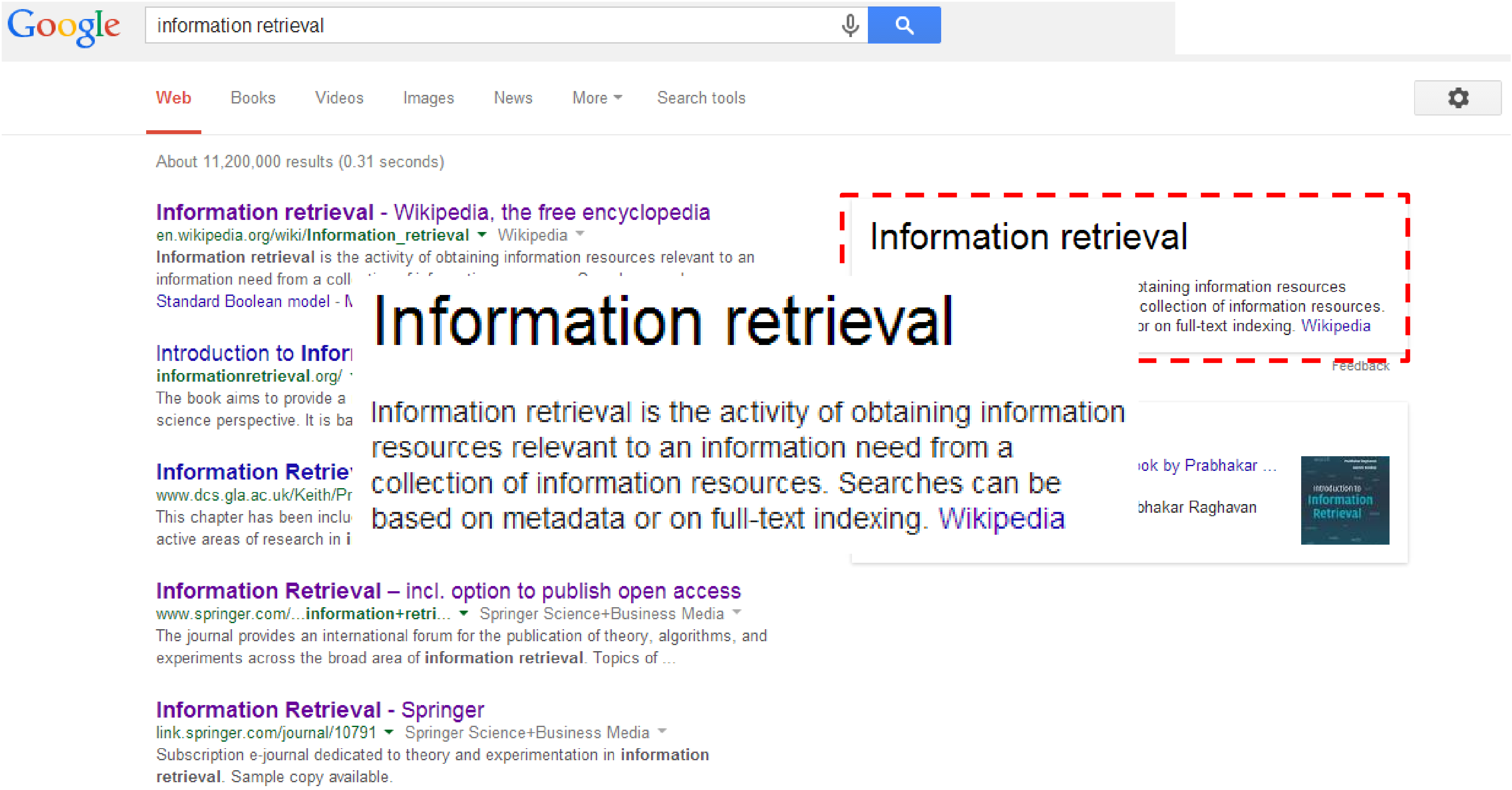
Information Retrieval

Lecture 1 Introduction

What is information retrieval?



About 11,200,000 results (0.31 seconds)

Information retrieval - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Information_retrieval** ▾ Wikipedia ▾
Information retrieval is the activity of obtaining information resources relevant to an information need from a coll
Standard Boolean model - M

Introduction to Infor
informationretrieval.org/
The book aims to provide a
science perspective. It is ba

Information Retrie
www.dcs.gla.ac.uk/Keith/Pr
This chapter has been inclu
active areas of research in i

Information Retrieval – incl. option to publish open access
www.springer.com/...**information+retri**... ▾ Springer Science+Business Media ▾
The journal provides an international forum for the publication of theory, algorithms, and experiments across the broad area of **information retrieval**. Topics of ...

Information Retrieval - Springer
link.springer.com/journal/10791 ▾ Springer Science+Business Media ▾
Subscription e-journal dedicated to theory and experimentation in **information retrieval**. Sample copy available.

Information retrieval

otaining information resources
collection of information resources.
or on full-text indexing. Wikipedia

Feedback

ok by Prabhakar ...
bhakar Raghavan

Information retrieval

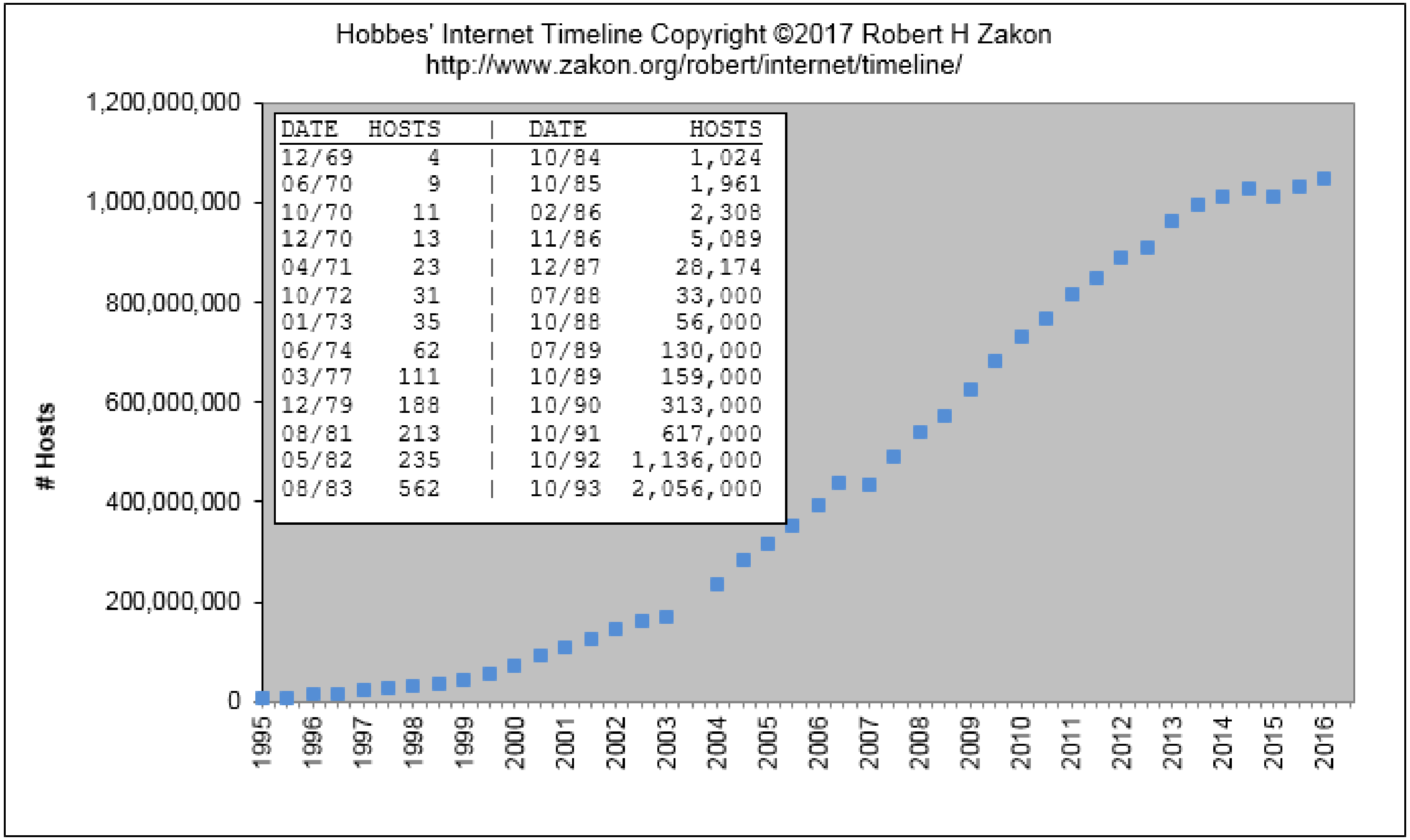
Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Wikipedia

Why information retrieval

- Information overload
  - *"It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information."* - wiki

