Distributed Information Systems: Spring Semester 2015

Quiz 4: Inverted Files + Unstructured P2P

`Student Name: ______________________________________________ Date: 16 Apr 2015`
`Student ID: ________________________________________________ Time: 11:15AM to 11:30AM`

Total number of questions: 8

Each question has a single answer!

1. In the physical representation of an inverted file, the size of the index file is typically in the order of: (where n is the number of documents)

   □ *a)* $O(log(n))$

   ☒ *b)* $O(\sqrt{n})$

   □ *c)* $O(n)$

   □ *d)* $O(n^2)$

2. Which of the following statements is **true** about posting files?

   □ *a)* Merging posting files has logarithmic complexity in the size of the posting files. n log (n/M): log x linear

   □ *b)* The values stored in posting files are the weights of a term with regard to a specific document.

   □ *c)* The posting files are always split into chunks to speed up the look up.

   ☒ *d)* The space complexity of the posting files is proportional to the document collection size (considering that all words are indexed).

3. Which of the following statements is **true**?

   □ *a)* The Map-reduce framework solves the problem of space requirement of posting files by compressing them in a single file.

   □ *b)* For large number of index terms (i.e. thousands and more), the storage required for the index file becomes much larger than what is required for posting files.

   □ *c)* Inverted files have been developed as alternatives for boolean retrieval and vector space retrieval.

   ☒ *d)* Inverted files are not optimized for searching dynamically changing text collections.

4. Consider the posting lists L_1 and L_2 (with the respective p_1 and p_2 pointers) corresponding to a two-term query. Suppose that both L_1 and L_2 contain the same n documents appearing in order of their highest score, and L_2 is the same as L_1, but shifted by $n/2$ (i.e., $L_1 = [d_1, d_2, \ldots d_n]$ and $L_2 = [d_{n/2+1}, d_{n/2+2}, \ldots d_n, d_1, \ldots d_{n/2}]$ respectively). How many steps are required by the pointer p_1 to find the top-k documents in the sequential search phase of the Fagin's algorithm? (*both n and k are even numbers*)

   □ *a)* k ☒ *b)* $\frac{n+k}{2}$ □ *c)* $\frac{n}{2}+k$ □ *d)* $n-k$

   use simple example

5. Which of the following statements about hierarchical overlay networks is **true**?

   □ *a)* The request latency is increased compared to unstructured overlay networks.

   □ *b)* Fault-tolerance is always better compared to unstructured overlay networks.

   ☒ *c)* Message flooding incurs less messages compared to unstructured overlay networks.

   □ *d)* The storage cost is equally distributed over all peers in the network.

6. Which of the following quantities **doesn't** follow a power law distribution?

   ☒ *a)* The number of outgoing links from a node in an unstructured overlay network.

   □ *b)* The amount of resources that was contributed to the Gnutella network by different users.

   □ *c)* The number of incoming links to a node in an unstructured overlay network.

   □ *d)* The number of visits to a website.

7. Consider an unstructured network formed by N nodes, with each node having out-degree d. When a node issues a query, it forwards the query message to all the d nodes it is connected to (as long as the TTL is greater than zero), then every first-hop node forwards the message to its own d neighbors, and so on. *Assume that the links never point back to the nodes that have already been visited.*

   In such a scenario, the maximum number of query messages sent overall with the expanding ring algorithm (with a starting TTL of 1 and maximum TTL of TTL) is:

   □ *a)* $TTL^2 * d$

   ☒ *b)* $TTL * d^1 + (TTL - 1) * d^2 + \ldots + d^{TTL}$

   □ *c)* $d * TTL^1 + (d - 1) * TTL^2 + \ldots + TTL^d$

   □ *d)* $(d/N)^{TTL}$

8. Which of the following is wrong in the context of unstructured peer-to-peer networks:

   □ *a)* k-Random walkers can be used instead of flooding in order to reduce the number of messages sent overall.

   □ *b)* The expanding ring algorithm makes use of the fact that the resources and queries are likely to be power-law distributed.

   ☒ *c)* The expanding ring algorithm achieves a lower search latency compared to the flooding algorithm.

   □ *d)* Replicating a resource across nodes in the network will result in a lower number of message required to find that specific resource.