self to full functional processing.

Q2. Define precision and recall?

Answer :

Precision

Precision indicates the quality of the result set. It is given by the ratio of no. of relevant document retrieval to the total no. of documents retrieval. As system is said to be good in terms of precision it is capable of capturing nine relevant documents among the ten documents retrieved. Hence the precision is 0.9.

Recall

Recall is associated with the quantity of relevant documents that exist in the document set. It is given by the ratio of total relevant documents captured to the total documents that are assumed to be relevant is document set. It is necessary to involve the computation of total no. of relevant documents and involves the tedious task of reading the complete documents collection.

Q3. Define functional overview?

Answer :

A total information an storage and retrieval system is composed of four major functional processes

(i) Item Normalization

(ii) Selective dissemination of Information

(iii) Archival Document Database Search and

(iv) An Index Database Search

along with the Automatic file build process that supports index files.

Q4. Write a short note multi-media adds an extra dimension to the normalization process.

Answer :

In addition to normalizing the textual input, the multi-media input also needs to be standardized. There are a lot of options to the stand and being applied to the normalization. If the input is video the likely digital standards will be either MPE G-2, MPEG-1, AVI or Real media MPE (motion picture Expert Group) standards are the most universal standards for higher quality video other real media is the most common standard for lower quality video being used on the internet. Audio standards are typically WAV or Real Media (Real Audio). Image vary form JPEG to BMP. In all of the cases for multi-media, the input analog source is incoded into a digital format.

Q5. Write a note an document database search?

Answer :

The Document Database Search Process (See figure in Long Question 5)

Provides the capability for a query to search against all items recieved by the system. The document Data base search process is composed of the search process, user entered queries (typically and hoc queries) and the document database which contains all items that have been received, processed and stored by the system. Each query is processed against the total document database. Queries differ form profiles in that they are typically short and focused on a specific area of interest. The document Database can be very large, hundreds of millions of items or more.

Q6. Write a short note on multimedia Database search?

Answer :

Form a system perspective, the multi-media data is not logically its own data structure, but an augmentation to the existing structures is the Information Retrieval System. It will reside almost entirely in the area described as the document Database. The correlation between the multi-media and the textual domains will be either via time or positional synchronization. Time synchronization is the example of transcribed text form audio or composite video sources. Positional synchronization is where the multi-media is localized by a hyperlink in a textual domains will be either via time or positional synchronization. Time synchronization is the example of transcribed text form audio or composite video sources. Positional synchronization is where the multi-media is localized by a hyperlink in a textual item. Making the multi-media data part of the document database also implies that the linking of it to private and public index files will also operate the same way as with text.

Q7. Write a short note on libraries?

Answer :

As the quantities of information grew exponentially, libraries were forced to make maximum use of electronic tools to facilitate the storage and retrieval process. Since the collection is digital and there is a world wide communications infrastructure available, the library no longer must own a copy of information as long as it can provide access. The indexing in one of the critical disciplines in library science and significant effort has gone into the establishment of indexing and cataloging standards.

Q8. Write a note on search capabilities.

Answer :

The objective of the search capability is to allow for a mapping between a user's specified need and the items in the information database that will answer that need. The search query statement is the means that the user employes to communicate a description of the needed information to the system. It can consist of natural language text in composition style and/or query terms with boolean logic indicators between them. Once concept that has occasionally been implemented in commercial systems, and holds significant potential for assisting in the location and ranking of relevant items, is the "weighting" of search terms.

Q9. Define the Fuzzy searches?

Answer :

Fuzzy searches provide the capability to locate spelling of words that are similar to the entered search term. This function is primarily used to compensate for errors is spelling of words. Fuzzy searching increases recall at the expense of decreasing precision. A Fuzzy search on the term " computer" would automatically include the following words form the information database. "computer" "Compiter" "Conputer" "Computer" "Compute". An additional enhancement may lookup the proposed alternative spelling and it is a valid word with a different meaning, include it is the search with a low ranking.

Q10. Write a short note on natural language queries.

