Introduction: Machine Learning in Digital Libraries

Compiled by Cornelia Caragea & Sujatha Das

Credits for slides: Hofmann, Mihalcea, Mobasher, Mooney, Schutze

August 18, 2014

Course Title: Case studies in applying Machine Learning for Document Analysis and Retrieval Tasks in Scientific Digital Libraries

Quick Survey

1. Background
   - CS/non-CS
   - Industry vs. graduate student
   - Coding experience
   - Prior course on IR?
   - Prior course in ML?
2. Expectations from RuSSIR
3. Expectation from this course

Machine Learning: Basics

What is “Learning"?

We “learn" many things:

- Motor skills: walk, ride a bicycle, drive, play tennis or golf, play the piano.
- Visual concepts: man-made objects, faces, natural objects.
- Language: Speech recognition, read and write natural languages
- Spatial knowledge: Navigate between spatial locations, physical layout of a room.
- Symbolic knowledge: algebra, arithmetic, calculus.
- Social rules: how to interact with people, animals, machines....

Abstract definition of “Learning"

Definition due to Herbert Simon (1980):

“Learning" denotes changes in a system that are adaptive in that they enable the system to perform the same task or similar tasks drawn from the same population better over time.

Well-posed learning problem

Definition due to Tom Mitchell (1998):

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Spam Filtering

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

1. Classifying emails as spam or not spam.
2. Watching you label emails as spam or not spam.
3. The number (or fraction) of emails correctly classified as spam/not spam.
4. None of the above - this is not a machine learning problem.

Fields of application

- **Biology**: Brain, Development, Evolution, Genetics, Neuroscience.
- **Information Theory**: *Coding Theory, Entropy.*
- **Linguistics**: Grammars, Language acquisition
- **Mathematics**: *Calculus, Linear Algebra, Optimization.*
- **Psychology**: Analogy, Concept Learning, Curiosity, Discovery, Memory, Reinforcement
- **Philosophy**: Causality, Induction, Theory Formation
- **Statistics**: *Probability Distributions, Estimation, Hypothesis Testing.*

Some applications of ML in practice

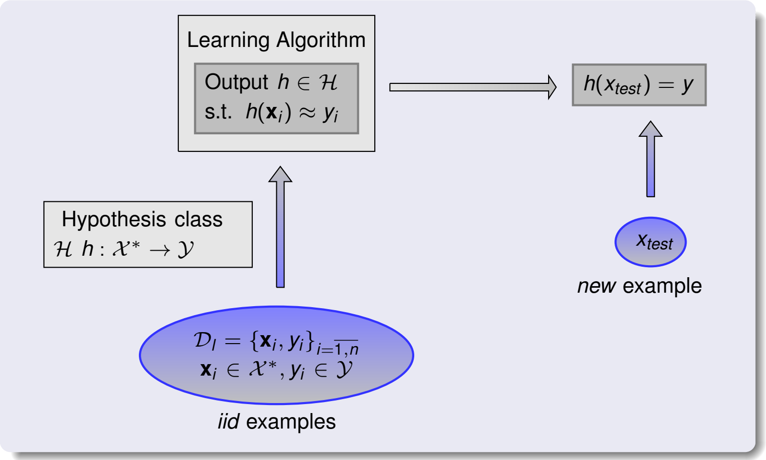
- “If you invent a breakthrough in artificial intelligence, so machines can learn, that is worth 10 Microsofts", Bill Gates quoted in NY Times, Monday March 3, 2004.

- Information extraction from the web: Google, Microsoft, Yahoo
- Spam filtering
- Speech/handwriting recognition
- Object detection/recognition
- Weather prediction
- Stock market analysis
- Search engines (e.g, Google)
- Ad placement on websites
- Adaptive website design
- Credit-card fraud detection
- Webpage clustering (e.g., Google News)
- Social Network Analysis
- Machine Translation (e.g., Google Translate)
- Recommendation systems (e.g., Netflix, Amazon)
- Predicting a protein’s functions
- Automatic vehicle navigation
- Performance tuning of computer systems
- Predicting good compilation flags for programs
- ... and many more

