ITS 413 – QUIZ 5 ANSWERS

First name: _____ Last name: _____

ID: _____ Total Marks: _____
out of 10

- Write your name and ID in the space provided at the top of the sheet.
- Answer the questions on this sheet(s) only, using the space given.
- When asked to *describe* or *explain* something, your answer must be clear, concise and unambiguous. Usually about 1 to 4 sentences.

Question 1 [4 marks]

List the name of any four of the five main components of a search engine architecture and describe briefly what they do.

Answer (4 of the 5 needed):

Crawler: traverses web pages by following links and collecting the web pages it visits

Indexer: collects information obtained from crawler and extracts content to be indexed, for example keywords.

Retrieval Engine: Parses the user's search query and then combines the other components to display results

Ranker: rank pages in order of importance.

Databases: stores the content and indexes obtained

Question 2 [2 marks]

Describe connectivity-based ranking and content-based ranking, and give an example of a criteria for each.

Answer:

Connectivity-based ranking: rank order of importance of pages based on the pages connections with other pages (e.g. number of links to a page), and users (e.g. number of hits of page).

Content-based ranking: rank order of importance of pages based on the content in the pages, and the relevance of that content to the user's search query. (e.g. number of keywords in page, location of keywords in page)

Question 3 [1 mark]

Describe a malicious technique that can be used to increase the ranking of a website.

Answers:

Keywords stuffing: Include many (hundreds) keywords in title, metatags, page (including to topics that aren't on your site) (easy to detect)

Crawler traps: software/pages that try to keep crawler on site (so crawler thinks its important site (easy to detect)

Ghost sites: create many simple one-page sites that link to your site

Blogs/Wikis: include many links in blogs/wikis to your website

Question 4 [1 mark]

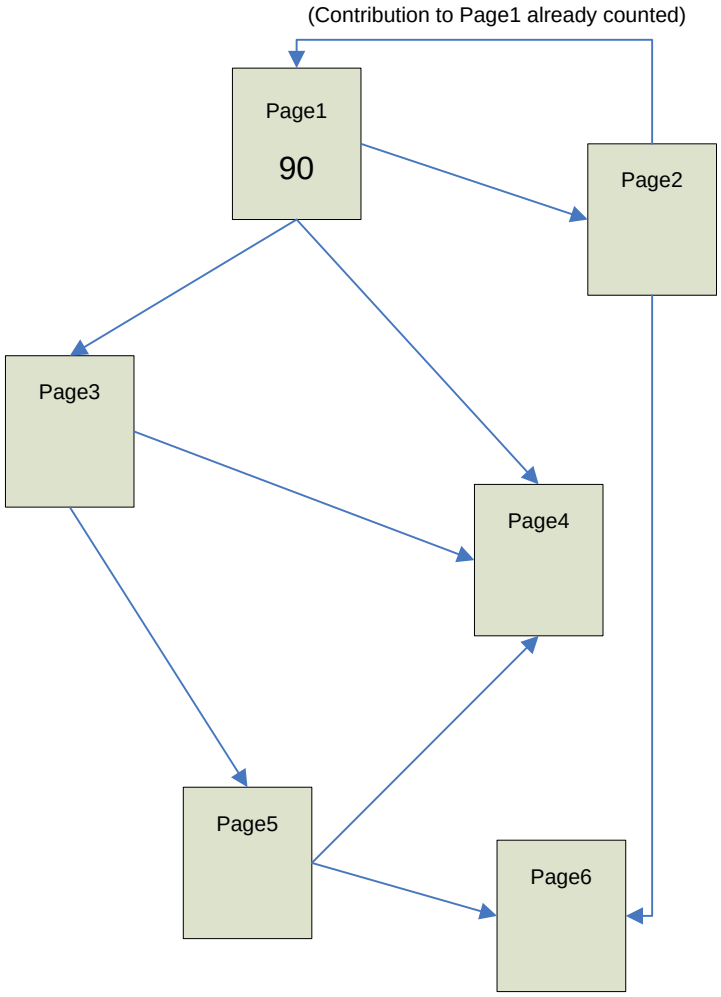
In comparison to client/server based systems, why is locating resources hard in P2P systems?

Answer:

There is no central node in a P2P system to store the index information (that is the mapping from resource to location); therefore the parts of the index is distributed across many nodes – locating resources requires locating the correct part of the index.

Question 5 [2 marks]

Calculate the PageRank for each page in the figure below, assuming that once the ranker has visited a page once to calculate the rank, it will not visit again. Also assume the PageRank of page 2 has already been counted towards the PageRank of page 1. You can write the PageRank of each page inside the page on the figure.



Answer: Page 3: 30; Page 4: 52.5; Page 2: 30; Page 5: 15; Page 6: 22.5