CSE 494 / 598
Information Retrieval, Mining and Integration

LYDIA MANIKONDA

HTTP://WWW.PUBLIC.ASU.EDU/~LMANIKON/
Today

• Field of Information Retrieval
• Course Logistics
• Some IR applications – Overview & Big themes
• Questionnaire
Food for thought

• Most used web search engine  Google

• Second most used search engine  Bing

• PageRank algorithm is used in this search engine  Google

• Yahoo stands for  Yet Another Hierarchically Organized Oracle

• Average length of a search query  2.4 words
Course Outcomes

- After this course, you should be able to answer:
  - How search engines work and why are some better than others
  - Can web be seen as a collection of (semi)structured data/knowledge bases?
  - How can we mine useful patterns from the pages/data of the web?
  - How can we exploit the connectedness of the web pages?
Main Topics

Approximately three halves plus a bit:

◦ Information retrieval
◦ Social Networks
◦ Information integration/Aggregation
◦ Information mining
◦ other topics as permitted by time
Books (or lack there of)

There are *no* required text books

- Primary source is a set of readings that I will provide (see “readings” button in the homepage)
  - *Relative importance of readings is signified by their level of indentation*

A good companion book for the IR topics

- *Intro to Information Retrieval* by Manning/Raghavan/Schutze (available online)
- Modern Information Retrieval (Baeza-Yates et. Al)

Other references

- Modeling the Internet and the Web by Baldi, Frasconi and Smyth
- Mining the web (Soumen Chakrabarti)
- Data on the web (Abiteboul et al).
- A Semantic Web Primer (Antonieu & van Haarmalen)
Pre-reqs

- Useful course background
  - CSE 310 Data structures
  - (Also 4xx course on Algorithms)
  - CSE 412 Databases
  - CSE 471 Intro to AI

- + some of that math you thought you would never use..
  - MAT 342 Linear Algebra
    - Matrices; Eigen values; Eigen Vectors; Singular value decomp
      - Useful for information retrieval and link analysis (pagerank/Authorities-hubs)
  - ECE 389 Probability and Statistics for Engg. Prob solving
    - Discrete probabilities; Bayes rule, long tail, power laws etc.
      - Useful for datamining stuff (e.g. naïve bayes classifier)
Topics Covered

• Introduction & themes
• Information Retrieval
• Indexing & Tolerant Dictionaries
• Correlation analysis and latent semantic indexing
• Link analysis & IR on web
• Social Network Analysis
• Crawling & Map Reduce

• Clustering
• Text classification
• Recommender Systems
• Specifying and Exploiting web
• Information Extraction
• Information Integration
What this course is not (intended to be)

[] there is a difference between training and education. If computer science is a fundamental discipline, then university education in this field should emphasize enduring fundamental principles rather than transient current technology.


• This course is not intended to
  • Teach you how to be a web master
  • Expose you to all the latest x-buzzwords in technology
    • XML/XSL/XPOINTER/XPATH/AJAX
      • (okay, may be a little).
  • Teach you web/javascript/java/jdbc etc. programming
Neither is this course allowed to teach you how to really make money on the web.
Grading etc.

- Projects / Homeworks (~45%)
- Midterm / final (~35%)
- Surprise Quizzes / Participation (~20%)
  - 1 class presentation
  - Reading (papers, web - no single text)
  - Class interaction (***VERY VERY IMPORTANT***)
    - will be evaluated by attendance, attentiveness, and occasional quizzes

471 and 598 students are treated as separate clusters while awarding final letter grades
(no other differentiation)
Course Logistics

• Piazza
  • http://piazza.com/asu/spring2016/cse494598

• Blackboard

• Course Website
  • http://www.public.asu.edu/~Imanikon/CSE494-598/

• Contact Info
  • Instructor: Lydia Manikonda
    • Email: Imanikon@asu.edu
  • TA: Vamsi Meduri
    • Email: vmeduri@asu.edu

*All emails must include "CSE494:IR“ or “CSE598:IR” in the subject title

• Questionnaire
Project (tentative)