Three fundamental problems in ML

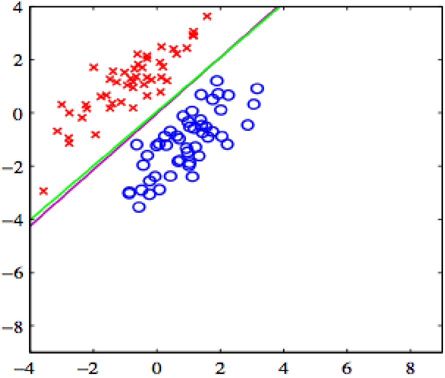
- **Classification:** Learning to predict discrete labels associated with given observations.
  - Binary classification: article related to politics or sports
  - Multiclass classification: digit recognition on postal addresses
- **Regression:** Learning to predict continuous outputs associated with given observations
  - Example: Predict the sales for a particular coffee-mix product
- **Unsupervised learning:** Learning to group objects into categories, without any training labels.
  - Examples: clustering search results into topics

Supervised framework



Learning = Search in Hypothesis Class

Linearly-separable classifiers



- Spam vs. not spam
- Tumor (malignant, benign)

Learning from relevant, labeled examples

- Distinguish a picture of me from a picture of someone else?
  - Provide examples pictures of me and pictures of other people and let a classifier learn to distinguish the two.
- Determine whether a sentence is grammatical or not?
  - Provide examples of grammatical and ungrammatical sentences and let a classifier learn to distinguish the two.
- Distinguish cancerous cells from normal cells?
  - Provide examples of cancerous and normal cells and let a classifier learn to distinguish the two.

Labeled data ("play" prediction)

Example dataset:

Class	**Outlook**	**Temperature**	**Windy?**
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

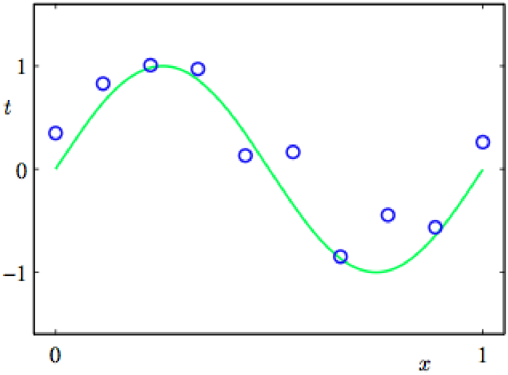
1. Class label (aka "label", denoted y)
2. Features (aka "attributes")
3. Feature values (aka "attribute values", denoted x)

   ⇒ Features can be binary, nomial or continuous

A *labeled* dataset is a collection of (x, y) pairs

Regression

Plot of a training data set of $N = 10$ points, shown as blue circles, each comprising an observation of the input variable x along with the corresponding target variable t. The green curve shows the function $\sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x, without knowledge of the green curve.



Regression in Medicine

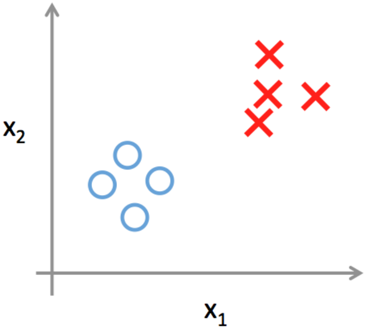
TABLE 1

Diabetes study: 442 diabetes patients were measured on 10 baseline variables; a prediction model was desired for the response variable, a measure of disease progression one year after baseline

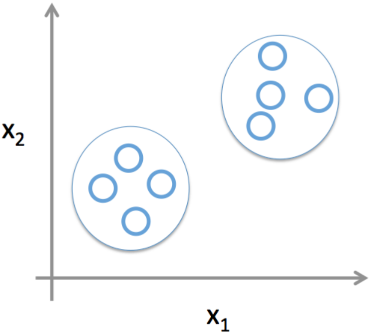
	AGE	SEX	BMI	BP	Serum measurements						Response
Patient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

[Efron et al., Least Angle Regression, Annals of Statistics, 2004]