Answer :

Rather than having the user enter a specific Boolean query by specifying search terms and the logic between them, Natural Language Queries allow a user to enter a prose statement that describes the information that the user wants to find. The longer the prose, the more accurate the results returned. The most difficult logic case associated with Natural Language Queries is the ability to specify negation in the search statement and have the system recognize it as negation.

Q11. Define Browse capabilities?

Answer :

Once the search is complete, Browse capabilities provide the user with the capability to determine which items are of interest and select those to be displayed. There are two ways of displaying a summary of the items that are associated with a query : Line item status and data visualization from these summary displays, the user can select the specific items and zone within the items for display.

Q12. Write a short note on vocabulary browse?

Answer :

Vocabulary Browse provides the capability to display is alphabetical sorted order words from the document database. Logically, all unique words in the database are kept in sorted order along with a count of the no. of unique items in which the word is found. The user can enter a word or word fragment and the system will being to display the dictionary around the entered text.

UNIT

2

CATALOGING AND INDEXING, DATA STRUCTURES

Short Questions with Answers

1. Write a short note on history and objectives of indexing?

nswer :

To understand the system design associated with creation and manipulation of the searchable data structures, s necessary to understand the objectives of the indexing process. Retrieving the history of indexing shows the pendency of information processing capabilities on manual and then automatic processing systems. Through cost the 1980's the goals of commercial information retrieval systems were constrained facilitating the manual indexing radigm.

2. Write a short note on precoordination and linkage.

nswer :

Linkages are used to correlate related attributes associated with concepts discussed in an item. This process creating term linkages at index creation times is called precoordination. When index terms are not coordinated at dex time, the coordination occurs at search time. This is called postcoordination, that is coordinating terms after ost) the indexing process. Postcoordination is implemented by "AND" ing index terms together.

3. Define the indexing by term?

nswer :

When the terms of the original items are used as a basic of the index process, there are two major techniques r creation of the index: statistical and natural language. Statistical techniques can be based upon vector models d probabilistic models with a special case being Bayesian models. They are classified as statistical because their lculation of weights use statistical information such as the frequency of occurrence of words and their distribution the searchable database natural language techniques also use some statistical information but perform more mplex pursing to define the final set of index concepts.

4. Write a short note on stemming Algorithm?

nswer :

The concept of stemming has been applied to information systems form their initial automation in the 960's. The original goal at stemming was to improve performance and require less system resources by reducing e number of unique words that a system has to contain with the continued significant increases in storage and mputing power, use of stemming for performance reasons is no longer as important. Stemming is now being viewed for the potential improvements it can make in recall versus its associated decline in precision. A system signer can trade off the increased overhead of processing query terms with trailing "don't cares" to include all of eir variants. The stemming process creates one large index for the system versus term masking which requires the erging of the indexes for every term that matches the search term.

Q5. Write the dictionary look-up stemmers?

Answer :

An alternative to solely relying on algorithms to determine a stem is to use a dictionary look-up mechanism. In this approach, simple stemming rules still may be applied. The rules are taken from those that have the fewest exceptions. But even the most consistent rules have exceptions that need to be addressed. The original term stemmed version of the term is looked up in a dictionary and replaced by the stem that best represents it. This technique has been implemented in the INQUERY and retrievalWare systems.

Q6. Define the conclusions?

Answer :

Frakes summarized studies of various stemming studies. He cautions that some of the authors failed to report test statistics, especially sizes, making interpretation difficult. Also some of the test sample sizes were so small as to make their results questionable. Frakes came to the following conclusion:

Stemming can affect retrieval and where effects were identified they were positive. There is little difference between retrieval effectiveness of different full stemmers with the exception of the Hafer and Weiss Stemmer.

Stemming is as effective as manual conflation.

Stemming is dependent upon the nature of the vocabulary.

Q7. Write a short note on B-Tree inversion lists.

