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
  - continously analyzing the articles in all US newpapers
  - 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

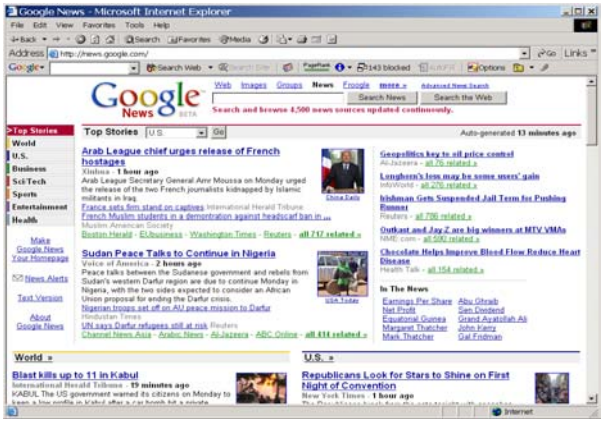
- authorship
- creation time and date
- hyperlinks

Usage

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

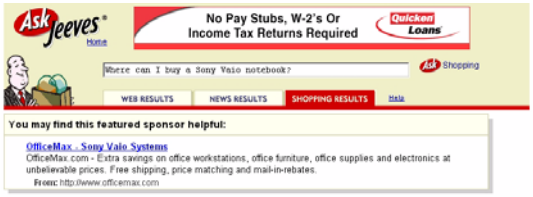
Text Classification: Yahoo!

Text Clustering: Google News



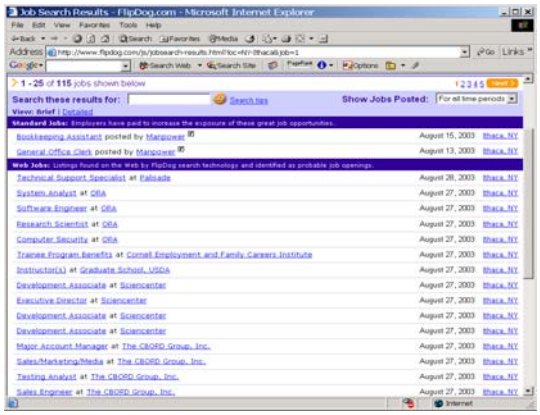
Information Retrieval: Google

Question Answering: AskJeeves

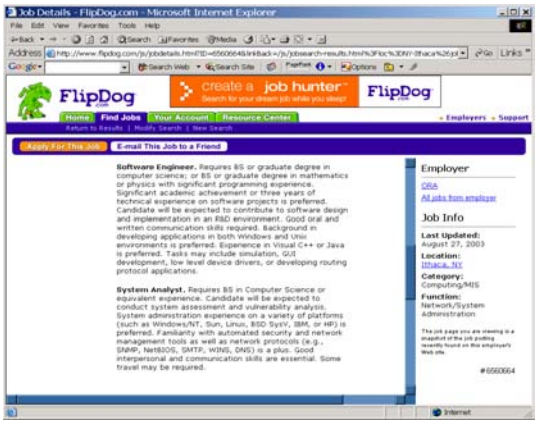


Information Extraction: Flipdog (I)

Information Extraction: Flipdog (II)



Information Extraction: Flipdog (III)



Information Extraction: ResearchIndex

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 Schutze. "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