Unsupervised learning

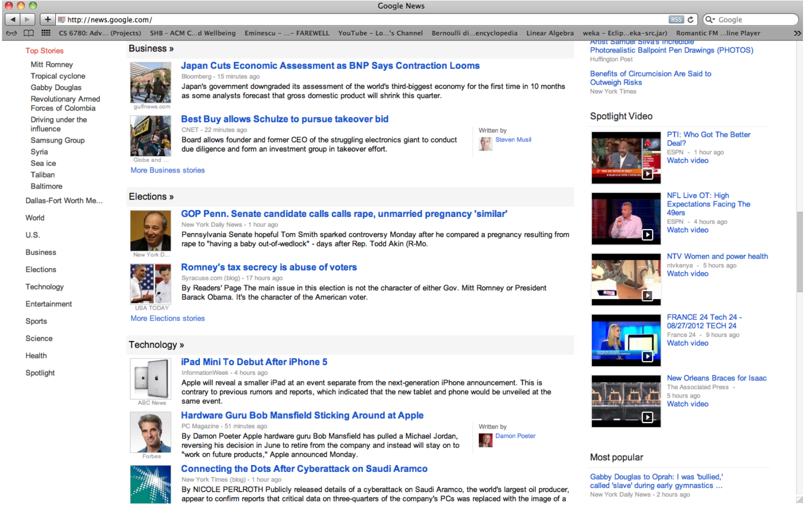


Supervised learning



Unsupervised learning

Clustering “news" articles



References

- *Pattern Recognition and Machine Learning*, Christopher Bishop.
- *Machine Learning*, Tom Mitchell.
- *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Trevor Hastie, Robert Tibshirani, Jerome Friedman.

Information Retrieval Systems: Basics

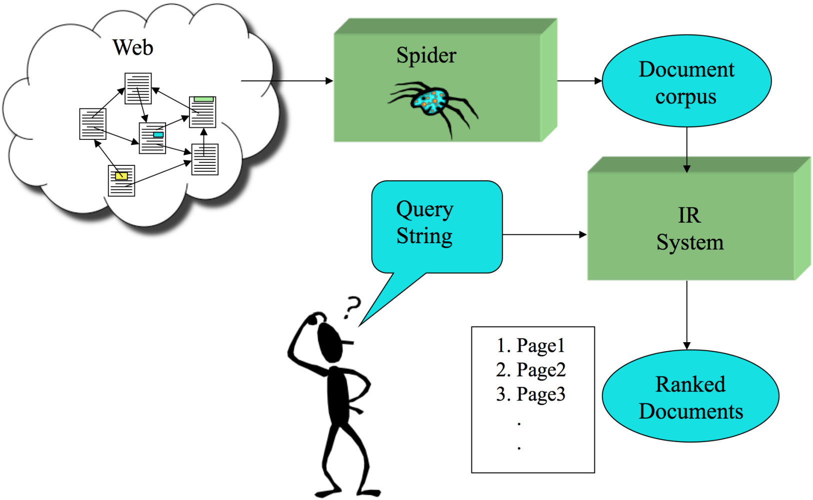
What is Information Retrieval (IR)

- The processing, indexing and retrieval of ~~textual~~ documents.
- 1. retrieving relevant documents to a query.
  2. retrieving from large sets of documents efficiently.

Key terms

- Query: a representation of what the user is looking for - can be a list of words or a phrase.
- Document: webpage/pdf/image...what user wants to retrieve
- Collection or corpus: a set of documents
- Index: a set of data structures that make querying efficient
- Term: word or concept that appears in a document or a query

Typical IR system architecture



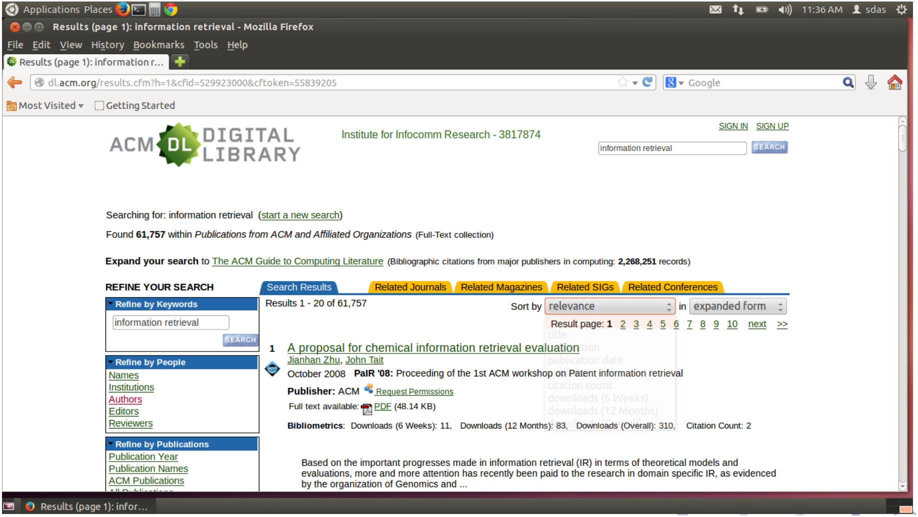
What is a Digital Library?

- An electronic library for a focused collection of digital objects
- Objects can include text, visual material, audio material ...(electronic media formats)
- A type of information retrieval system.
- Examples: CiteSeer^x, PubMed, ACM DL, LawNet ...