Why information retrieval

- Information overload



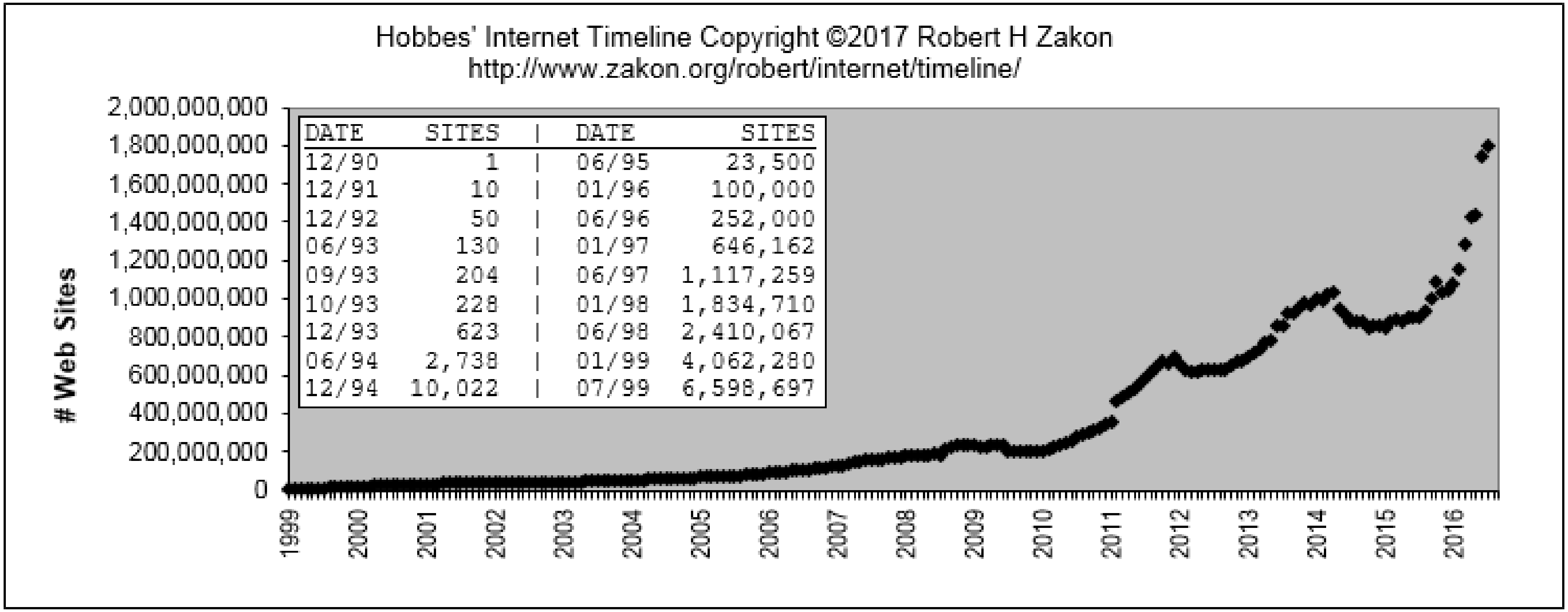
Hobbes' Internet Timeline Copyright ©2017 Robert H Zakon
http://www.zakon.org/robert/internet/timeline/

DATE	HOSTS	DATE	HOSTS
12/69	4	10/84	1,024
06/70	9	10/85	1,961
10/70	11	02/86	2,308
12/70	13	11/86	5,089
04/71	23	12/87	28,174
10/72	31	07/88	33,000
01/73	35	10/88	56,000
06/74	62	07/89	130,000
03/77	111	10/89	159,000
12/79	188	10/90	313,000
08/81	213	10/91	617,000
05/82	235	10/92	1,136,000
08/83	562	10/93	2,056,000

Figure 1: Growth of Internet

Why information retrieval

- Information overload



Hobbes' Internet Timeline Copyright ©2017 Robert H Zakon
http://www.zakon.org/robert/internet/timeline/

DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

Figure 2: Growth of WWW

Why information retrieval

- Handling unstructured data
  - Structured data: database system is a good choice
  - Unst
    - Te lio, video…
    - “ as
      ur
    - U

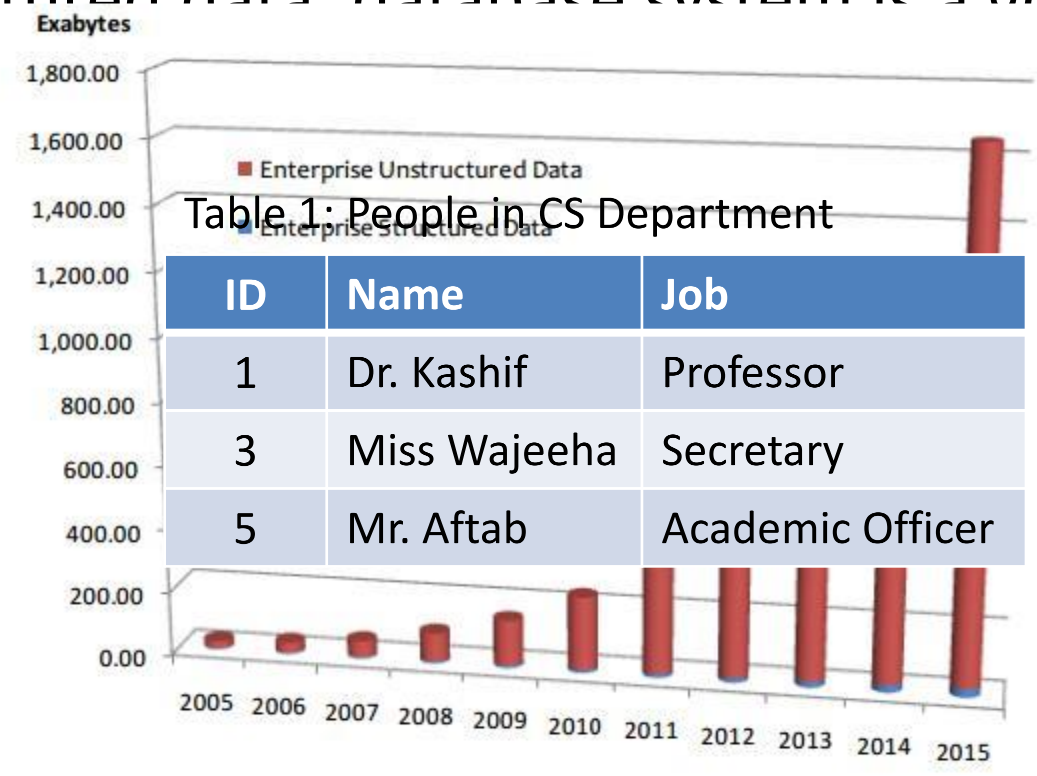
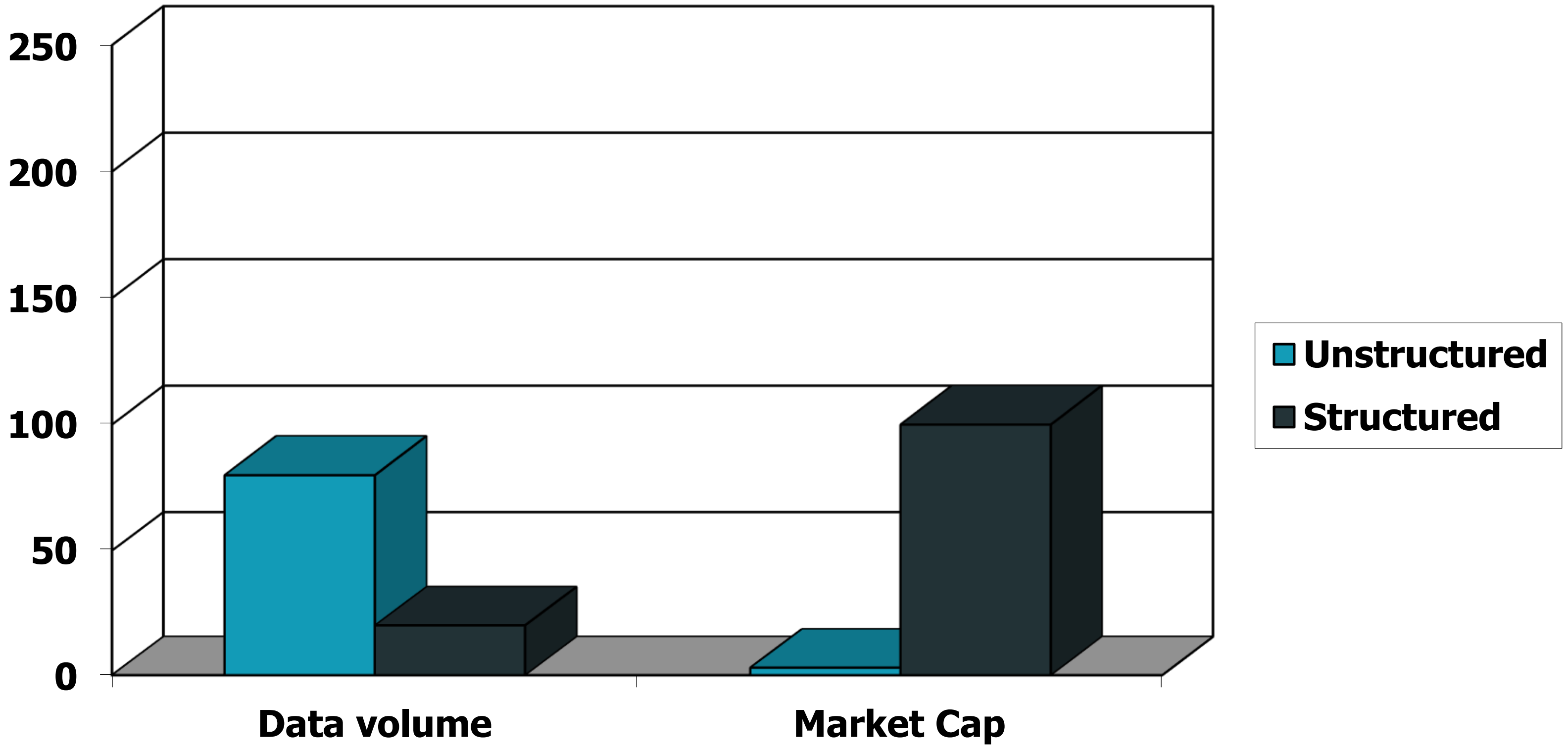


