NTNU
Norges teknisk-naturvitenskapelige universitet

Fakultet for informasjonsteknologi, matematikk og elektroteknikk

Institutt for datateknikk og informasjonsvitenskap

Contact during the exam:
Dept. of Computer and Information Science
Heri Ramampiaro, 73593440/99027656

EXAM IN IT2801 INFORMATION RETRIEVAL

Thursday 24th of May 2007.
Duration: 09.00 – 13.00 (4 hours)

ENGLISH

Allowed aid: D – No printed or written materials allowed. **Only** approved calculator is allowed.

Result deadline: **14th of May 2007**.

Give **short and concise** answers to all questions. Short sentences are preferred rather than long explanation.

Problem I (25%)

1. Explain how the information retrieval process is built up. You must draw here a block diagram for this process. Hint: This does not deal with the text operations. (5%)
2. Why is the concept of ***index terms*** is important in information retrieval? What are the most important criteria for the choice of appropriate index terms? Justify your answer. (4%)
3. Explain which text operations one have to execute before an information search can take place? What are the purposes with these operations? (5%)
4. What are the main differences between ”***Suffix arrays***” and ”***Suffix tree***”? Explain briefly the principle behind them. You have to use examples or illustration to support your explanation. (5%)
5. “Inverted files” is an indexing method. Explain how you would use ”***inverted files***” to index documents. You may use an example in your explanation. (6%)

Problem II (25%)

1. Explain what are the main differences between ***multimedia*** and ***text*** retrieval. (3%)
2. Make taxonomy for the multimedia data model. Hint: The model is divided into several layers. You are expected to come up with at least one example of multimedia object type for each layer. (6%)
3. What are the purpose with features extraction? Give examples of features if we are talking about (a) image retrieval and (b) audio retrieval. (5%)
4. Audio can be classified to speech and music.
    a. Show how you can use the differences between the two to classify audio by using step-by-step classification. Draw a flowchart diagram to show the steps. (4%)
    b. Is possible to distinguish two music types (genre) based on the frequency spectrum? Justify your answer. (3%)

5. Provide a list of relevant features that can be used in video retrieval. What are the purposes with “segmentation” within video retrieval? (4%)

Problem III (30%)

1. Compare the probabilistic and vector space model. What model would you prefer if you were building a text retrieval system? You have to explain here the strength and the weaknesses of each of the models. (6%)
2. ***Precision*** and ***recall*** are two standards and most used measures to evaluate an IR system. What are the drawbacks in using these measures? To gain a better result, a concept of use-oriented measure has been suggested. Describe and explain the 4 relevant concepts here. (5%)
3. Explain briefly the principle behind User Relevant Feedback. You may use figure to support you explanation. (4%)
4. What are the main differences between ”automatic local analysis” and ”automatic global analysis”? Describe first when/where are these concepts used/relevant. (4%)
5. Construct a signature file for the following text. Assume that you can divide the text into 4 blocks. In any cases, you have to do any assumption on the signature function (and the like) you find appropriate: ”A tornado can cause sever destructions. Twister is a movie about a tornado” (6%)
6. Explain the principle behind a method for term re-weighting and query expansion for the vector space model. Hint: ”***Standard_Rochio***”. (5%)

Problem IV (20%)

Answer with ***correct/wrong*** (with a brief ***explanation***) on the following statements. Each **correct** and **justified** answer will be given **2** points. **Each wrong answer** gets **-1.5** point. While **unexplained** and **no answer** is **0** point.

1. Colour histograms are not useful for video retrieval.
(Correct/Wrong)
2. ”E-measure” can be seen as a generalisation of ”Harmonic Means”.
(Correct/Wrong)
3. A search engine with the ”Harvest” architecture is a kind of centralised web search engine architecture.
(Correct/Wrong)
4. Thesaurus is a tool for query refinements and is often used in connection to automatic local analysis.
(Correct/Wrong)
5. ”Gatherer” is often used as ”crawler” within a distributed web search engine.
(Correct/Wrong)
6. Image features cannot be used in video retrieval process.
(Correct/Wrong)
7. An SQL database is well suited for content-based image retrieval.
(Correct/Wrong)
8. Entropy is a concept that is often used to characterise the contents of information.
(Correct/Wrong)
9. With a Dictionary-based method we can achieve compression by replacing groups of following symbols (or phrases) with pointers to an entry in the dictionary.
(Correct/Wrong)
10. Micon is an important feature for images and used in image retrieval. It can be seen as analogous with r-frames within video retrieval.
(Correct/Wrong)