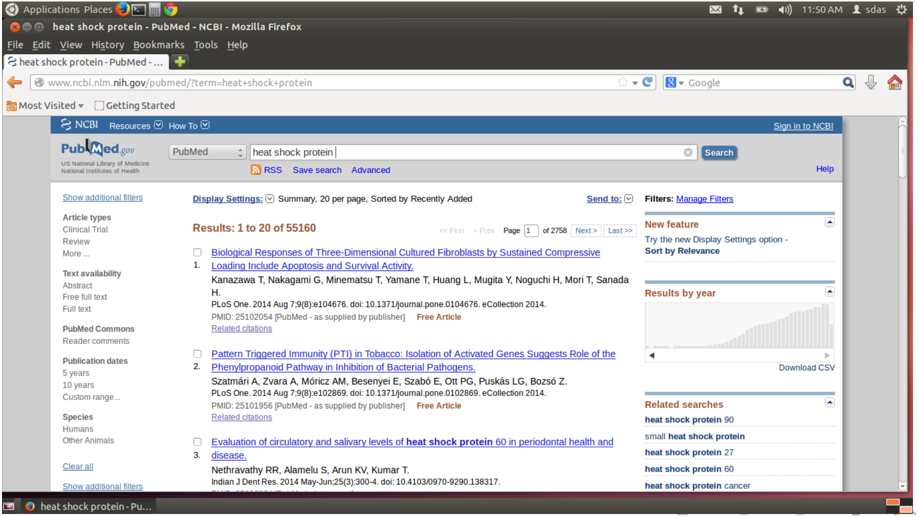
Web Search (Google) vs. Digital Libraries

- "All the Web" vs. domain-specific collections/special types of documents
- Everybody vs. users with "special" needs
- "Documents" vs. "Documents, Authors, Connections..."
- For a DL (or a typical IR system)
  - Must assemble a document corpus (spidering the Web or from trusted sources)
  - Document collections need to be constantly updated
  - Different types of search, ranking, and visualization tasks