Answer :

Rather than using a dictionary to point to the inversion list, B-Tree can be used. The inversion lists may be at the leaf level or referenced in higher level pointers. Figure 2.4 shows how the words in figure 2.3 (See long Question 14) would appear. A B-tree of order m is defined as:

- A root node with between 2 and 2m keys.
- All other internal nodes have between m and 2m keys.
- All keys are kept in order from smaller to larger.
- All leaves are at the same level or differ by at most one level.

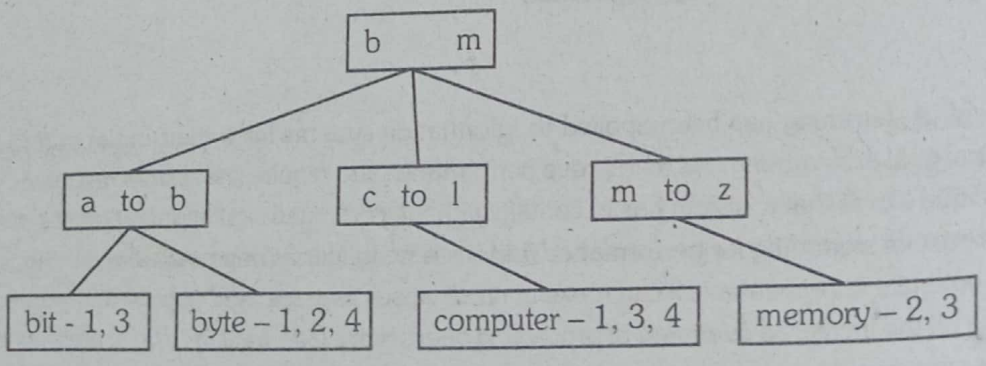

Figure: B-Tree Inversion Lists

Q8. List out the N-Grams uses?

Answer :

The uses of N-Grams are:

1. The mechanism was widely used for identifying and rectifying the errors in spellings.
2. It was used for performing the text compressions.
3. It was used for obtaining the information regarding authorization of the documents.

Q9. Write about advantage and disadvantage of N-Gram?

Answer :

Advantage

They place a finite limit on the number of searchable token. $Maxseg_n = (\lambda)^n$ maximum number of unique n - grams that can be generated. "n" is the length of n-grams λ number of process able symbols.

Disadvantage

Longer the n-gram the size of inversion list increase. Performance has 85% precision.

Q10. Write a short note an advantages applications of signature file?

Answer :

Signature files provide a practical solution for storing and locating information in a number of different situations.

Signature files have been applied as medium size database, databases with low frequency of terms, worm devices, parallel processing machines, and distributed environments.

Q11. Definition of hypertext structure?

Answer :

The hypertext data structure is used extensively in the internet environment and requires an electronic media storage for the item. Hypertext allows one item to reference another item via an imbeded pointer. Each separate item is called a node and the reference pointer is called a link. The referenced item can be of the same or a different data type that the original. Each node is displayed by a viewer that is defined for the file type associated with the node.

Example

Hpertext markup language (HTML) defines the internal structure for information exchange across the world wide web on the internet.

Q12. Write about the definition of a discrete hidden markov model and steps?

Answer :

A more formal definition of a discrete hidden markov model is summarized by Mittendorf and Schauble as consisting of the following:

1. $S = \{S_0, ..., S_{n-1}\}$ as a finite set of states where δ_0 always denotes the initial state. Typically the states a interconnected such that any state can be reached from any other state.

2. $V = \{V_0, ..., V_{m-1}\}$ is a finite set of output symbol. This will correspond to the physical output from th system being modeled.

3. $A = S \times S$ a transition probability matrix where aij represents the probability of transitioning from state i state j such that $\sum_{j=0}^{n-1} a_{i,j} = 1$ for all i = 0, ...>, n – 1. Every value in the matrix is a positive value between and 1. For the case where every state can be reached from every other state every value in the matrix will b non-zero.

4. $B = S \times V$ is an output probability matrix where element $b_{j,k}$ is a function determining the probability ar $\sum_{k=0}^{m-1} b_{j,k} = 1$ for all j = 0, ..., n–1.

