IV2002 Internet Search and Monitoring Techniques 7.5 credits

Teknik för internetsökning och omvärldsbevakning

This is a translation of the Swedish, legally binding, course syllabus.

Establishment

Course syllabus for IV2002 valid from Autumn 2009

Grading scale

A, B, C, D, E, FX, F

Education cycle

Second cycle

Main field of study

Specific prerequisites

For single course students:

Prerequisites (Förkunskaper): Apart from a completed upper secondaryeducation and very good knowledge of English, basic knowledge of databases and programming is recommended. It is assumed that the students use the Internet and search engines on a regular basis.

Language of instruction

The language of instruction is specified in the course offering information in the course catalogue.

Intended learning outcomes

The course gives an insight into the techniques for information searching and monitoring applied on the Internet. After the course is finished, the students should be able to:

- Compare the different models of information retrieval (IR) and explain their pro and cons.
- Explain how a Business Intelligence system works, strength and weaknesses.
- Measure the quality of an information retrieval tool.
- Explain and choose among different approaches and techniques of automated query analysis.
- Make a specification of a Business Intelligence (BI) system that fulfills certain requirements.
- Use the terminology and concepts in information retrieval and business intelligence
- Know how to convey this information to fellow students/colleagues.

Course contents

Fundamentals of Information Retrieval: Boolean, term weight- and vector-space text retrieval models; document similarity measures; quality measures - precision and recall; index of documents and its access methods; morphologic and semantic analysis in text retrieval.
Query analysis: Processing the search word and index using word stemming, query expansion, fuzzy matching, compound splitting and compound joining that increase the quality of search. Other techniques are automatic translation of search words to other languages to make cross language information retrieval.
Information clustering and presentation: Sorting of text flows using automatic clustering and semi automatic clustering. Automatic document summarization removes redundant information from a document and creates a shorter summarized document. Multi document summarization summarizes several documents to one document. Using machine translation to present results in the users native language.
Search Engines: Architecture of a search engine; crawlers and features that hinder crawling; keyword-based retrieval; link analysis and PageRank; optimization of websites for search engines (Search Engine Optimization) and search engine spamming; paid listing; meta-search engines; web directories. Furthermore, there exist authoritative information accessible over the Internet and not visible to ordinary search engines. This material resides on the "invisible web", which is largely comprised of content-rich databases from universities, libraries, associations, businesses, and government agencies.
Monitoring tools: News archives and indexing tools, news alerts and agents, and RSS based news surveillance tools.
Question-Answering Systems deliver the answer to the question the user has in mind while searching, instead of a ranked list of documents. The three main question-answering approaches are based on Natural Language Processing, Information Retrieval, and question templates.

Disposition

Half speed
Credits (p): 7,5
Lectures: 14 lectures x 2 hours
Assignments: 3
Laborations: 3 occasions x 2 hours
Groups
Laborations are carried out in groups of maximum two students. Assignments in groups of maximum four students.
Laborations are carried out at university at fixed times under supervision of the course managers.
The assignments are carried out at home but there are occasions of supervision where the students can ask questions and get support from the teacher.
Distance students
The distance student must only participate physically for the exam, the rest of the tasks are solved completely at distance.
The distance education students must be present at the campus for the exam, the rest of the tasks can be solved using electronic means of communication.
If a distance student has no possibility to form a group then the student is allowed to solve all tasks alone, http://www.dsv.su.se/~eriks/66BI/66BIdist.html

Course literature

Preliminary:

- R. Baeza-Yates, B. Ribeiro-Neto: Modern Information Retrieval, Harlow Addison-Wesley, 1999
- Våge, L., Dalianis, H. och Iselid, L.: Informationssökning på Internet (Upplaga: 1:a), Studentlitteratur, 2003, 91-44-03178-5
- Mark Levene: An Introduction to Search Engines and Web Navigation, Addison Wesley, 2005
- Mike Moran, Bill Hunt: Search Engine Marketing, Inc., IBM Press, 2005

Course compendium.

Examination

- INL1 - Assignment, 4.5 credits, grading scale: P, F
- TEN1 - Examination, 3.0 credits, grading scale: A, B, C, D, E, FX, F

Based on recommendation from KTH's coordinator for disabilities, the examiner will decide how to adapt an examination for students with documented disability.

The examiner may apply another examination format when re-examining individual students.

If the course is discontinued, students may request to be examined during the following two academic years.

Assignments: 3
Laborations: 3

Written examination
The exam has a grading of (A/B/C/D/E/Fx/F). The assignments and laborations together correspond to 4.5 hp. To pass the course the student must pass the exam, the assignments, and the laborations. The assignments and laborations have the grading (P/F).
If the student is close to pass then we will give the possibility to make a complement of the examinations so the student can pass. The grading of the whole course is based on the exam.

Weighting of the grading
The grade of the written examination is the grade of the course. The laborations and assignments are compulsory but are not included in the grade.

Grading criteria: Written examination

A, VG+
Independent novel thinking using the basic concepts correctly to construct new functionality in IR and BI-system.

B, VG
Combine basic concepts to new functionality. Make detailed schema of IR and BI systems. All tasks carried out in time.

C, G+
Define all basic concepts in IR and BI. Describe the functionality of IR, query analysis, automatic summarization, clustering, (without errors).

D, G
Define all basic concepts in IR and BI. Describe the functionality of IR, query analysis, automatic summarization, clustering, (with few errors).

E, G-
Define some basic concepts in IR and BI. Describe the basic functionality of IR, query analysis, automatic summarization, clustering, (with some errors).

Fx, Rest (in English residue)
Define basic concepts erroneous, do not know basic functionality of IR, query analysis, automatic summarization, clustering.

F, U
Do not have knowledge in basic concepts.

Grading criteria: Assignments and laborations
Each assignment and laboration gets a grade either G or U, when obtaining a U the student has to make the task again. All assignments and laborations gets a G implies a G ('passed') on the Assignments and laborations part of the course.

Passed-P
Assignments and laborations fulfilled correctly.

Not passed-F (in English residue)
Not carried out assignments and laborations or lot of errors.

Ethical approach

- All members of a group are responsible for the group's work.
- In any assessment, every student shall honestly disclose any help received and sources used.
- In an oral assessment, every student shall be able to present and answer questions about the entire assignment and solution.