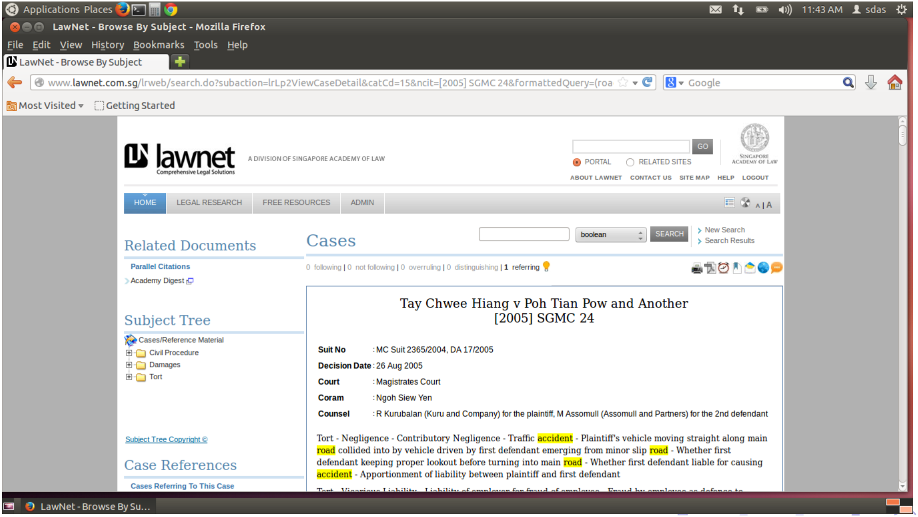
ACM Digital Library



PubMed


Applications Places
11:50 AM sdas
heat shock protein - PubMed - NCBI - Mozilla Firefox
File Edit View History Bookmarks Tools Help
heat shock protein - PubMed - ...
www.ncbi.nlm.nih.gov/pubmed/?term=heat+shock+protein
Google
Most Visited
Getting Started
NCBI Resources How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
heat shock protein
Search
RSS Save search Advanced
Help
Show additional filters
Article types
Clinical Trial
Review
More ...
Text availability
Abstract
Free full text
Full text
PubMed Commons
Reader comments
Publication dates
5 years
10 years
Custom range...
Species
Humans
Other Animals
Clear all
Show additional filters
Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to:
Results: 1 to 20 of 55160
<< First < Prev Page 1 of 2758 Next > Last >>
1. Biological Responses of Three-Dimensional Cultured Fibroblasts by Sustained Compressive Loading Include Apoptosis and Survival Activity.
Kanazawa T, Nakagami G, Minematsu T, Yamane T, Huang L, Mugita Y, Noguchi H, Mori T, Sanada H.
PLoS One. 2014 Aug 7;9(8):e104676. doi: 10.1371/journal.pone.0104676. eCollection 2014.
PMID: 25102054 [PubMed - as supplied by publisher] Free Article
Related citations
2. Pattern Triggered Immunity (PTI) in Tobacco: Isolation of Activated Genes Suggests Role of the Phenylpropanoid Pathway in Inhibition of Bacterial Pathogens.
Szatmári A, Zvara A, Móricz AM, Besenyei E, Szabó E, Ott PG, Puskás LG, Bozsó Z.
PLoS One. 2014 Aug 7;9(8):e102869. doi: 10.1371/journal.pone.0102869. eCollection 2014.
PMID: 25101956 [PubMed - as supplied by publisher] Free Article
Related citations
3. Evaluation of circulatory and salivary levels of heat shock protein 60 in periodontal health and disease.
Nethravathy RR, Alamelu S, Arun KV, Kumar T.
Indian J Dent Res. 2014 May-Jun;25(3):300-4. doi: 10.4103/0970-9290.138317.
Filters: Manage Filters
New feature
Try the new Display Settings option - Sort by Relevance
Results by year
Download CSV
Related searches
heat shock protein 90
small heat shock protein
heat shock protein 27
heat shock protein 60
heat shock protein cancer
heat shock protein - Pu...

LawNet


Applications Places
11:43 AM sdas
LawNet - Browse By Subject - Mozilla Firefox
File Edit View History Bookmarks Tools Help
LawNet - Browse By Subject
www.lawnet.com.sg/lrweb/search.do?subaction=lrLp2ViewCaseDetail&catCd=15&ncit=[2005] SGMC 24&formattedQuery=(roa
Google
Most Visited
Getting Started
lawnet
Comprehensive Legal Solutions
A DIVISION OF SINGAPORE ACADEMY OF LAW
GO
PORTAL
RELATED SITES
ABOUT LAWNET CONTACT US SITE MAP HELP LOGOUT
HOME LEGAL RESEARCH FREE RESOURCES ADMIN
Related Documents
Parallel Citations
Academy Digest
Subject Tree
Cases/Reference Material
Civil Procedure
Damages
Tort
Subject Tree Copyright ©
Case References
Cases Referring To This Case
Cases
boolean
SEARCH
New Search
Search Results
0 following | 0 not following | 0 overruling | 0 distinguishing | 1 referring
Tay Chwee Hiang v Poh Tian Pow and Another
[2005] SGMC 24
Suit No :MC Suit 2365/2004, DA 17/2005
Decision Date : 26 Aug 2005
Court :Magistrates Court
Coram :Ngoh Siew Yen
Counsel :R Kurubalan (Kuru and Company) for the plaintiff, M Assomull (Assomull and Partners) for the 2nd defendant
Tort - Negligence - Contributory Negligence - Traffic accident - Plaintiff's vehicle moving straight along main road collided into by vehicle driven by first defendant emerging from minor slip road - Whether first defendant keeping proper lookout before turning into main road - Whether first defendant liable for causing accident - Apportionment of liability between plaintiff and first defendant
LawNet - Browse By Su...

Components in an IR system

- Crawl/Acquire* documents that need to be indexed in the system.
- Index/Search retrieves documents that contain a given query token from the inverted index.
- Rank scores all retrieved documents according to a relevance metric.
- Visualize manages interaction with the user:

Typical IR Search

- Given:
  - A corpus
  - A user query in the form of a textual string
- Find:
  - A ranked set of documents that are relevant to the query

Relevance and Ranking

- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (information need)
- Main relevance criterion: an IR system should fulfill a user's information need
- Relevance is "hard to measure"
- Measures such as Precision, Recall, Mean Reciprocal Rank, NDCG on benchmark collections (example, from TREC)

Information Retrieval

- The processing, indexing and retrieval of documents.
- 1. retrieving relevant documents to a query.
  2. retrieving from large sets of documents efficiently.
- Matching documents and queries
  - Handling vocabulary mismatch (“PRC" vs “China")
  - Handling ambiguity (“bat", “jaguar")

Implementation and User Experience concerns

- Fast search (efficient data structures such as inverted indices)
- What queries are possible?
- How many results?
- Query suggestions?
- Show similar searches?
- Cluster results, other visualizations?