One project with 3 parts
- Extending and experimenting with a mini-search engine
  - Project description available online (tentative)
    - *if you did search engine implementations already and would rather do something else, talk to me*

Expected background
- Competence in JAVA programming
  - (Gosling level is fine; Fledgling level probably not..).
  - *We will not be* teaching you JAVA
  - We don’t have TA resources to help with debugging your code.
Sociological issues

• Attendance in the class is *very* important
  • 2 surprise quizzes

• Active concentration in the class is *very* important
  • Not the place for catching up on Sleep/State-press reading

• Interaction/interactiveness is highly encouraged both in and outside the class
  • Offline discussions on Piazza
Academic Integrity

• Ethical violations include cheating on exams, plagiarism, reuse of assignments, improper use of the internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition.

• Academic honesty is required in all work you submit to be graded. You must not look at any other solutions (including program code) to your homework problems or similar problems.

• Falsifying program output of results is prohibited.

• Students who cheat will suffer a serious course grade penalty in addition to being reported to university officials.

• For more details: https://provost.asu.edu/academicintegrity
  • Graduate: https://graduate.asu.edu/academic-integrity
Honor Code/Trawling the Web

Almost any question I can ask you is probably answered somewhere on the web!
- May even be on my own website
- Even if I disable access, Google caches!

...You are still required to do all course related work (homework, exams, projects etc) **yourself**
- Trawling the web in search of exact answers considered academic plagiarism
- *If in doubt, please check with the instructor*
Occupational Hazards..

Caveat: Life on the bleeding edge
- 494 midway between 4xx class & 591 seminars
  - It is a “SEMI-STRUCTURED” class.
- No required text book (recommended books, papers)
- Need a sense of adventure
  - ..and you are assumed to have it, considering that you signed up voluntarily

Being offered for the 8th time.. and it seems to change every time..
- I modify slides until the last minute...
  - To avoid falling asleep during lecture...

Silver Lining?
--Audio & Video Recordings online
Life with a homepage..

I will not be giving any handouts

◦ All class related material will be accessible from piazza or web page
  ◦ Home works may be specified incrementally
    ◦ (one problem at a time)
  ◦ The slides used in the lecture will be available on the class page
    ◦ Due to the success of this course in the past years, the slides/other material will be based on the ones Prof. Subbarao Kambhampati used in Spring 2015 (these are available on the homepage)
      ◦ However I reserve the right to modify them until the last minute (and sometimes beyond it).
    ◦ When printing slides avoid printing the hidden slides
Information Retrieval
Web as a collection of information

• Web viewed as a large collection of__________
  • Text, Structured Data, Semi-structured data
  • (connected) (dynamically changing) (user generated) content
  • (multi-media/Updates/Transactions etc. ignored for now)

• So what do we want to do with it?
  • Search, directed browsing, aggregation, integration, pattern finding

• How do we do it?
  • Depends on your model (text/Structured/semi-structured)
Structure

A generic web page containing text
[English]

An employee record
[SQL]

A movie review
[XML]

How will search and querying on these three types of data differ?
Structure helps querying

• Expressive queries
  • Give me all pages that have key words “Get Rich Quick”
  • Give me the social security numbers of all the employees who have stayed with the company for more than 5 years, and whose yearly salaries are three standard deviations away from the average salary
  • Give me all mails from people from ASU written this year, which are relevant to “get rich quick”
How to get Structure?

When the underlying data is already structured, do unwrapping
- Web *already* has a lot of structured data!
- Invisible web...that disguises itself

..else *extract* structure
- Go from text to structured data (using quasi NLP techniques)

..or *annotate* metadata to add structure
- Semantic web idea..

