

Representing and Accessing Digital Information

30. August, 2004
Mondays and Wednesdays, 2:55-4:10
Rhodes 484

Outline:

- Instructor: Thorsten Joachims
- Overview of material covered
- Syllabus
- Reference Material
- Prerequisites
- Grading

Instructor

Thorsten Joachims

- 4153 Upson Hall
- Email: tj@cs.cornell.edu
- Office hours: Monday 16:15-17:00, Wednesday 16:15-17:00

Broad Research Interests:

- machine learning and statistical learning theory
- information retrieval

Examples:

- text classification
- information systems that learn by observing users
- similarity metrics for natural language
- predicting complex objects (e.g. trees, alignments)

Information Access Tasks Covered in CS630

- documents/texts in natural languages and semi-structured data
 - unknown and not predefined structure
 - could be in multiple languages
 - no or little operational semantics
- well defined tasks (classification, topic tracking, etc.)
- typically large quantities of data, for example
 - continuously analyzing the articles in all US newspapers
 - extracting information from the WWW
- methods perform the task without fully understanding the text
 - not full natural language understanding
 - use statistical techniques and machine learning
- user modelling
 - patterns in user behavior / homogeneous groups

Layers of Information

Content

- text in document
- images

Meta-data

- authorship
- creation time and date
- hyperlinks

Usage

- number of visits (over time)
- keywords used in search for document
- documents visited by same user in same session

Text Classification: Yahoo!



Text Clustering: Google News



Information Retrieval: Google

Google Advanced Search Preferences Language Tools Search Tips
support vector machine Google Search

Web Images Groups Directory
Searched the web for support vector machine Results 1 - 10 of about 370,000 Search took 0.07 seconds

Kernel Machines
Description: A central source of information on kernel based methods, including support vector machines, Gaussian...
Category: Computers > Artificial Intelligence > Neural Networks
www.kernel-machines.org/~tk - Cached - Similar pages

Support Vector Machines - The Book - Support Vector
... OCTOBER 2001: FOURTH REPRINT NOW AVAILABLE. This book is the first comprehensive introduction to Support Vector Machines (SVMs), a new generation learning...
Description: First comprehensive introductory book to the field of Support Vector Machines, a novel machine learning...
Category: Computers > Artificial Intelligence > Machine Learning > Publications > Books
www.support-vector.net/~tk - 25 Aug 2002 - Cached - Similar pages

svm.first.gmd.de -> www.kernel-machines.org [Translate this page]
svm.first.gmd.de/~tk - 25 Aug 2002 - Cached - Similar pages

Support Vector Machine
... Support Vector Machine. The most recent SVM light page can now be found at http://ovmsight.joachims.org/
Older versions are still available from here...
Description: Large-scale support vector machine training software.
Category: Computers > Artificial Intelligence > Neural Networks > Software
www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light_eng.html - 12k - Cached - Similar pages

SVM-Light Support Vector Machine
... SVM-Light Support Vector Machine Hier finden Sie Informationen zu den folgenden Themen: Thorsten Joachims, SVM-light, SVM-light, SVM-light, Support Vector...
ovmsight.joachims.org/~tk - 25 Aug 2002 - Cached - Similar pages

Question Answering: AskJeeves

AskJeeves.com No Pay Stubs, W-2's Or Income Tax Returns Required QuickStart Loans

Where can I buy a Sony Vaio notebook? ASK Shopping

WEB RESULTS NEWS RESULTS SHOPPING RESULTS

You may find this featured sponsor helpful:

OfficeMax - Sony Vaio Systems
OfficeMax.com - Extra savings on office workstations, office furniture, office supplies and electronics at unbelievable prices. Free shipping, price matching and mail-in rebates.
From: http://www.officemax.com

Information Extraction: Flipdog (I)

Find a Job - FlipDog.com - Microsoft Internet Explorer

FlipDog WAL-MART Get Certificate From TheUseshul

Step 1 Location: Where do you want to work?
Step 2 Category: What type of work?
Step 3 Employer: Which employer?

Keywords: Search, NY
Location: Search, NY
Category: All Categories
Employer: All Employers

115 jobs found

Information Extraction: Flipdog (II)

Job Search Results - FlipDog.com - Microsoft Internet Explorer

1 - 25 of 115 jobs shown below

Search these results for: Search Jobs Posted: For all time periods

View: Brief | Detailed

Standard Jobs: Employers have paid to increase the exposure of these great job opportunities.