5. The initial state distribution.

UNIT 3

AUTOMATIC INDEXING, DOCUMENT AND TERM CLUSTERING

Short Questions with Answers

Q1. Write a short note on statistical strategy?

Answer :

Covers the broadest range of indexing techniques and common in commercial systems. Basis for statistical is frequency of occurrence of processing token (words/phrases) within documents and within database. The phrases are the domain of searchable values. Statistics that are applied to the event data are probabilistic, Bayesian, vector spaces, neutral net.

Q2. What is the problems with the vector model?

Answer :

The problems with the vector model an assignment of a weight for a particular processing token to an item. Each processing token can be viewed as a new semantic topic. A major problem comes in the vector model when there are multiple topics being discussed in a particular item. For examples assume that an item has on in depth discussion of "oil" in "Mexico" and also "coal" in "Pennsylvania". The vector model does not have a mechanism to associate each energy source with its particular geographic area. There is no way to associate correlation factors between terms since each dimension in a vector is independent of the other dimensions.

Q3. Define natural language?

Answer :

The goal of natural language processing is to the semantic information in addition to the statistical information to enhance the indexing of the item. This improves the precision of searches, reducing the no. of false hits a user reviews. The semantic information is extracted as a results of processing the language rather than treating each word as an independent entity. The simplest output of this process results in generation of phrases that become includes to an item.

Q4. Write a short note on natural language processing?

Answer :

Natural language processing not only produce, more accurate term phases, but can provide higher level semantic information identifying relationship between concepts. System adds the functional processes relationship concept detectors, conceptual graph generators and conceptual graph matchers that generate higher level linguistic relationships including sematic and is course level relationship. During the first phase of this approach. The processing tokens in the document are mapped to subject codes. These codes equate to index term assignment and have some similarities to the concept based systems.

The next phase, is called the text structure, which attempts to identify general discourse level areas within an item.

Q5. Write a short note an clustering.

Answer :

Clustering of words originated with the generation of thesauri. Thesaurus, coming from the Latin wo meaning "treasure", is similar a dictionary is that it stores words. Instead of definitions, it provides the synony and antonyms for the words. Its primary purpose is to assist authors in selection of vocabulary. The goal clustering is to provide a grouping of similar objects into a "Class" under a more general little. Clustering also allo linkages between clusters to be specified. The term class is frequently used as a synonym for the term cluster.

Q6. Write the additional decisions for thesaurus?

Answer :

Word Coordination Approach

Specify of phases as well as individual terms are to be clustered.

Word Relationships

1. Equivalence, hierarchical, non-hierarchial
2. Parts-wholes, collocation, paradigmatic, Taxonomy and synonym, and Antonym
3. Constrated words, child - of, parent - of, is - part - of, has part.

Momograph Resolution

A homograph is a word that has multiple, completely different meanings.

Vocabulary Constraints

1. Normalization constrain the thesaurus to stems vs complete words.
2. Specificity : Eliminated specific words or use general terms for class identifies.

Q7. Write a note on thesaurus generation?

Answer :

1. Automatically generated thesauri contain classes that reflect the use of words in the corpus.
2. The classes do not naturally have a name, but are just a groups of statistically similar terms.
3. Basic idea for term clustering : The more frequently two terms co - occur in the same items, the more like they are abut the same concept.
4. Term - clustering, algorithms differ by the completeness with which terms are correlated.
5. The more complete the correlation, the higher the time and computational overhead to create the cluste

Q8. Write a short note a one pass assignments?

Answer :

1. Minimum overhead : Only one pass of all of the terms is used to assign terms to classes.
2. Algorithm
   - The first term is assigned to the first class
   - Each additional term is compared to the centroid of the existing classes.
   - A threshold is chosen. If the item is greater than the threshold, it is assigned to the class with the high similarity.
   - A new centroid has to be calculated for the modified class.

- If the similarity to all of the existing centroids is less than the threshold, the term is the first item in a new class.
- This process continues until all items a reassigned to classes.

