IV2038 Web-mining 7,5 hp

Web-mining

När kurs inte längre ges har student möjlighet att examineras under ytterligare två läsår.

Fastställande

Kursplan för IV2038 gäller från och med VT09

Betygsskala

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

Utbildningsnivå

Avancerad nivå

Huvudområden

Särskild behörighet

Undervisningsspråk

Undervisningsspråk anges i kurstillfällesinformationen i kurs- och programkatalogen.

Lärandemål

The course intends to give an insight into techniques for data mining applied on Internet related data, and for what they can be used. After the course is finished the student should be able to:
1. Identify and differentiate between application areas for web content mining, web structure mining and web usage mining.

2. Describe key concepts such as deep web, surface web, semantic web, web log, hypertext, social network, information synthesis, corpora and evaluation measures such as precision and recall.
3. Discuss the use of methods and techniques such as word frequency and co-occurrence statistics, normalization of data, machine learning, clustering, vector space models and lexical semantics.
4. In detail explain the architecture and main algorithms commonly used by web mining applications.
5. Appropriately select between different approaches and techniques of web mining for e.g. sentiment analysis, targeted marketing, linguistic forensics, topic/trend-detection-tracking and multi-document summarization (information aggregation).
6. Apply human language technology tools such as tokenizers, stemmers, part-of-speech taggers, noun phrase chunkers and shallow parsers on different types of web content gathered from for instance e-commerce sites.
7. Perform analysis of linguistically processed data using a suitable statistical classifier.
8. Set requirements to, compare and assess the quality of existing web mining tools.
9. Analyze and explain what web mining problems are satisfiably solved, what is worked upon at the research frontier and what still lies beyond the current state-of-the-art.
10. Independently solve a well defined practical web mining problem using tools and techniques introduced in the course, or analyze it through theoretical studies seeking information beyond the course literature.
11. Convey the outcome own work on web mining orally and in written form to fellow peers using relevant and appropriate terminology.

Kursinnehåll

Internet contains a huge amount of information, which is rapidly growing at an ever increasing pace. People, organizations and corporations from the whole world are adding different types of information to the web continuously in various languages. The web therefore contains potentially very interesting and valuable information. This course will investigate various techniques for processing the Web in order to extract such information, refine it and make it more structured, thus making it both more valuable and accessible. These techniques are often referred to as web mining techniques.
The domains within the Internet that we will study are databases, e-commerce web sites, wikis, virtual communities and blogs. Semantic web and Web 2.0 are two other concepts that are relevant for the course. Web mining is considered to contain three main areas, namely web content mining, web structure mining and web usage mining. Web structure mining is closely related to information search techniques, and web usage mining to opinion mining or sentiment analysis. Also related is the automatic construction of sociograms. Web content mining can for example be used to find the cheapest airline tickets, by monitoring all web based databases of all airlines in order to attempt to find the lowest common denominator of all databases.
Techniques
Web mining techniques explored in the course are human language technology, machine learning, statistics, information retrieval and extraction, text mining, text summarization, automatic classification, clustering, wrapper induction, normalization of data, match cardinality of data in different databases, interface matching, schema matching, sentiment analysis, opinion mining, extraction of comparatives, forensic linguistics etc.

Kursupplägg

Half speed
Credits: 7,5 hp
Lectures: approx. 12 lectures x 2 hours
Lab exercises: approx. 5 occasions x 3 hours
Project and seminar task: approx. 1 lecture x 2 hours

Kurslitteratur

Preliminary:

Bing Liu: Web Data Mining - Exploring Hyperlinks, Contents and Usage Data ISBN 354037881

Compendium of selected scientific papers.

Examination

- LAB1 - Laborationer, 2,5 hp, betygsskala: P, F
- PRO1 - Projektuppgift och seminarium, 2,5 hp, betygsskala: P, F
- TEN1 - Tentamen, 2,5 hp, betygsskala: A, B, C, D, E, FX, F

Examinator beslutar, baserat på rekommendation från KTH:s handläggare av stöd till studenter med funktionsnedsättning, om eventuell anpassad examination för studenter med dokumenterad, varaktig funktionsnedsättning.

Examinator får medge annan examinationsform vid omexamination av enstaka studenter.

Written examination part 1 validates goal 1-3
Written examination part 2 validates goal 3-5
Lab exercises validates goal 6-7
Project and seminar task validates goal 5, 8, 9 and 10-11
Tentamina/Written exams:
The written examination is divided into two parts, where the student must reach at least 75 % correctness on the first part in order to pass. The second part of the written examination determines, together with any bonus points from the lab exercises, the passing grade A-E.
Laborationer/Laboratory work:
Lab exercises and seminar task are all carried out in groups of maximum two students. Lab exercises are carried out at the university at fixed times under supervision of the course managers. Single exercises can not be carried out after the current course ends. If the student does not finish all five exercises within the time frame of the course, the course managers may demand that the student carries out all exercises the next time the course is given.
Projektarbete/Project work:
The project and seminar task is carried out in groups of 1-3 students and is examined in two stages, where the first is an oral presentation at one occasion where fellow students can ask questions and criticize and the second as a written report of approximately 8-10 pages to be handed in. We encourage that Swedish and international students mix in groups in order to obtain higher language competence. Part of this task may contain some Swedish text.
Bonuspoäng/Bonus system:

We apply a bonus system in order to encourage the students to spread the study load evenly over the course. For each lab exercise that is finished in time 1 bonus point is awarded to the second part of the written examination. The deadline for each exercise is 11:59 pm the day the supervised lab exercise is scheduled, although it is recommended to present the solution to the exercise during the supervised lab. These bonus points are valid for one year, and are added to the points awarded on the written examination.

Övriga krav för slutbetyg

To pass the course the student needs to pass all the examination parts. The final course grade is based on the grade of the written exam.

Etiskt förhållningssätt

- Vid grupparbete har alla i gruppen ansvar för gruppens arbete.
- Vid examination ska varje student ärligt redovisa hjälp som erhållits och källor som använts.
- Vid muntlig examination ska varje student kunna redogöra för hela uppgiften och hela lösningen.