Table 1: People in CS Department

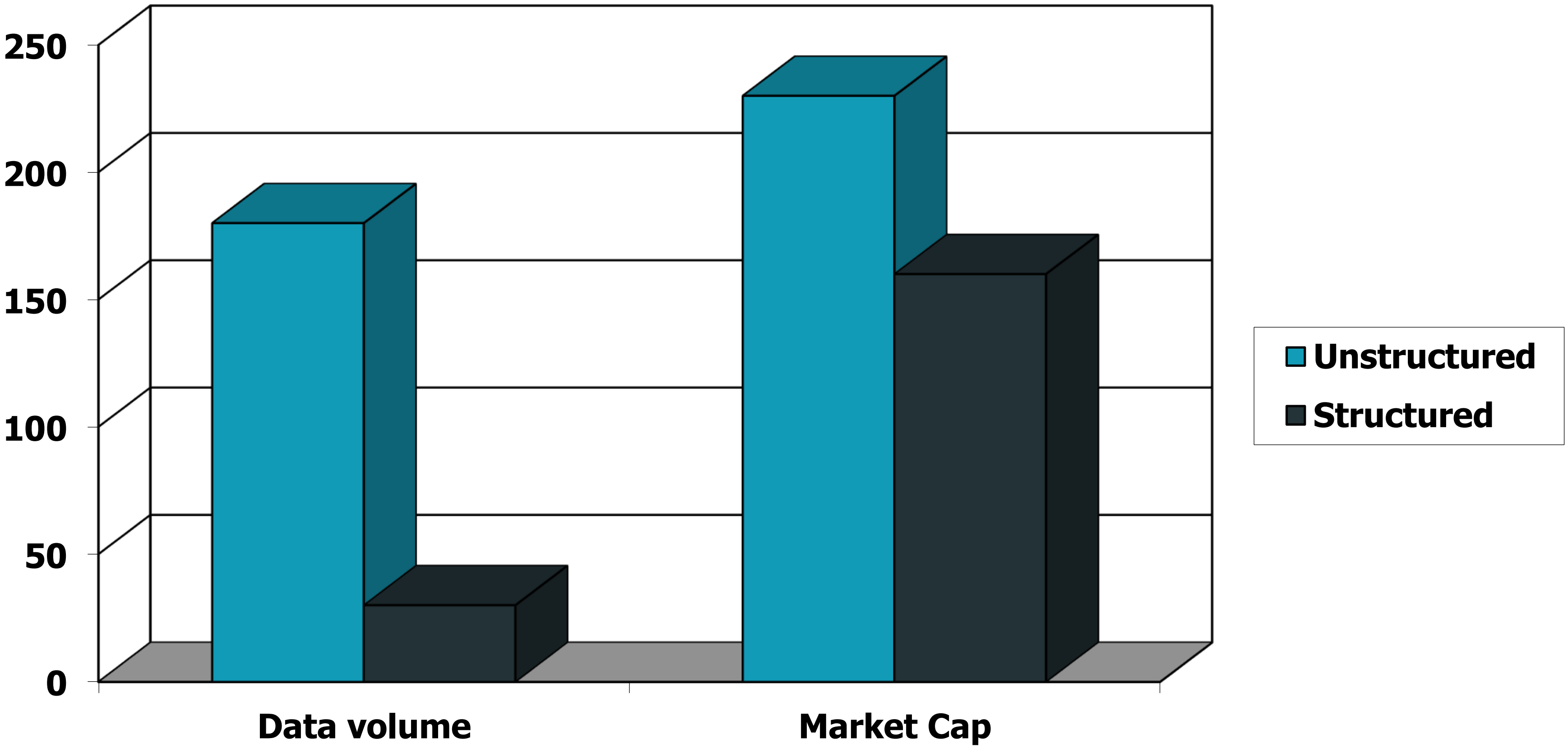
ID	Name	Job
1	Dr. Kashif	Professor
3	Miss Wajeeha	Secretary
5	Mr. Aftab	Academic Officer

Total Enterprise Data Growth 2005-2015, IDC 2012

Unstructured (text) vs. structured (database) data in the mid-nineties



Unstructured (text) vs. structured (database) data today



Why information retrieval

- An essential tool to deal with information overload

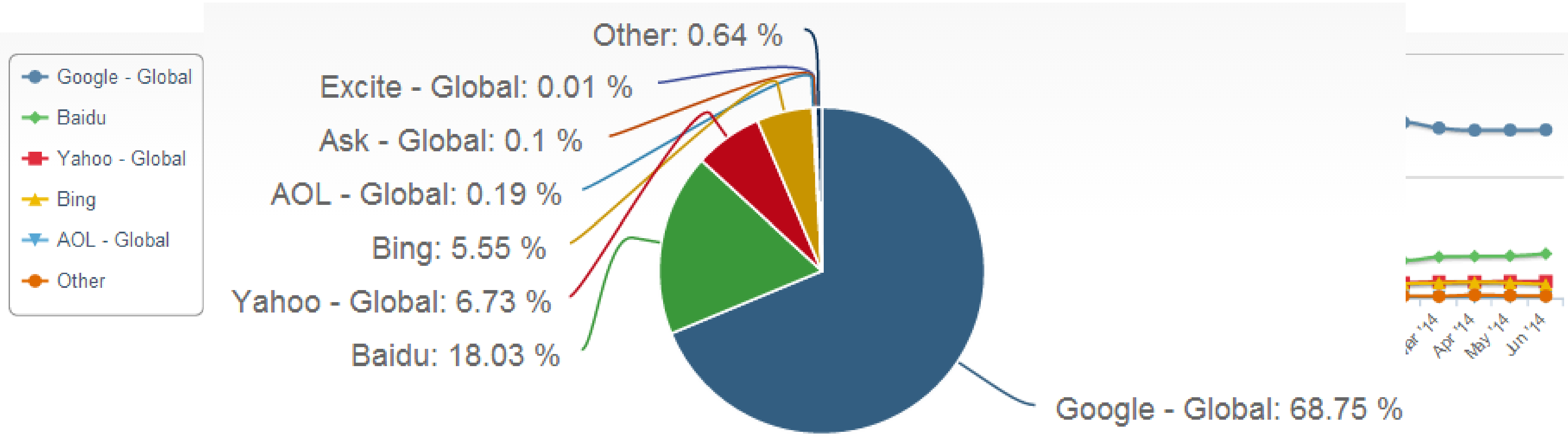
You are here!