Example with threshold of 10

Class generated

- Class 1 = Term1, Term 3, Term 4
- Class 2 = Term2, Term6, Term 8
- Class 3 = Term 5
- Class 4 = Term 7

Centroids Values Used

- Class 1(Term1, Term3) = 0, 7/2, 3/2, 0, 4/2
- Class 2(Term1, Term 3, Term4) = 0, 10/3, 3/3, 3/3, 7/3
- Class 3(Term 2, Term 6) = 6/2, 3/2, 0/2, 1/2, 6/2
- Minimum computation on the order of o(n)
- Does not produce optimum clustered classes
- Different classes can be produced if the order in which the items are analyzed changes.

Q9. Define item clustering?

Answer :

Clustering of items is very similar to term clustering for the generation of thesauri, manual item clustering is inherent is any library or filing system. In this case someone reads the item and determines the category or categories to which it belongs. When physical clustering occurs, each item is usually assigned to one category. With the advent of indexing, an item is physically stored in a primary category, but it can be formed in other categories as defined by the index terms assigned to the item.

Q10. List the objectives of creating a Hierarchy of cluturs?

Answer :

List

1. Reduce the overhead of search
2. Provide for visual representation of the information space.
3. Expand the retrieval of relevant items.

Q11. Write a short note on Hierarchical agglomerative clustering method (MACM)?

Answer :

For Refer answer from Long Q.No. 20.

Q12. Write about the analysis of MACM?

Answer :

1. Wards method typically took the longest to implement.
2. Single link and complete linkage are somewhat similar in run time.
3. Clusters found in the single link clustering tend to be fair broad in nature and provide lower effectiveness.
4. Choosing the best cluster as the source of relevant documents results in very close effectiveness results for complete link, ward's and group average clustering.
5. A consistent drop in effectiveness for single link clusturing in noted.

UNIT 4

USER SEARCH TECHNIQUES AND INFORMATION VISUALIZATION

Short Questions with Answers

Q1. Write a short note on the final level of binding.

Answer :

The final level of binding comes as the search is applied to a specific database this binding is based upon the statistics of the processing tokens in the database and the semantics used in the database. This is especially true in statistical and concept indexing systems. Some of the statistics used in weighting are based upon the current indexing systems. Some examples are document frequency and total frequency that are used as the basis for indexing are determined by applying a statistical algorithm against a representative sample of the database. Natural language indexing techniques tend to use the most corpora-independent algorithms.

Q2. Write about the similarity measures and Ranking?

Answer :

Searching in general is concerned with calculating in similarity between a user's search statement and the items in the database.

Restricting the similarity measure to passages gains significant precision with minimal impact on recall.

Once items are identified so possibly relevant to the user's query, it is best to present the most likely relevant items first-Ranking is a scalar number that represents how similar an item is to the query.

Q3. Define HMM?

Answer :

We present a new method for information retrieval using hidden Markov models (HMM). We develop a general framework for incorporating multiple word generation mechanisms within the same model. We then demonstrate that an extremely simple realization of this model substantially outperforms standard.

Q4. Write a short note on page Ranking Algorithms?

Answer :

With the growing number of web pages and users on the web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines needs to be more efficient in its processing way and output. Web mining techniques are employed by the search engines to extract relevant documents from the web database documents and provide the necessary and required information to the users. The search engines become very successful because of its page Rank algorithm. Page Rank algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some Ranking algorithms depend only on the link structure of the documents i.e., their popularity scores, where as other look for the actual content of the documents.

Q5. Write a short note on Distance Rank algorithm?

Answer :

The main goal of this ranking algorithm is computed on the basis of the shortest logarithmic distance between two pages and ranked according to them so that a page with smaller distance to assigned a higher rank. The advantage of this algorithm is that, being less sensitive, it can find pages faster with high quality and more quickly with the use of distance based solution as compared to other algorithms. If the some algorithms provide quality output then that has some certain limitations. So the limitation for this algorithm is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages. This distance rank algorithm adopts the page Rank properties i.e. the rank of each page is computed as the weighted sum of ranks of all incoming pages to that particular page. Then, a page has a high rank value if it has more incoming links on a page.

