Introduction to IR and IE

Vasudeva Varma

Goal of this course

- To introduce and provide *hands-on exposure* in the areas related to *Information Access*.

- To provide necessary background for potential *research* students in the areas of Information Access technologies.

Course Topics and Roadmap

- Introduction (4)
- IR Fundamentals (9)
  - Models,
  - Scoring functions
  - Index design
  - Crawling
  - IR Evaluation
- NLP/Text Mining for IR (4)
- Machine Learning & IR (9)
- Information Extraction (3)
  - IE Fundamentals
  - Named Entity Recognition
- Information Access and IR Applications (9)
  - Summarization
  - Social Computing

Instructors

- Vasudeva Varma
- Manish Gupta
- Niyati Chaya

- PhD students of IREL
  - Pulkit Parikh
  - Harika Abburi
  - Vijay Sarathi
  - …

Course Administration

- Teaching Assistants:
  - Teaching Associate: Vivek Anand
  - Mentors: All PhD/Senior MS Students of IREL
  - Teaching Assistants: TBA
- Tutorial: Every Thu/Fri ??
- TA Office Hours: TBA

Grading

- Quiz/In Class Activities: 10%
- Assignments 15%
- Project 60% (20+40)
- Term Paper 15%

Text/Reference Books/material

- Introduction Information Retrieval – Chris Manning et al (the Stanford IR Book)
- Search Engines IR in Practice – Bruce Craft et al

Projects

- Mini Project – Individual Project (4 Weeks)
  - Two deliverables
- Major Project – Teams with 3-5 members each (10 weeks)
  - Three deliverables

Mini Project – 4 Weeks

- Objective: Design and develop a scalable and efficient search engine using the Wikipedia data.
- Features:
  - Dump of Wikipedia as document repository
  - Results obtained in less than a sec (even for long queries)
  - Supports field queries (ex: title)
  - Index size should be less than 1/4 of the data size.
  - You have to build your own indexing mechanism
    - i.e. you cannot use Nutch or Lucene to index the Wikipedia data.
- Platform:
  - OS: Preferably Linux
  - Languages: Java/C++/Python

Mini Project evaluation

- The evaluation will be done on 4 parameters:
  - Search time,
  - Search efficiency
  - Indexing time
  - Index Size
- You can use compression techniques
- Explore several ranking functions (tf,tf-idf, normalized tf, normalized idf etc) and
- Create a secondary index if required.

Mini Project Deadlines

- First evaluation: 29th August
  - Indexing time and efficiency will be evaluated.
- Second (Final) evaluation: 7th September
  - Dummy queries will be provided before August 26th
  - All four evaluation parameters will be considered

Major Project (10 Weeks)

- Team project (4 members) – Constrained choice
- Advanced topics
- Well defined project
- Major implementation component
- Three deliverables
  - Scope document Sep/26
  - End-to-end system – MVP Oct/25
  - Complete system - Demo/presentation Video, Code, Report… Nov/14

Follow the course on…

Web: http://moodle.iiit.ac.in

Information-Retrieval-Course-at-IIIT-Hyderabad

Key to Knowledge Kingdom!

- Search engine is a lens through which we see (or don’t see) information
  - hence effects what we learn and decide
- What if a search engine took money to suppress the listings of its competitors?
- What if search engines owned by larger corporations promote their own sites
- Influence on the politics, religion, … all aspects of life

Paid Search!

- Paid search (sponsored links) combines
  - Targeted ads
  - Performance based ads
- “Nothing more valuable than the user at the moment of desire”

Evolution of Search Engines

- Crawling and Indexing
- Topic directories
- Clustering and Classification
- Hyperlink analysis
- Resource discovery and vertical portals
- Semantic Web
- ???

New Players in Search War

- Facebook, Twitter, Amazon, Apple
- These are Web 2.0 technologies and new approach to get the attention of the users and also dig into the revenues of Google
- New possibilities are in Web 3.0 players

Search
Social
CLR Search.com
backpage.com post ads for free
ebay
BEAT you There.com BETA EVERYTHING REAL ESTATE
craigslist
facebook
3DVista
FRONTDOOR POWERED BY HGTV
POINT2HOMES
househunting.ca BETA
THESE TIMES DEMAND THE TIMES.
The New York Tim
HOMES
UNIQUE GLOBAL ESTATES
HOMESZ BETA Your Real Estate Network
Vast
hotpads.com The place to find your place
LiveDeal
Home
INGUAL RKETING ROUP
Google Base BETA
North Texas House Hunters
reamf
ldWide Real Estate For Sale
MEXICO BUYER'S GUIDE
Y
trulia BETA real estate search
cyberhomes by Fidelity National Financial
ousing Pages
the search
e

Observations

- Very exciting area
- Very dynamic
- We have seen only the tip of the iceberg
  - Future is even more exciting

thank you

Vasudeva Varma

@devvarma

vv@iiit.ac.in www.iiit.ac.in/~vv