History of information retrieval

- Catalyst
  - Industry: web search engines
    - WWW unleashed explosion of published information and drove the innovation of IR techniques
    - Lycos (started at CMU) was launched and became a major commercial endeavor in 1994
    - Booming of search engine industry: *Magellan*, *Excite*, *Infoseek*, *Inktomi*, *Northern Light*, *AltaVista*, *Yahoo!*, *Google*, and *Bing*

Major players in this game

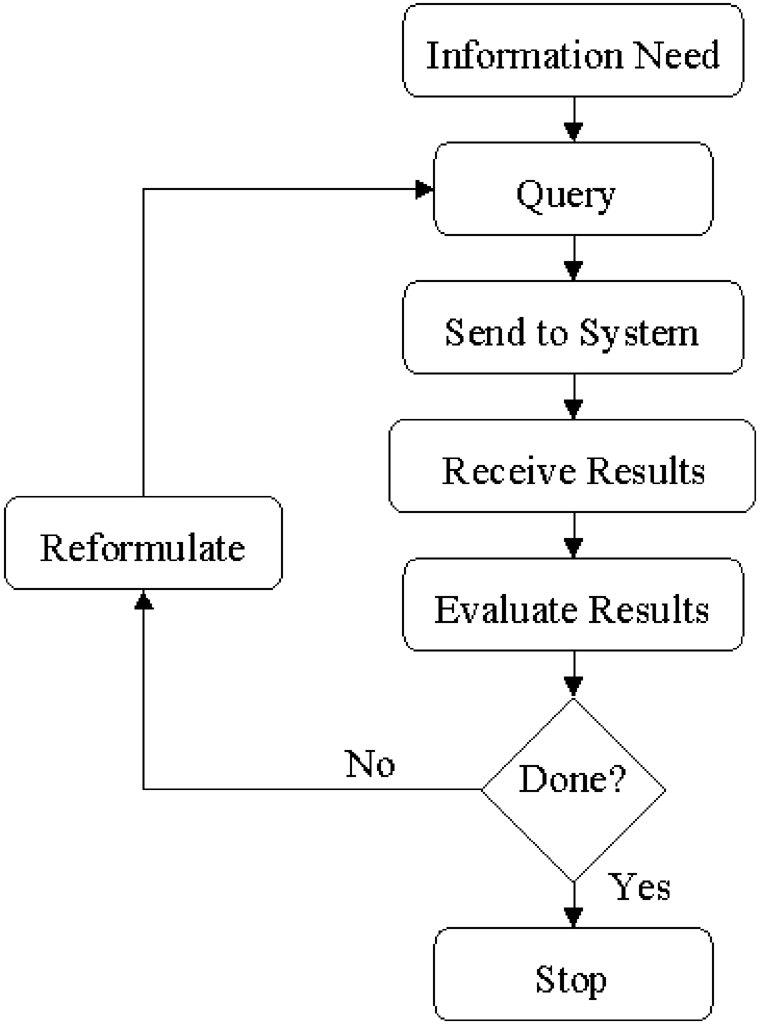
- Global search engine market
  - By http://marketshare.hitslink.com/search-engine-market-share.aspx



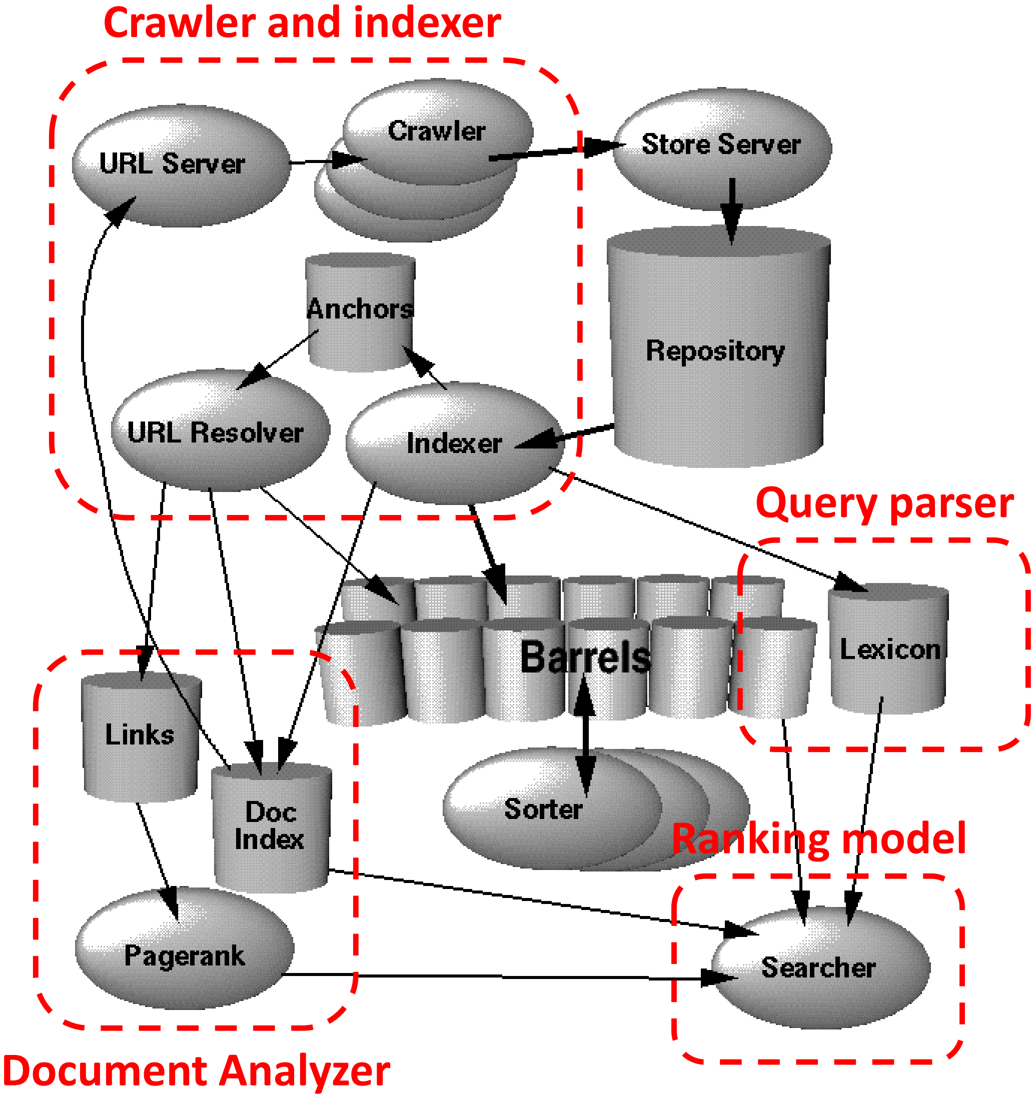
How to perform information retrieval

- Information retrieval when we did not have a computer

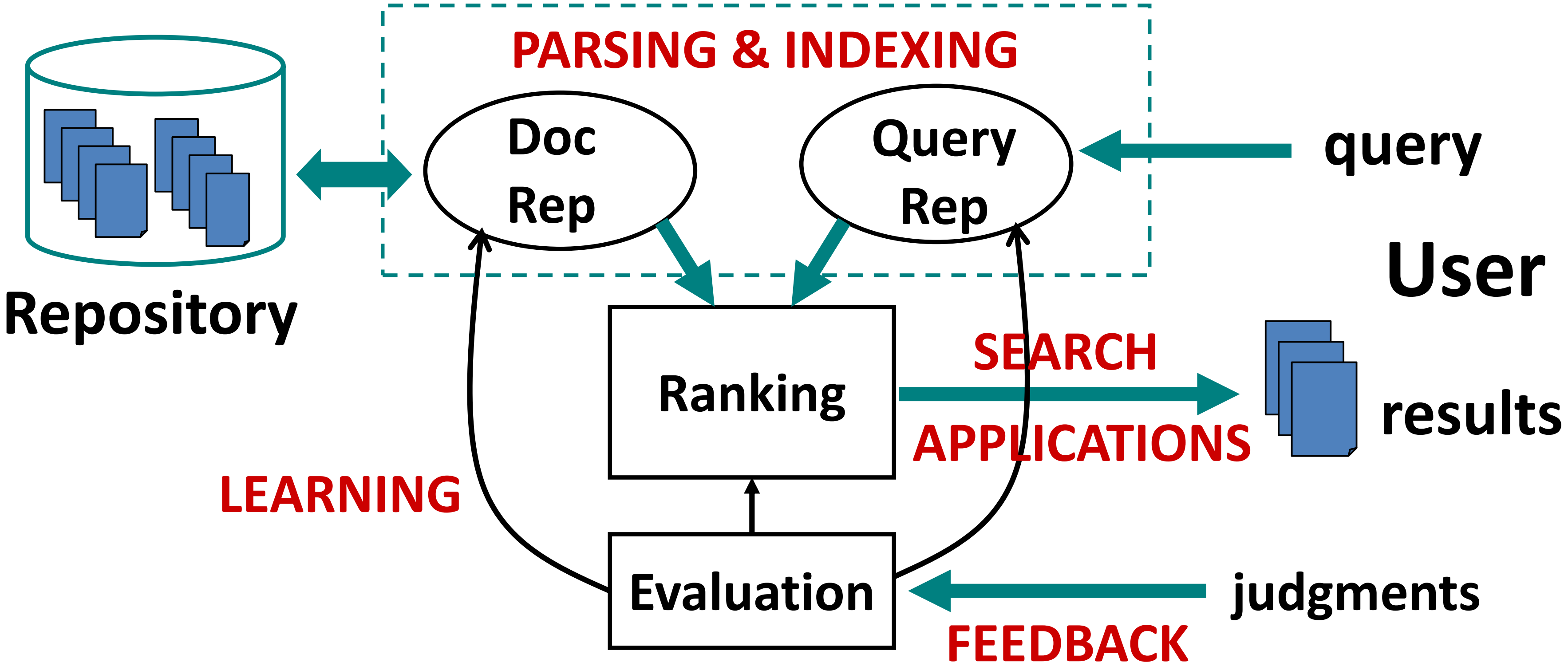
The Standard Retrieval Interaction Model



How to perform information retrieval



How to perform information retrieval



We will cover:

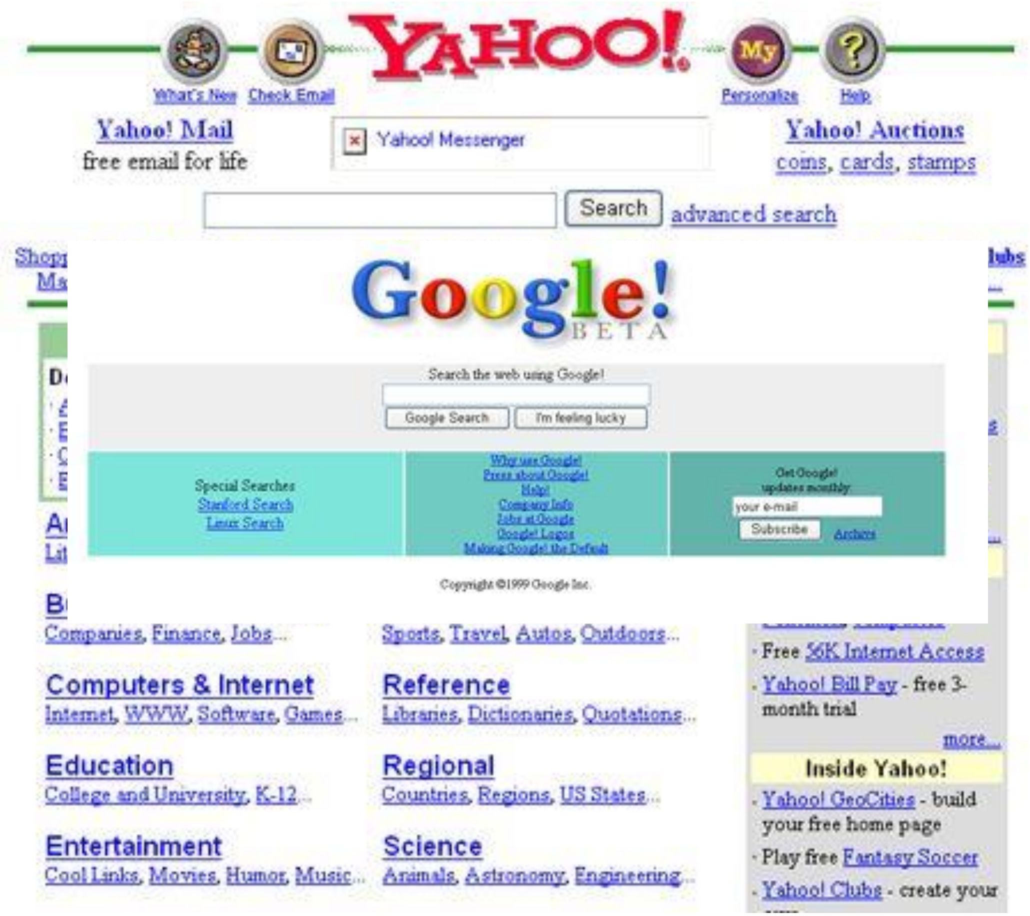
1) Search engine architecture; 2)Retrieval models;
3) Retrieval evaluation; 4) Relevance feedback;
5) Link analysis; 6) Search applications.

Core concepts in IR

- Query representation
  - Lexical gap: say v.s. said
  - Semantic gap
- Document representation
  - Specific data structure for efficient access
- Retrieval model
  - Algorithms that find the ***most relevant*** documents for the given information need

A glance of modern search engine

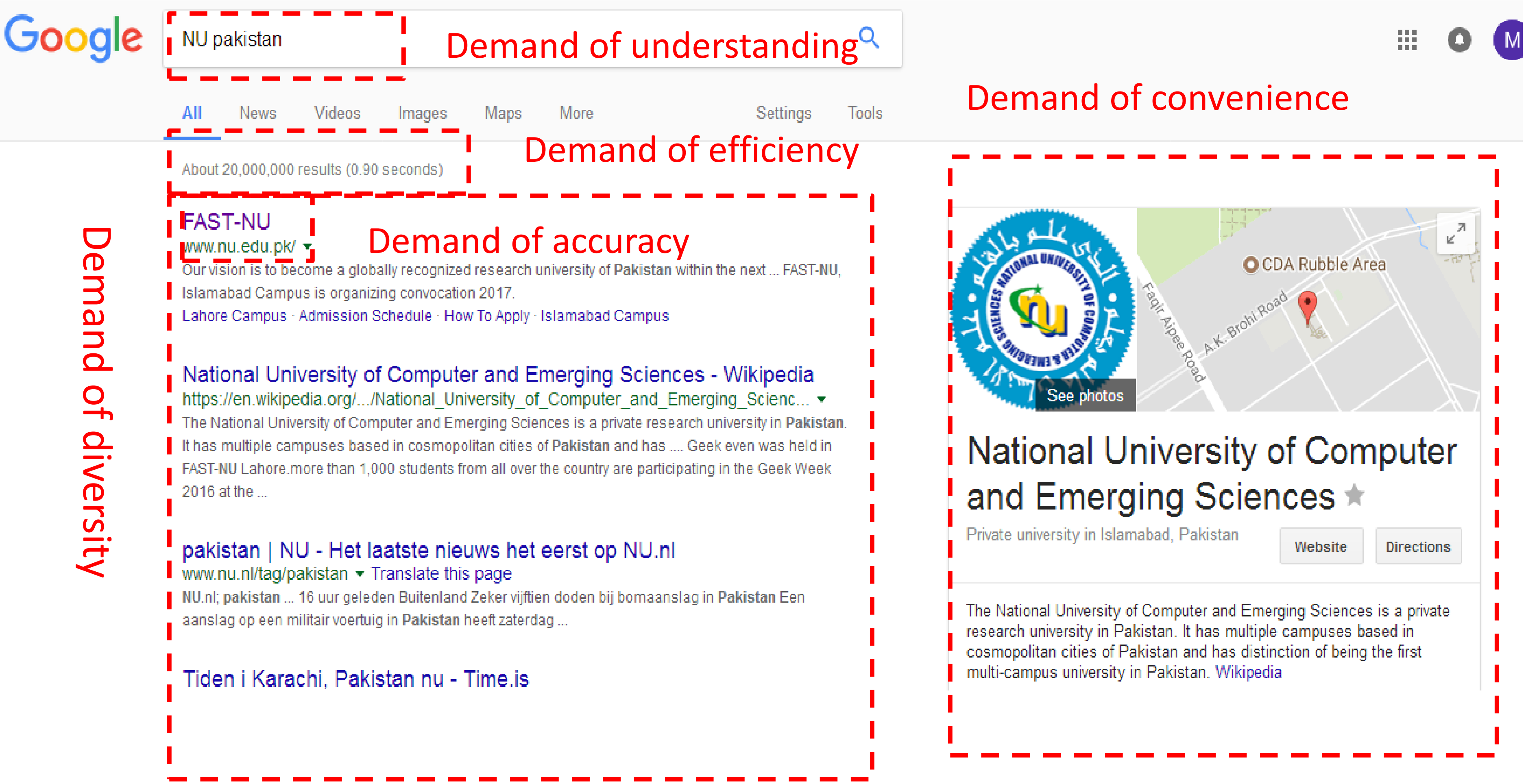
- In old times



A glance of modern search engine

In modern time

A glance of modern search engine



NU pakistan

About 20,000,000 results (0.90 seconds)

FAST-NU
www.nu.edu.pk/
Our vision is to become a globally recognized research university of **Pakistan** within the next ... FAST-**NU**, Islamabad Campus is organizing convocation 2017.
Lahore Campus · Admission Schedule · How To Apply · Islamabad Campus

National University of Computer and Emerging Sciences - Wikipedia
https://en.wikipedia.org/.../National_University_of_Computer_and_Emerging_Scienc...
The National University of Computer and Emerging Sciences is a private research university in **Pakistan**. It has multiple campuses based in cosmopolitan cities of **Pakistan** and has .... Geek even was held in FAST-**NU** Lahore.more than 1,000 students from all over the country are participating in the Geek Week 2016 at the ...

pakistan | NU - Het laatste nieuws het eerst op NU.nl
www.nu.nl/tag/pakistan Translate this page
NU.nl; **pakistan** ... 16 uur geleden Buitenland Zeker vijftien doden bij bomaanslag in **Pakistan** Een aanslag op een militair voertuig in **Pakistan** heeft zaterdag ...