Job Title	Posted by	Posted Date	Backlist
Bookkeeping Assistant	posted by Manager	August 15, 2003	Backlist
General Office Clerk	posted by Manager	August 13, 2003	Backlist
Technical Support Specialist w/ Database		August 26, 2003	Backlist
System Analyst at OSA		August 27, 2003	Backlist
Software Engineer at OSA		August 27, 2003	Backlist
Research Scientist at OSA		August 27, 2003	Backlist
Computer Security at OSA		August 27, 2003	Backlist
Travels Program Benefits at Cornell Employment and Family Careers Institute		August 27, 2003	Backlist
Industrialist at Graduate School, USGA		August 27, 2003	Backlist
Development Associate at Sciencenter		August 27, 2003	Backlist
Executive Director at Sciencenter		August 27, 2003	Backlist
Development Associate at Sciencenter		August 27, 2003	Backlist
Development Associate at Sciencenter		August 27, 2003	Backlist
Major Account Manager at The CBSD Group, Inc.		August 27, 2003	Backlist
Sales/Marketing/Media at The CBSD Group, Inc.		August 27, 2003	Backlist
Testing Analyst at The CBSD Group, Inc.		August 27, 2003	Backlist
Sales Engineer at The CBSD Group, Inc.		August 27, 2003	Backlist

Information Extraction: Flipdog (III)

Job Details - FlipDog.com - Microsoft Internet Explorer

FlipDog create a job hunter search for your search job what you want?

Software Engineer. Requires BS or graduate degree in computer science; or BS or graduate degree in mathematics or physics with significant programming experience. Significant academic achievement or three years of technical experience on software projects is preferred. Candidate will be expected to contribute to software design and implementation in an R&D environment. Good oral and written communication skills required. Background in developing applications in both Windows and Unix environments is preferred. Experience in Visual C++ or Java is preferred. Tasks may include simulation, GUI development, low level device drivers, or developing routing protocol applications.

System Analyst. Requires BS in Computer Science or equivalent experience. Candidate will be expected to conduct system assessment and vulnerability analysis. System administration experience on a variety of platforms (such as Windows/NT, Sun, Linux, BSD, IRIX, SMI, or HP) is preferred. Familiarity with automated security and network management tools as well as network protocols (e.g., SNMP, NetBIOS, SMTP, WINS, DNS) is a plus. Good interpersonal and communication skills are essential. Some travel may be required.

Employer: OSA
All jobs from employer

Job Info:
Last updated: August 27, 2003
Location: Ithaca, NY
Category: Computing/MS
Functions: Network/System Administration

Information Extraction: ResearchIndex

Text Categorization with Support Vector Machines: Learning with Many Relevant Features (1998) (More Corrections) (189 citations)

Proceedings of ECML-98, 10th European Conference on Machine Learning

View or download: [all.cs.umford.edu/~joachims_98a.ps.gz](#), [cornell.edu/People/~joachims_98a.ps.gz](#), [Cached](#), [ES.gz](#), [PS](#), [PDF](#), [DjVu](#), [Image](#), [Update](#), [Help](#)

From: [ai.infomark.uni-joeachims.eng.fhnw.ch](#)
Homepages: [T.Joachims](#), [HPSearch](#), [Update Links](#)

(Enter summary) Rate this article: 1 2 3 4 5 (prev) Comment on this article

Abstract: This paper explores the use of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyzes the particular properties of learning with text data and identifies why SVMs are appropriate for this task. Empirical results support the theoretical findings. SVMs achieve substantial improvements over the currently best performing methods and behave robustly over a variety of different learning tasks. Furthermore, they are fully automatic, eliminating the need for manual... [Updated]

Context of citations to this paper: [More](#)

... from the Canadian Parliament corpus (Hansard) was used in Support Vector Machine (SVM) text classification [2] of Reuters 21578 corpus [3] (Table 6) in this experimental setting the intersection of vector spaces of the Hansard, 5159 English words from the first 1000 French...