References

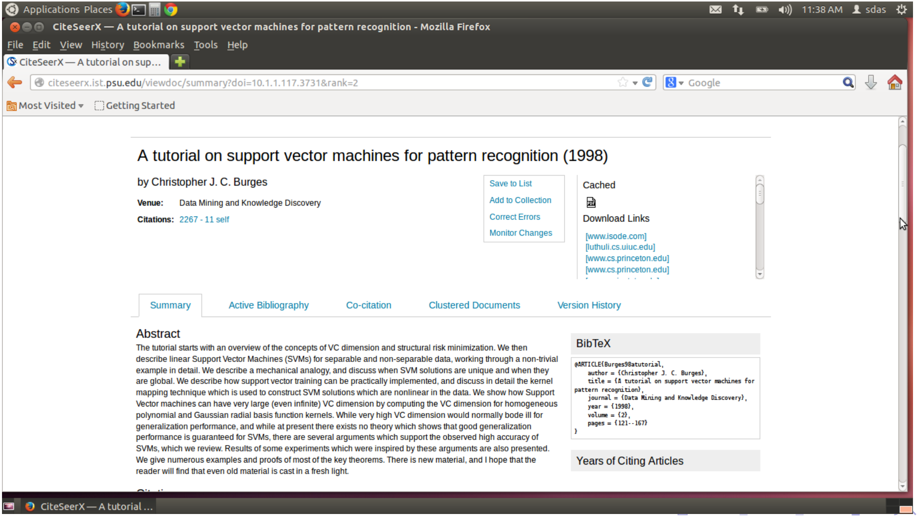
- “Introduction to Information Retrieval", *C.D. Manning, P. Raghavan, H. Schütze*
- “Mining the Web: Discovering Knowledge from Hypertext", *Soumen Chakrabarti*
- “Search Engines: Information Retrieval in Practice", *Bruce Croft, Donald Metzler and Trevor Strohman*

ML in a practical IR system (CiteSeer)

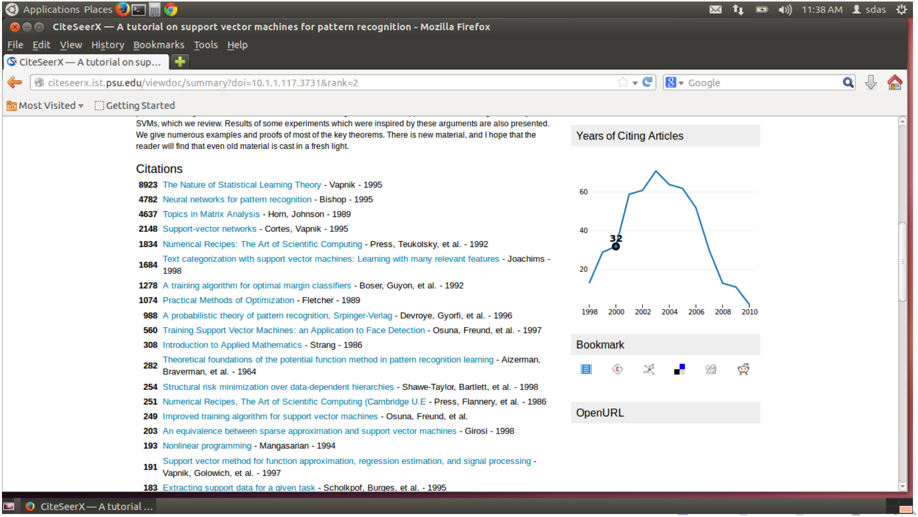
What is CiteSeer/CiteSeer^x?

- Scientific digital library for Computer and Information Science
- Indexes (free) PostScript and PDF research articles on the Web
- Automated techniques for acquiring and harvesting research articles
- Several functionalities: citation indexing, metadata extraction, author disambiguation, citation statistics and trends