Q6. Write a note on probabilistic Relevance feedback?

Answer :

The probabilistic relevance feedback is based on the user feedback of documents relevance. So, after user can give relevance feedback of certain documents relevant or non - relevant. Then we can indicate it with a Boolean indicator variable 'R' of a document relevance. The probability of a term t to appear in a document relevance is,

$$\hat{p}(x_t = 1/R = 1) = |VR_t| / |VR|$$

$$\hat{p}(x_t = 0 / R = 0) = (df_t - |VR_t|) / (N - |VR|)$$

From the 'N' number of documents, where the number of df_t contains the term "t' within the known relevant document set 'VR' and subset VR_t, the estimation of probability is $p(x_t = 1)$. The assumption of the set of relevant documents as a small subset of the set of all documents makes the probability estimation $p(x_t = 1)$ possible.

Q7. Write about relevance feedback on the web?

Answer :

The relevance feedback on a web page is not providing by the users that much. Because the people on a web page using advanced search interfaces to complete their search in a single interaction, with respect to the time. The lack off up take also probably reflects the RF (Relevance Feedback) hard to explain to the average users and RF recall enhancing, and web users are, concerning rarely to go sufficient recall. From all of the web search users, only 4% of user query session using RF option, and about the 70% of the users only looked at the top first resultant page and did not pursue things any further.

Now - a - days a thread of work is the use of click stream data to provide indirect RF when a user click on a resultant link it automatically redirect to the other page or window. It is the most commonly using thread for RF indirectly.

Q8. Write the Indirect Relevance Feedback (RF)?

Answer :

Indirect RF is also called as implicit feedback then the explicit feedback approach. The implicit feedback collection easy from the users when it is difficult to collect in explicit relevante feedback on a web search.

This approach will work by collecting the click rates on pages globally, by the user's clicks. When a user want to use an relevant information then the user will click on that particular links to loss at more, frequently then those links were assumed to indicate that the pages was likely relevant to the query. This approach was introduced by Direct flit search engine. An example for this indirect feedback of relevance is clickstream and this is the of form of the general area of clickstream mining.

Q9. Explain in brief about relevance feedback?

Answer :

Relevance feedback is the most preferred utility that enables the user to retrieve the desired information in several passes. At every pass, the user enters a query and obtains a result set containing the documents that match the query. The user then selects the documents that are relevant to the generated query. Depending upon this selection the initial query will be upgraded by adding additional terms for better results. More over, the process of upgrading the query can also include assigning new rights to the existing terms depending upon the feedback of the user. In this way, the end of every pass, the query gets revised that ultimate leads to better search results.

Q10. Write a note on searching the Internet and Hypertext?

Answer :

The internet has multiple different mechanisms that are the basis for search of items. The primary techniques are associated with serves on the internet that create indexes of items on the internet and allow search of them. Some of the most commonly used nodes are YAHOO, Alta Vista and Lycos. In all of these systems there are active processes that visit a large number of internet sites and retrieve textual data which retrieve data and their general philosophy on user access. LYCOS (http : // www. lycos. com) and Alta Vista automatically go out to other Internet sites and return the text at the sites for automatically indexing (http : // www. alta vista. digital. com).

Q11. Write a short note on cognition and perception?

Answer :

The user - machine interface has primarily focused on a paradigm of a type writer.

As computers displays became ubiquitous, man - machine interfaces focused on treating the display as an extension of paper with the focus on consistency of operations.

The advent of WIMP interfaces and the evolution of the interface focused on how to represent to the user what is taking place in the computer environment.

Extending the HCI to improve the information flow, thus reducing wasted user overhead in locating needed information.

Although the major focus is an enhanced visualization of information, other senses are also being looked at for future interfaces.

The audio sense has always been part of simple alerts in computers.

The sounds are now being replaced by speech in both input and output interfaces.

The tactile (touch) sense is being addressed in the experiments using virtual Realty (VR).

Short Questions with Answers

Q1. Write a note on text search techniques ?