Tiden i Karachi, Pakistan nu - Time.is

National University of Computer and Emerging Sciences ★
Private university in Islamabad, Pakistan

The National University of Computer and Emerging Sciences is a private research university in Pakistan. It has multiple campuses based in cosmopolitan cities of Pakistan and has distinction of being the first multi-campus university in Pakistan. Wikipedia

Demand of understanding

Demand of convenience

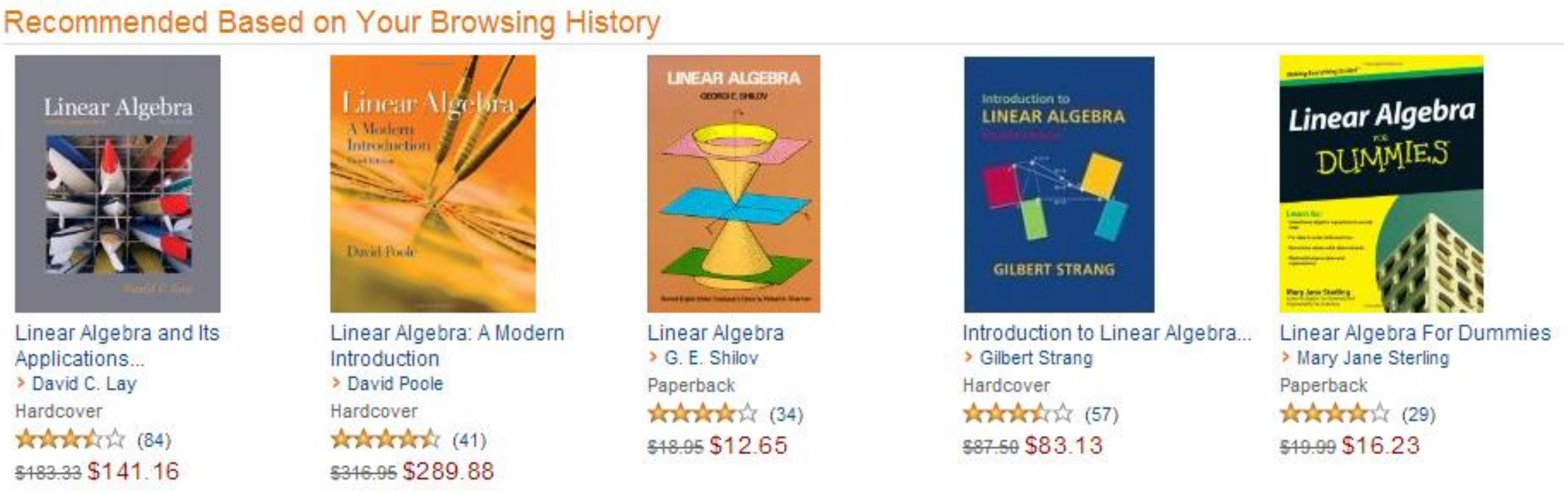
Demand of efficiency

Demand of accuracy

Demand of diversity

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Recommendation



Recommended Based on Your Browsing History

Linear Algebra and Its Applications...
› David C. Lay
Hardcover
★★★½☆ (84)
~~$183.33~~ $141.16

Linear Algebra: A Modern Introduction
› David Poole
Hardcover
★★★★½ (41)
~~$316.95~~ $289.88

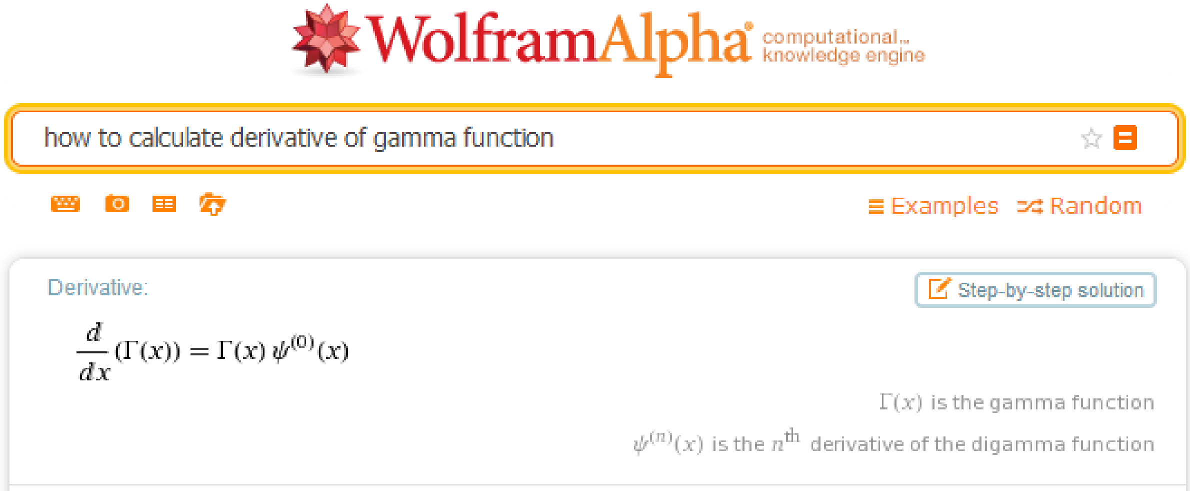
Linear Algebra
› G. E. Shilov
Paperback
★★★★☆ (34)
~~$18.95~~ $12.65

Introduction to Linear Algebra...
› Gilbert Strang
Hardcover
★★★½☆ (57)
~~$87.50~~ $83.13

Linear Algebra For Dummies
› Mary Jane Sterling
Paperback
★★★★☆ (29)
~~$19.99~~ $16.23

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Question answering



how to calculate derivative of gamma function

Derivative:

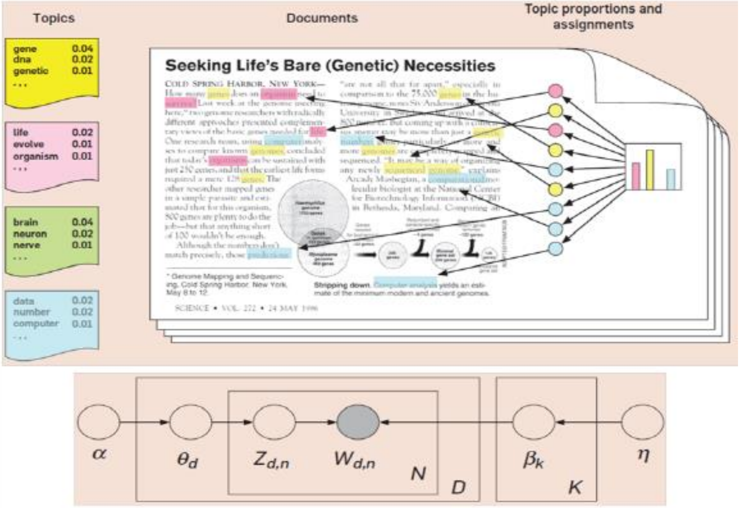
$$\frac{d}{dx}(\Gamma(x)) = \Gamma(x)\,\psi^{(0)}(x)$$

$\Gamma(x)$ is the gamma function

$\psi^{(n)}(x)$ is the n^{th} derivative of the digamma function

IR is not just about web search

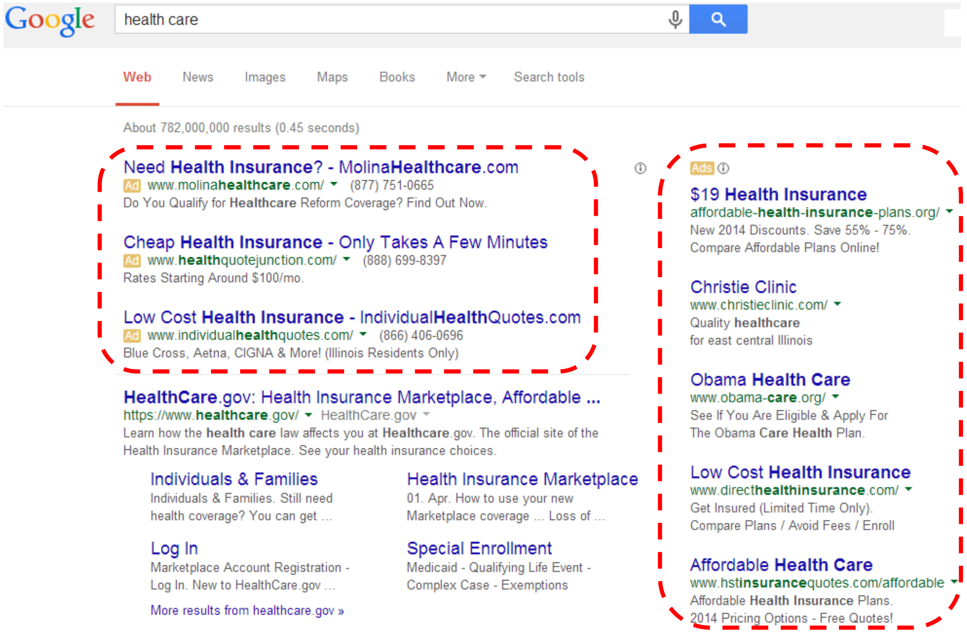
- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Text mining



[D.M Blei, Probabilistic Topic Models. Communications of the ACM, 2012]

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Online advertising



health care

About 782,000,000 results (0.45 seconds)

Need **Health Insurance**? - Molina**Healthcare**.com
Ad www.molina**healthcare**.com/ (877) 751-0665
Do You Qualify for **Healthcare** Reform Coverage? Find Out Now.