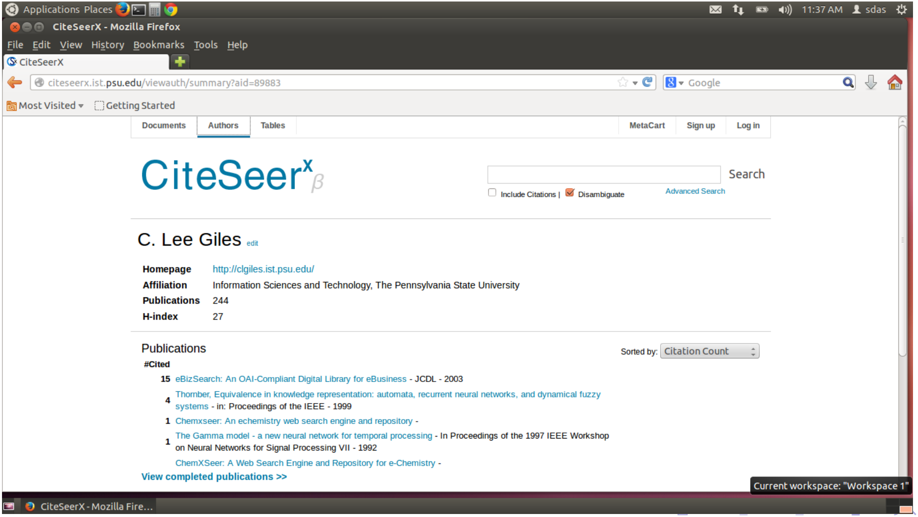
CiteSeer: Document Search and Metadata



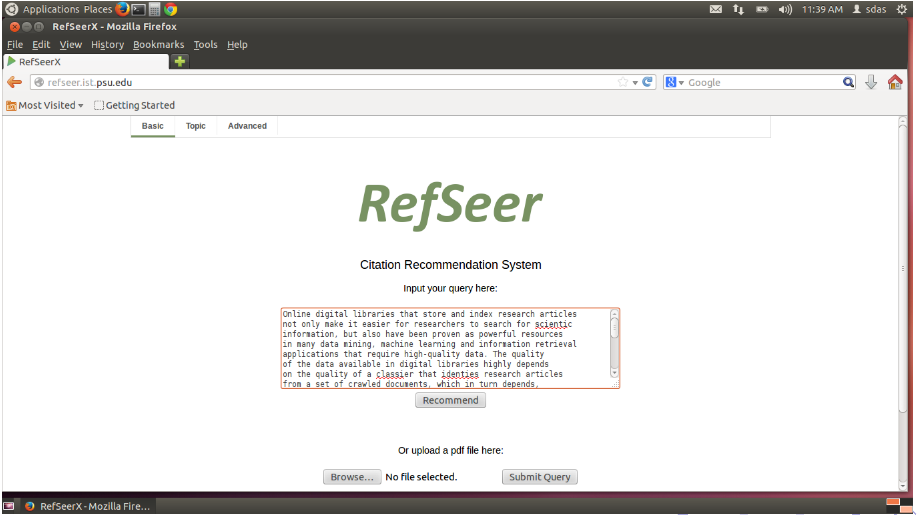
CiteSeer: Citations and Trends



Author Disambiguation



Reference Recommendation



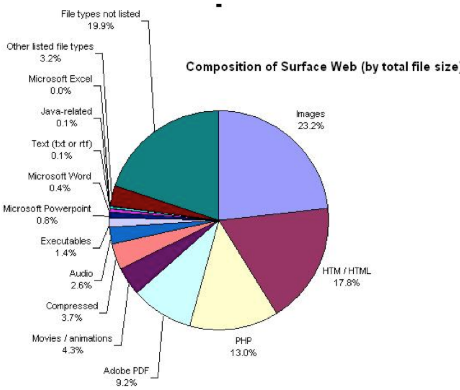
Upcoming Lectures/Hands-on

- How are we applying Machine Learning techniques for these tasks in CiteSeer?
  - Day 1: Introduction + Heritrix (crawling) exercise
  - Day 2: Classification + Weka exercise
  - Day 3: Pagerank/graph-based analysis + Gephi demo
  - Day 4: Topic Modeling + Mallet (LDA) exercise
  - Day 5: Information Extraction + OpenCalais demo
- Requirements for exercises: Familiarity with Java and Linux environments

Crawling the Web

The Web (Corpus) by the Numbers

- 43 million web servers
- 167 Terabytes of data
  - About 20% text/html
- 100 Terabytes in "deep Web"
- 440 Terabytes in emails



[Lyman & Varian: How much Information? 2003]

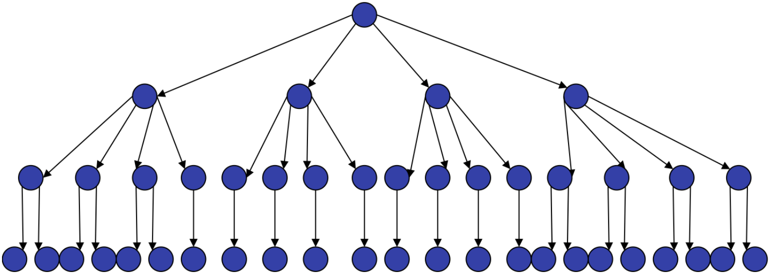
http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Spiders (Robots/Bots/Crawlers)

- **Spidering** represents the main difference between traditional IR and IR these days.
  - Start with a comprehensive set of root URL's from which to start the search.
  - Follow all links on these pages recursively to find additional pages.
  - Index/Process all **novel** found pages in an inverted index as they are encountered.
  - May allow users to directly submit pages to be indexed (and crawled from).