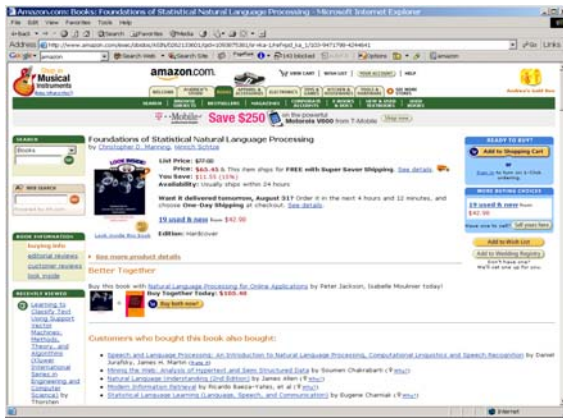
... efficiency and predictive accuracy [16] in recent years. Joachims has done much research on the application of SVM to text categorization [16]. His SVM system published via http: www. cs. uni. dortmund. de/ FORSCHUNG/ VERFAHREN/ SVM_ LIGHT/ svm. light. eng. html is used in our...

Cited by: [More](#)
JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, - Video And Web (Context)
Any A Tool For Automatic Report Generation - Marat Kasnikava And (Context)
Predicting The Sub-Cellular Location Of - Proteins From Text (Context)

Similar documents (at the sentence level):
41.3% Text Categorization with Support Vector Machines: Learning with... - Joachims (1998) (Context)

Active bibliography (related documents): [More All](#)
••• Text learning and related intelligent agents - Mladenic (1999) (Context)
••• Learning to Recognize 3D Objects - Roth, Yang, Ahuja (2000) (Context)
••• Text Categorization: A Survey - Asa, Elkel (1999) (Context)

Product Recommendations: Amazon.com



Overview of the Syllabus (I)

- **Information Retrieval Basics:** vector space model, inverted indexes, statistical properties of text, evaluation in information retrieval (4 lectures)
- **WWW Structure:** co-citation analysis, Pagerank (1-2 lectures)
- **Text Classification:** support vector machines, naive bayes, k-nearest neighbor, feature selection (4 lectures)
- **Text Clustering:** k-means clustering, hierarchical agglomerative clustering (2 lectures)
- **Latent Semantic Analysis:** (1 lecture)

Overview of the Syllabus (II)

- **Semi-Structured Data and Semantic Web:** schemas, XML databases and queries, XML information retrieval (1-2 lectures)
- **Information Extraction:** system architecture, hidden markov models, part-of-speech tagging, named entity detection, learning extraction patterns (3-4 lectures)
- **Usage Data:** clickthrough data, navigation paths, personalized retrieval functions (2 lectures)
- **Recommender Systems:** product recommendations, item-to-item similarity, user groups (2 lectures)
- **Document Summarization:** single- and multi-document summarization, summarization evaluation (1 lecture)

Support System

Handouts:

- readings
- slides
- homework assignments

Course WWW page:

- <http://www.cs.cornell.edu/Courses/CS630/2004fa>
- syllabus
- homework assignments / slides / research papers

Office Hours:

- Thorsten Joachims: 4153 Upson Hall, tj@cs.cornell.edu, Mondays 16:15-17:00, Wednesdays 16:15-17:00

Further Reference Material

- Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley, 1999.
- Christopher Manning and Hinrich Schütze. "Foundations of Statistical NLP", MIT Press, 1999.
- Ian H. Witten, Alistair Moffat, and Timothy C. Bell, "Managing Gigabytes: Compressing and Indexing Documents and Images", 2nd edition, Morgan Kaufmann, 1999.
- Karen Sparck Jones and Peter Willett (editors), "Readings in Information Retrieval", Morgan Kaufman, 1997.
- Thorsten Joachims, "Learning to Classify Text using Support Vector Machines", Kluwer, 2002.
- Tom Mitchell, "Machine Learning", McGraw Hill, 1997.

Prerequisites

Any of the following:

- CS472 Artificial Intelligence
- CS478 Machine Learning
- CS578 Emp. Methods in Machine Learning
- CS678 Advanced Topics in Machine Learning
- CS674 Natural Language Processing
- equivalent of any of the above
- permission from the instructor

Assignments

Homework

- ~3 homework assignments
- some programming, some conceptual
- some group work (I will make clear when group work is allowed)

Reading

- ~6 critiques of selected readings and research papers
- max. 1 page
- individual, not group work

All assignments due at start of class. Assignments turned in late will be penalized one full grade (e.g. A-->B) for every 24 hours of delay.

Copying / cheating / cooperating / helping may result in automatic failure of the course => academic integrity policy on WWW page.

Grading

Grades will be determined as follows:

- 25%: mid-term exam
- 25%: final project
- 30%: homework assignments
- 10%: critiques of selected readings and research papers
- 10%: class participation

Roughly: A=90-100; B=80-90; C=70-80; D=60-70; F= below 60