Structure is so important that we are willing to pay people to add structure or hope that people will be disciplined enough to annotate their pages with structure.
Adapting old disciplines for Web-age

Information (text) retrieval
- Scale of the web
- Hyper text/Link structure
- Authority/hub computations

Social Network Analysis
- Ease of tracking/centrally representing social networks

Databases
- Multiple databases
  - Heterogeneous, access limited, partially overlapping
  - Network (un)reliability

Datamining [Machine Learning/Statistics/Databases]
- Learning patterns from large scale data
Information Retrieval

Traditional Model
- Given
  - a set of documents
  - A query expressed as a set of keywords
- Return
  - A ranked set of documents most relevant to the query
- Evaluation:
  - Precision: Fraction of returned documents that are relevant
  - Recall: Fraction of relevant documents that are returned
  - Efficiency

Web-induced headaches
- Scale (billions of documents)
- Hypertext (inter-document connections)
- Bozo users
- Decentralization (lack of quality guarantees)
  - Hard for users to figure out quality

& simplifications
- Easier to please “lay” users

Consequently
- Ranking that takes link structure into account
  - Authority/Hub
  - Emphasis of precision over recall
- Focus on trustworthiness in addition to relevance
- Indexing and Retrieval algorithms that are ultra fast
Social Networks

Traditional Model

- Given
  - a set of entities (humans)
  - And their relations (network)
- Return
  - Measures of centrality and importance
  - Propagation of trust (Paths through networks)
- Many uses
  - Spread of diseases
  - Spread of rumours
  - Popularity of people
  - Friends circle of people

Web-induced headaches

- Scale (billions of entities)
- Implicit vs. Explicit links
  - Hypertext (inter-entity connections easier to track)
  - Interest-based links

& Simplifications

- Global view of social network possible...

Consequently

- Ranking that takes link structure into account
  - Authority/Hub
- Recommendations (collaborative filtering; trust propagation)
Information Integration
Database Style Retrieval

• Traditional Model (relational)
  • Given:
    • A single relational database
      • Schema
    • Instances
    • A relational (sql) query
  • Return:
    • All tuples satisfying the query

• Evaluation
  • Soundness/Completeness
  • efficiency

• Web-induced headaches
  • Many databases
    • With differing Schemas
  • all are partially complete
  • overlapping
  • heterogeneous schemas
  • access limitations
  • Network (un)reliability

• Consequently
  • Newer models of DB
  • Newer notions of completeness
  • Newer approaches for query planning
Learning Patterns (Web/DB mining)

• Traditional classification learning (supervised)
  • Given
    • a set of structured instances of a pattern (concept)
  • Induce the description of the pattern

• Evaluation:
  • Accuracy of classification on the test data
  • (efficiency of learning)

• Mining headaches
  • Training data is not obvious
  • Training data is massive
  • Training instances are noisy and incomplete

• Consequently
  • Primary emphasis on fast classification
    • Even at the expense of accuracy
  • 80% of the work is “data cleaning”
Future of the Internet

• Domination of Mobile Devices (cellphone, etc)
• Link-Spamming (Arms race to bias SE ranking)
• Local Search, Digital Earth
• Image & Video search
• Social news (Digg / Twitter)
• Crowd Sourcing
• What else?
Big Ideas and Cross Cutting Themes
Finding “Sweet Spots” in computer-mediated cooperative work

It is possible to get by with techniques blythely ignorant of semantics, when you have humans in the loop

- All you need is to find the right sweet spot, where the computer plays a pre-processing role and presents “potential solutions”
- ...and the human very gratefully does the in-depth analysis on those few potential solutions

Examples:
- The incredible success of “Bag of Words” model!
- Bag of letters would be a disaster ;-)
- Bag of sentences and/or NLP would be good
  - ..but only to your discriminating and irascible searchers ;-)
Collaborative Computing
AKA Brain Cycle Stealing
AKA Computizing Eyeballs

A lot of exciting research related to web currently involves “co-opting” the masses to help with large-scale tasks:
- It is like “cycle stealing”—except we are stealing “human brain cycles” (the most idle of the computers if there is ever one ;-)
- Remember the mice in the Hitch Hikers Guide to the Galaxy? (..who were running a mass-scale experiment on the humans to figure out the question..)
- Collaborative knowledge compilation (wikipedia!)
- Collaborative Curation
- Collaborative tagging
- Paid collaboration/contracting

Many big open issues:
- How do you pose the problem such that it can be solved using collaborative computing?
- How do you “incentivize” people into letting you steal their brain cycles?
  - Pay them! (Amazon mturk.com )
  - Make it fun (ESP game)
Another thread of exciting research is driven by the realization that WEB is not random at all!