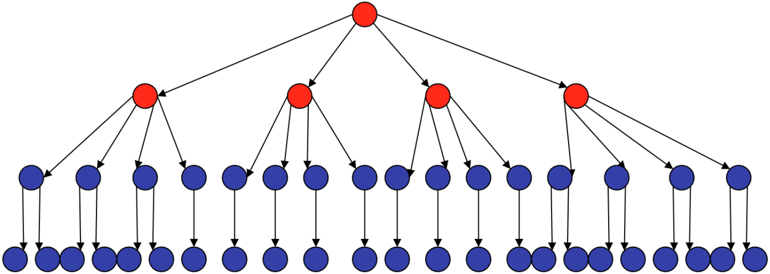
Search Strategies

- Breadth-first Search



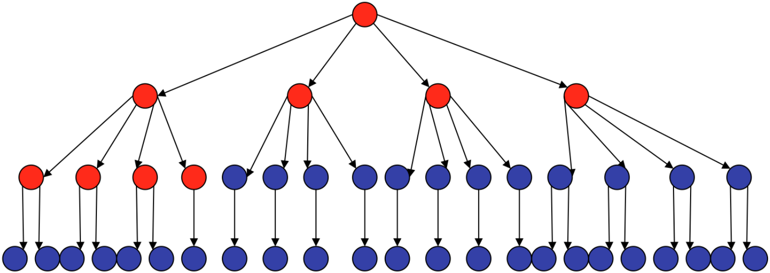
Search Strategies

- Breadth-first Search



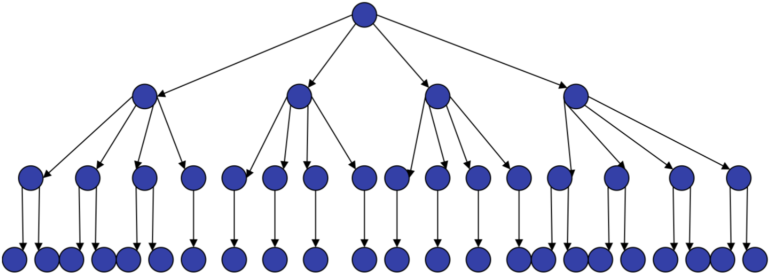
Search Strategies

- Breadth-first Search



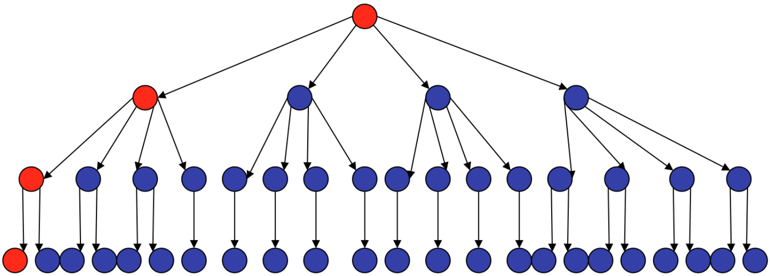
Search Strategies

- Depth-first Search



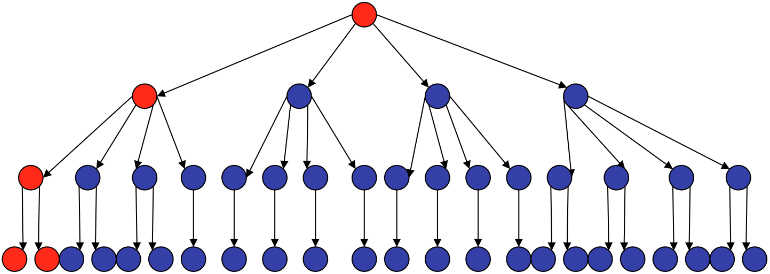
Search Strategies

- Depth-first Search



Search Strategies

- Depth-first Search



Some challenges/concerns

- Must detect when revisiting a page that has already been spidered (web is a graph not a tree, link canonicalization).
- Must efficiently index visited pages to allow rapid recognition test.
- Restricting the crawl (robots.txt, content/anchor-text decide-rules)
- How often should we crawl?
- Directed/Focused Spidering

Hands-on with Heritrix