Cheap **Health Insurance** - Only Takes A Few Minutes
Ad www.**health**quotejunction.com/ (888) 699-8397
Rates Starting Around $100/mo.

Low Cost **Health Insurance** - Individual**Health**Quotes.com
Ad www.individual**health**quotes.com/ (866) 406-0696
Blue Cross, Aetna, CIGNA & More! (Illinois Residents Only)

HealthCare.gov: Health Insurance Marketplace, Affordable ...
https://www.**healthcare**.gov/ HealthCare.gov
Learn how the **health care** law affects you at **Healthcare**.gov. The official site of the Health Insurance Marketplace. See your health insurance choices.

Individuals & Families
Individuals & Families. Still need health coverage? You can get ...

Health Insurance Marketplace
01. Apr. How to use your new Marketplace coverage ... Loss of ...

Log In
Marketplace Account Registration - Log In. New to HealthCare.gov ...

Special Enrollment
Medicaid - Qualifying Life Event - Complex Case - Exemptions

More results from healthcare.gov »

Ads

$19 **Health Insurance**
affordable-**health-insurance**-plans.org/
New 2014 Discounts. Save 55% - 75%.
Compare Affordable Plans Online!

Christie Clinic
www.christieclinic.com/
Quality **healthcare**
for east central Illinois

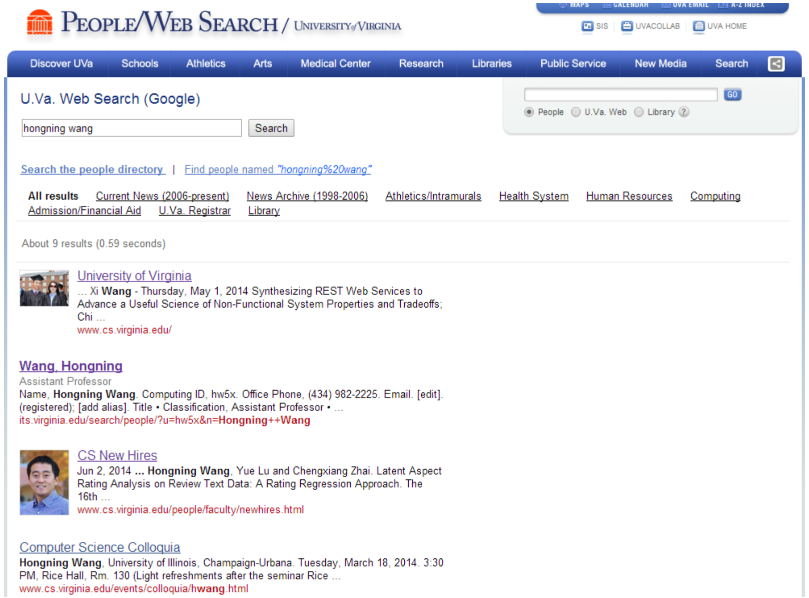
Obama **Health Care**
www.obama-**care**.org/
See If You Are Eligible & Apply For
The Obama **Care Health** Plan.

Low Cost **Health Insurance**
www.direct**healthinsurance**.com/
Get Insured (Limited Time Only).
Compare Plans / Avoid Fees / Enroll

Affordable **Health Care**
www.hst**insurance**quotes.com/affordable
Affordable **Health Insurance** Plans.
2014 Pricing Options - Free Quotes!

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Enterprise search: web search + desktop search



U.Va. Web Search (Google)

hongning wang

Search the people directory | Find people named *"hongning%20wang"*

All results Current News (2006-present) News Archive (1998-2006) Athletics/Intramurals Health System Human Resources Computing Admission/Financial Aid U.Va. Registrar Library

About 9 results (0.59 seconds)

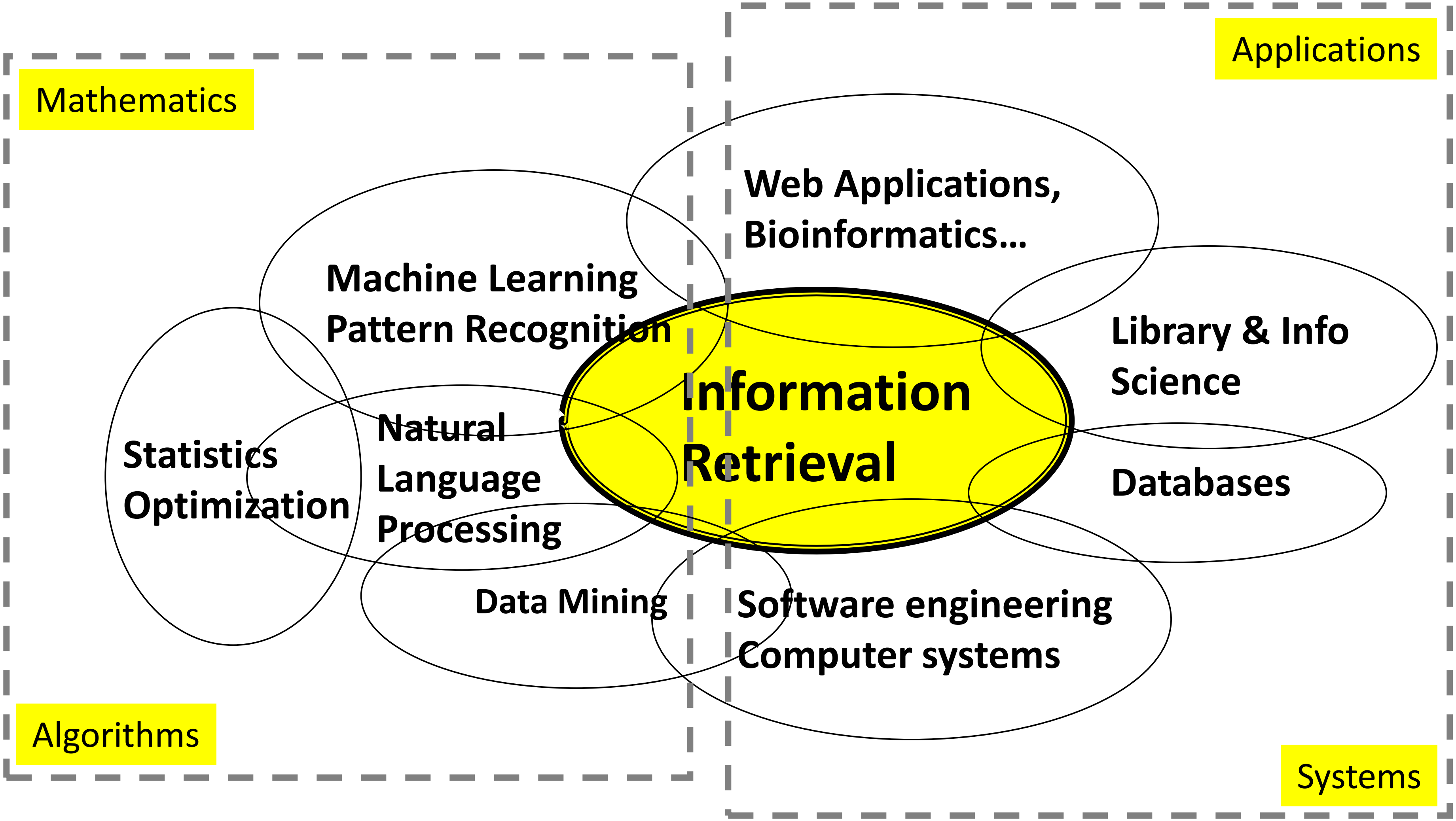
University of Virginia
... Xi **Wang** - Thursday, May 1, 2014 Synthesizing REST Web Services to Advance a Useful Science of Non-Functional System Properties and Tradeoffs; Chi ...
www.cs.virginia.edu/

Wang, Hongning
Assistant Professor
Name, **Hongning Wang**. Computing ID, hw5x. Office Phone, (434) 982-2225. Email. [edit]. (registered); [add alias]. Title • Classification, Assistant Professor • ...
its.virginia.edu/search/people/?u=hw5x&n=**Hongning++Wang**

CS New Hires
Jun 2, 2014 **... Hongning Wang**, Yue Lu and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. The 16th ...
www.cs.virginia.edu/people/faculty/newhires.html

Computer Science Colloquia
Hongning Wang, University of Illinois, Champaign-Urbana. Tuesday, March 18, 2014. 3:30 PM, Rice Hall, Rm. 130 (Light refreshments after the seminar Rice ...
www.cs.virginia.edu/events/colloquia/h**wang**.html

Related Areas



IR v.s. DBs

- Information Retrieval:
  - Unstructured data
  - Semantics of object are subjective
  - Simple key work queries
  - Relevance-drive retrieval
  - Effectiveness is primary issue, though efficiency is also important

- Database Systems:
  - Structured data
  - Semantics of each object are well defined
  - Structured query languages (e.g., SQL)
  - Exact retrieval
  - Emphasis on efficiency

IR and DBs are getting closer

- IR => DBs
  - Approximate search is available in DBs
  - Eg. in mySQL

```
mysql> SELECT * FROM articles
    -> WHERE MATCH (title,body)
       AGAINST ('database');
```