- It is written by humans
- ...so analyzing its structure and content allows us to tap into the collective unconscious ..
  - Meaning can emerge from syntactic notions such as “co-occurrences” and “connectedness”

Examples:
- Analyzing term co-occurrences in the web-scale corpora to capture semantic information (gmail)
  - Statistical machine translation with massive corpora
- Analyzing the link-structure of the web graph to discover communities
  - DoD and NSA are very much into this as a way of breaking terrorist cells
- Analyzing the transaction patterns of customers (collaborative filtering)
Water Flows Downhill: Lesson Plan, Activity, or Teaching Idea...

Children will experiment with different containers to see if water flows up or down.


Cached - Similar pages

BBC NEWS | Science/Nature | Water 'flows' on Mars

Water flows downhill, Nasa. Briny water flows downhill. She speculates

How does Dyson make water go uphill?

James Dyson's uphill water feature has been the striking image of this year's Chelsea Flower Show. But how did he do it?

http://news.bbc.co.uk/2/hi/uk_news/3046791.stm - 35k - Cached - Similar pages

bit-tech.net | Water flows uphill without pump

Scientists have made water roll uphill, leading to the possibility of an entirely pump-less watercooling system. Could watercooling soon replace...
Water's getting aggressive
It’s a Jungle out there (adversarial Web & Arms Race)

- Web is authority-free zone!
  - Anyone can put up any information and get indexed..
  - Everyone is trying to trip you up... (snopes.com)

- Need to keep “adversarial” aspect constantly in view
  - Adversarial IR (focus on Trust in addition to Relevance)
  - Adversarial mining (the class is being changed even as you are learning)
    - Classic example: Spam mail
"water flows downhill" search results:

A to Z Teacher Stuff :: Water Flows Downhill
Water Flows Downhill Grade Level(s): Preschool, K By: Emily, Kindergarten Teacher. Children will experiment with different containers to see if water flows...
atozteacherstuff.com/pages/515.shtml - Cached

Water Flows Downhill - YouTube
5 min - Jul 30, 2009 - Uploaded by nol97
http://tcb.us/nol - It is amazing that the average person has to learn over and over again that water flows downhill.
www.youtube.com/watch?v=k30OudevmEk

More videos for "water flows downhill" »

Water flows downhill | Hypercrit
Water flows downhill. Posted on July 5, 2011 by Michael Becker ... I would like to preface what follows with this fact: Water flows downhill. ...
www.hypercrit.net/2011/07/05/water-flows-downhill/ - Cached

Guatemala water flows up hill - YouTube
1 min - Dec 27, 2007 - Uploaded by kimbrough15
this isn't an illusion or fake there actually are real places where water flows uphill. jannunez1 month ago...
www.youtube.com/watch?v=Kpso5dCB-08

Minecraft Water Flows Uphill - YouTube
4 min - Jan 18, 2011 - Uploaded by Scatscycle01
I made the water flow uphill... this could be used to a river going upwards. I will expand on it later.
www.youtube.com/watch?v=8lIlQNTM7aWE

More videos for "water flows uphill" »

Water flows uphill without pump | bit-tech.net
May 1, 2006 – Water flows uphill without pump. Scientists have made water roll uphill, leading to the possibility of an entirely pump-less watercooling ...
www.bit-tech.net/news/2006/05/01/water_rolls_u... - Cached - Similar

BBC NEWS | UK | How does Dyson make water go uphill?
May 21, 2003 – James Dyson's uphill water feature has been the striking image of this year's Chelsea Flower Show. But how did he do it?
news.bbc.co.uk/2/hi/uk/3046791.stm - Cached

lies.com » Water Flows Uphill in British Garden
May 22, 2003 – Water Flows Uphill in British Garden. British inventor and vacuum-cleaner magnate James Dyson has created a really cool illusion as part of...
www.lies.com/wp/..../water-flows-uphill-in-british-garden/ - Cached - Similar
Learnings from this Lecture

• Overview of Information Retrieval

• Big themes

• Next class:
  • Traditional IR, vector space representations