Answers to quizzes

5.1 INDEXING FOR INFORMATION RETRIEVAL

A posting indicates...

1. The frequency of a term in the vocabulary
2. The frequency of a term in a document
3. The occurrence of a term in a document
4. The list of terms occurring in a document

Answer 3

A posting corresponds to the occurrence (or location) of a term in a document. It may contain in its payload the frequency of occurrence of the term, so Answer 2 might be true in special cases, but not in general.

When indexing a document collection using an inverted file, the main space requirement is implied by ...

1. The access structure
2. The vocabulary
3. The index file
4. The postings file

Answer 4

The sizes of the access structure, the vocabulary and the index file all are proportional to the size of the vocabulary which follows for reald world datasets Heap's law and are therefore of order O(n^0.5). The postings file always has a size of order O(n).

Using a trie in index construction …

1. Helps to quickly find words that have been seen before
2. Helps to quickly decide whether a word has not seen before
3. Helps to maintain the lexicographic order of words seen in the documents
4. All of the above

Answer 4

When receiving the next word during the indexing process first the word is searched in the trie constructed so far. So the trie helps to decide whether the word is already known, and if it is known find the associated data. The true can also be used to derive the list of words in. the vocabulary in lexicographical order by traversing the trie. So no sorting of the vocabulary is required.

Maintaining the order of document identifiers for vocabulary construction when partitioning the document collection is important ...

1. in the index merging approach for single node machines
2. in the map-reduce approach for parallel clusters
3. in both
4. in neither of the two

Answer 1

In the index merging approach documents are traversed in order and the order is maintained when generating the postings for individual words, which is important for avoiding expensive sorting. In the map-reduce indexing approach postings are sent over the network to the reducer nodes in an arbitrary order, since the infrastructure is controlling the communication of the messages. Therefore, the reducer nodes have to reestablish the order of document identifiers.

When compressing the adjacency list of a given URL, a reference list

1. Is chosen from neighboring URLs that can be reached in a small number of hops
2. May contain URLs not occurring in the adjacency list of the given URL
3. Lists all URLs not contained in the adjacency list of given URL
4. All of the above

Answer 2

In fact, the reference list can contain URLs that are not contained in the page of the given URL. Answer 1 is wrong since the reference list is chosen from a window in the lexicographical ordering of URLs. Answer 3 does not make sense, as it says that the reference list contains all URLs, except those in the given page.

Which is true?

1. Exploiting locality with gap encoding may increase the size of (the representation of) an adjacency list
2. Exploiting similarity with reference lists may increase the size of (the representation of) an adjacency list
3. Both of the above is true
4. None of the above is true

Answer 1

It is possible that with a very unfortuntate distribution of integers the size of the representation might increase with gap encoding (though it is not straightforward to find such a counter-example).

The encoding using a reference list might naturally increase the size of the representation of the adjacency list, if the reference list is poorly chosen. However, in the practical implementation only reference lists will be chosen that lead to an improvement in storage size.

When applying Fagin's algorithm for a query with three different terms for finding the k top documents, the algorithm will scan ...

1. 2 different lists
2. 3 different lists
3. k different lists
4. it depends how many rounds are taken

Answer 2

For answering a query with three terms, the posting lists of the three terms need to be inspected, therefore 3 different lists have to be scanned.

With Fagin's algorithm, once k documents have been identified that occur in all of the lists ...

1. These are the top-k documents
2. The top-k documents are among the documents seen so far
3. The search has to continue in round-robin till the top-k documents are identified
4. Other documents have to be searched to complete the top-k list

Answer 2

In Fagin's algorithm round-robin traversal of posting lists stops once k documents have been seen in all lists. At this point also other documents may have been seen. The top-k documents are among all the documents that have been seen, but not necessarily those that have been seen in all lists.