NTNU
Norges teknisk-naturvitenskapelige universitet

Fakultet for informasjonsteknologi, matematikk og elektroteknikk

Institutt for datateknikk og informasjonsvitenskap

Cantact durng the exam:
Dept. of Computer and Information Science
Robert Neumayer / Heri Ramampiaro, 73593440

IT2801 INFORMATION RETRIEVAL EXAM

Friday, 22th of May 2009.
Duration: 09.00 – 13.00 (4 hours)

ENGLISH

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

Result deadline: 12th of June 2009.

This exam was quality assured by Professor Kjetil Nørvåg.

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

Problem I (25%)

1. What are the purposes with **information retrieval (IR) systems**? Explain the differences between data and information retrieval.
2. What is the role of index terms in the context of information retrieval? Explain how the "inverted index works. Use illustration/example in your explanation.
3. Explain how the Boolean similarity model works. What are the reasons Boolean queries can have limited application areas?
4. Explain the challenges with the WWW seen from IR perspective.
5. Explain how an IR system can be evaluated and how can these be done?

Problem II (25%)

1. Explain the necessary steps that are required from a collection of documents is available to these documents are indexed (Hint: Indexing pipeline).
2. Explain the concept "*edit distance*". What is it used for? Assuming a text string s_i, where $len(s_i)$ is length of s_i Show that the "*edit distance*" between s_1 and s_2 can never be bigger than max($len(s_1)$, $len(s_2)$).
3. Featue is a central concept within multimedia IR. What are the purposes of features? Which three requirements must feature extraction meet? Provide examples of features for video IR.

4. Explain the concept "Jaccard Coefficient". What is the reason that this is not suitable for ranking of search results?

Problem III (30%)

1. Compare the **probability – Okapi BM25** and the **vcctor-based models – TF/IDF**. What model would you prefer if you were building a text retrieval system? Justify your answer. (Hint: emphasise on the principles, weaknesses, and advantages).
2. Test collections are often used in evaluation of IR systems. Give examples of existing test collections. How are they used? What is *R-precision, F-measure, and MAP (mean average precision)*?
3. Extension of queries.
    a. What are the aims of query extension? Explain.
    b. Explain briefly the principle behind **Rocchio's** method for User Relevant Feedback (URF). You may use a figure to support you explanation.
    c. What are the differences between automatic local analysis and automatic global analysis?
4. Explain two different indexing methods for indexing of text documents.
5. There are two possible ways to summarise search results. What are these, and how do they work?

Problem IV (20%)

Answer with correct/wrong 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** gives **0** point.

1. Pixel-to-pixel comparison of two images is suitable to compute the similarity between the two pictures.
   (Correct/Wrong)
2. Two audio files can be compared by using frequency spectrum of the audios.
   (Correct/Wrong)
3. R-frame and R-precision are two measures that can be used for evaluation of IR systems.
   (Correct/Wrong)
4. Statistical Thesaurus is a thesaurus that we can use to improve retrievals or extension of queries, when "*local automatic analysis*" is applied.
   (Correct/Wrong)
5. Harvest and Crawler have the same functions but used in two different web search engine architectures.
   (Correct/Wrong)
6. Video information cannot be retrieved by methods made for images or audio.
   (Correct/Wrong)
7. Ranking of results in vector-based and probabilistic-based retrieval system use the same principle.
   (Correct/Wrong)
8. Audio retrieval systems can use techniques known from text retrieval systems.
   (Correct/Wrong)
9. Movement information is not useful as feature for retrieval of video.
   (Correct/Wrong)
10. Image histogram can be used to retrieve images. In addition, it can be used in video retrieval.
    (Correct/Wrong)