Answer :

The basic concept of a text scanning system is the ability for one or more users to inter queries, and text to be searched is accessed and compared to the every terms. When all of the text has been accessed, the query is complete one advantage of this type architecture is that as soon as an items is identified as satisfying a query, the results can be presented to the user for retrieval.

Q2. Write a short note one software text search.

Answer :

In Software Streaming techniques, the item to be searched is read into memory and then the algorithm is applied Although nothing in the architecture described above prohibits software streaming from being applied to many simultaneous search against the same item, it is more frequently used to resolve particular search against a particular item. There are four major algorithms associated with software text search. There are four approach, Kunth - Morris -putt, Boyer - Moore, Shift - oR algorithm, and Rabin-Karp.

Q3. Write the main features of Brute force algorithm.

Answer :

⇒ No Preprocessing phase.

⇒ Constant Extra Space needed.

⇒ Always Shifts the window by Exactly 1 position to the right.

⇒ Comparisons can be done in any order.

⇒ Searching phase in 0 (mn) time complexity.

⇒ 2n Expected text characters Comparisons.

Q4. Write the Karp -Rabin algorithm main features ?

Answer :

⇒ Uses an hashing function.

⇒ Preprocessing phase in 0 (m) time complexity and constant space.

⇒ Searching Phases in 0 (mn) time complexity.

⇒ 0 (n+m) Expected running time.

Q5. Give the main features of Shift -OR algorithm ?

Answer :

⇒ Uses bitwise techniques.

⇒ Efficient if the pattern length is no longer than the memory - ward size of the machine.

⇒ Preprocessing Phase in 0 (m + σ) time and space complexity.

⇒ Searching Phase in 0 (n) time complexity (independent from the alphabet size and the patter length).

⇒ Adapts easily to approximate string matching.

Q6. How does brute force work?

Answer :

A brute force attack is an illegal, "black - hat " attempts by a hacker to obtain a password or a PIN. It use several repetitive trial - and - error attempts to guess the password to break into a website or a service. Thes attempts are quick and vigorous and are carried out by bots.

Q7. What is brute force algorithm with example ?

Answer :

Brute force algorithms refers to a programming style that does not include any shortcuts to improv performance, but instead relies on sheer computing power to try all possibilities until the solution to a problem i found. A classic Example is the travelling salesmen problem (TSP).

Q8. What is the terms complexity of KMP algorithm?

Answer :

The Worst case complexity of the naive algorithms 0 (m(n – m + 1)). The time complexity of KMP algorithm is 0 (n) in the worst case. The Naive pattern searching algorithm doesnot work well in cases where we see many matching characters followed by a mis matching character.

Q9. What is Spurious hit in Rabin Karp algorithm ?

Answer :

The Rabin - Karp algorithm uses a rolling hash to detect the presence of a desired substring Because it's a hash function, it maps many different strings to the same hash value. If the rolling. Hash produces a candidate match due to this hash collision, which turns out not to be a string match, that is a "Spurious hit".

Q10. What type of algorithm is the Rabin and Karp algorithm ?

Answer :

The Rabin - Karp algorithm is a string - searching algorithm that uses hashing to find patterns in strings. A String is an abstract data type that consists of a sequence of characters. Letters, words, Sentences, and more can be represented as strings, String matching is a very important application of computer science.

Q11. Write a short note on Hardware text search system ?

Answer :

Software text search is applicable many circumstances but has encountered restrictions on the ability to handle many search terms Simultaneously against the same text and limits due to I/O speeds. One approach that off loaded the resource intensive searching from the main processors was to have a specialized hardware machine to perform the searches and pass the results to the main computer which supported the user interface and retrieval of hits. Since the search is hardware based, Scalability is achieved by increasing the no.of hardware search devices.

Q12. Write a short note on multimed in information retrieval ?

Answer :

Multimedia information retrieval is the process of satisfying a user's stated information need by identifying all relevant text, graphics, audio (speech and non-speech audio), imagery, or video documents or portions of documents from a document collection.