- DBs => IR
  - Use information extraction to convert unstructured data to structured data
  - Semi-structured representation: XML data; queries with structured information

IR v.s. NLP

- Information retrieval
  - Computational approaches
  - Statistical (shallow) understanding of language

- Natural language processing
  - Cognitive, symbolic and computational approaches
  - Semantic (deep) understanding of language

IR and NLP are getting closer

- IR => NLP
  - Larger data collections
  - Scalable/robust NLP techniques, e.g., translation models

- NLP => IR
  - Deep analysis of text documents and queries
  - Information extraction for structured IR tasks

Course Learning Objectives

- Enable students to understand the common algorithms and techniques for information retrieval (document indexing and retrieval, query processing, etc)
- Introduce the quantitative evaluation methods for the IR systems and data mining techniques
- Enable students to implement a basic textual information retrieval system using Java or Python
- Introduce the popular probabilistic retrieval methods and ranking principles
- Introduce the techniques and algorithms existing in practical retrieval and data mining systems such as those in web search engines and recommender systems

Course Outline

Inverted Index Construction

Posting Lists, Dictionary

Text Preprocessing

Tokenization Stopping, stemming

Retrieval Models (Vector Space Models)

Vector-space model, Cosine Similarity, Tf-Idf, BM25

Retrieval Models (Language Models)

Smoothing Methods

Relevance Feedback

IR Evaluation/ Measures

Ranking measures: R-prec, Mean Average Precision, nDCG, Reciprocal Rank

Web Retrieval

Link analysis, Markov Chains, PageRank

Recommendation Systems/ Collaborative Filtering

Semantic Similarity Measures

Word Net, Skipgrams

Word Embedding

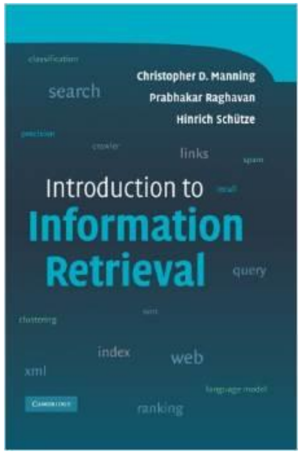
Text Classification

Naive Bayes, KNN

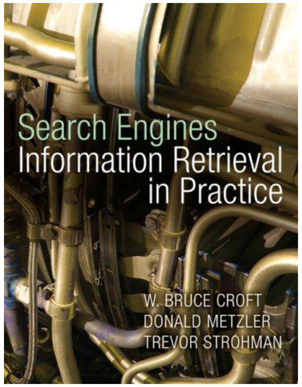
Clustering

K-means clustering

Text books



- ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

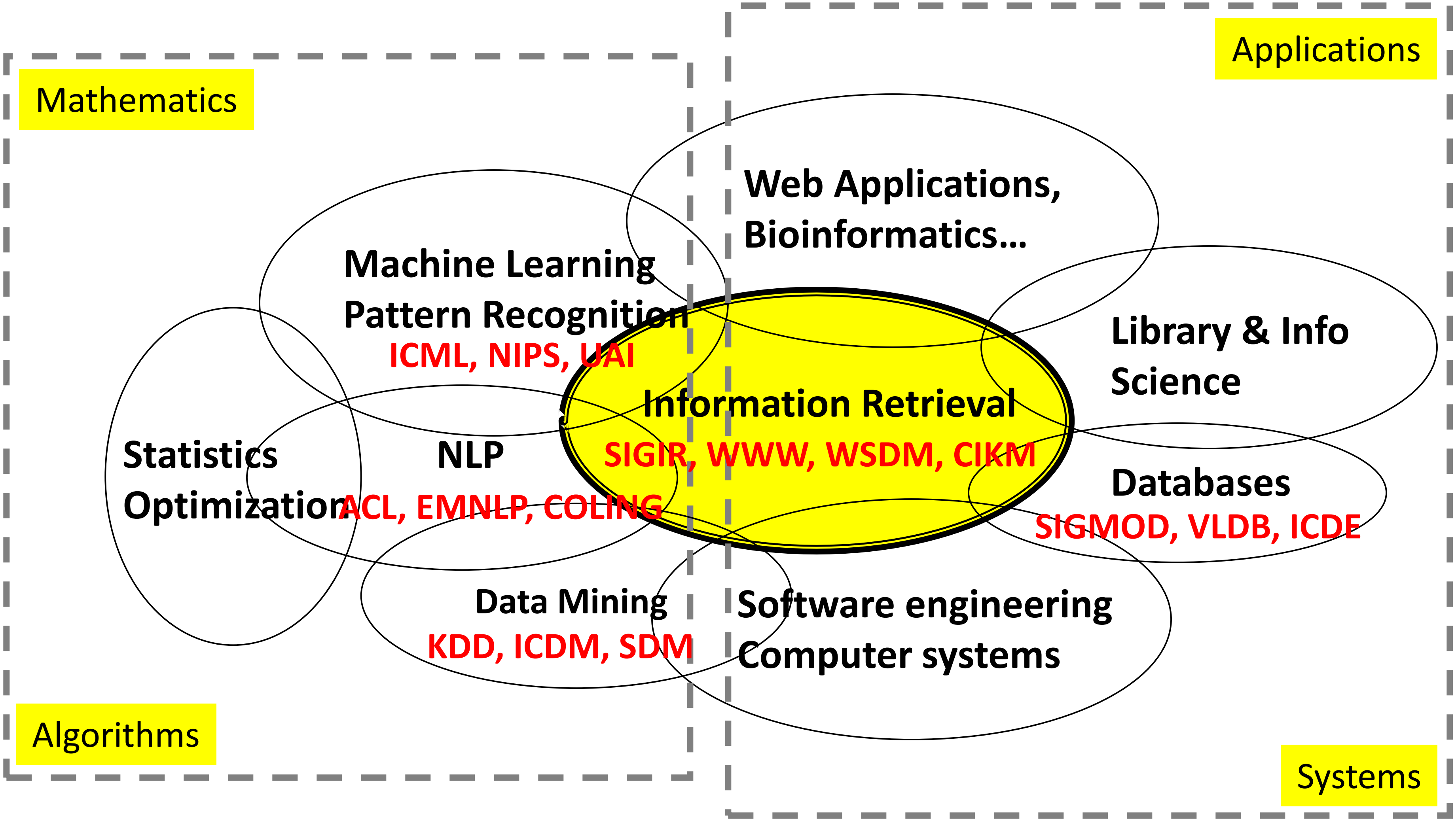


- ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

You should know

- IR originates from library science for handling unstructured data
- IR has many important application areas, e.g., web search, recommendation, and question answering
- IR is a highly interdisciplinary area with DBs, NLP, ML, HCI

What to read?


Mathematics
Applications
Web Applications,
Bioinformatics…
Machine Learning
Pattern Recognition
ICML, NIPS, UAI
Library & Info
Science
Information Retrieval
SIGIR, WWW, WSDM, CIKM
Statistics
Optimization
NLP
ACL, EMNLP, COLING
Databases
SIGMOD, VLDB, ICDE
Data Mining
KDD, ICDM, SDM
Software engineering
Computer systems
Algorithms
Systems

Top Conferences and Journals in IR Field

- SIGIR: One of the most important and influential conference in IR field (attract more attention from academia), proceedings of publications can be found here.
- WWW: Another most important and influential conference in IR field (attract more attention from industry), proceedings of publications can be found here.
- WSDM: A new but quickly raising conference in the field, attracting attentions from both industry and academia. Proceedings of publications can be found here.
- CIKM: A major conference in IR field. Proceedings of publications can be found here.
- ECIR Conference Proceedings

- TOIS: One of major journals for IR field.
- Information Processing and Management (Journal)
- Knowledge and Data Engineering (Journal)
- Information Retrieval (Journal)
- Information Science (Journal)
- Knowledge Based systems (Journal)

IR Toolkits

- Lucene (Apache)
- Lemur & Indri (CMU/Univ. of Massachusetts)
- Terrier (Glasgow)
- MeTA (University of Illinois)
- RankLib (A collection of learning-to-rank algorithms University of Massachusetts Amherst)
- General Information Retrieval Systems

NLP-related Resources

- Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources
- Stanford NLP parser (Stanford University NLP group)
- OpenNLP (Apache)
- LingPipe (Jave-based)
- NLTK(Python-based)

Machine Learning Toolkits

- Weka (A rich collection of machine learning algorithms, Machine Learning Group at the University of Waikato)
- Mallet (An alternative package for Weka, developed by Andrew McCallum at University of Massachusetts Amherst)
- LibSVM (A collection of SVMs, developed by Chih-Chung Chang and Chih-Jen Lin at National Taiwan University)
- SVM-light (Another collection of SVMs, developed by Thorsten Joachims at Cornell University)
- GraphLab (Large-scale machine learning package)
- mahout (Apache large-scale machine learning package)
- Topic Models (David Blei's collection of various topic models)

Percentage Grade Distribution

	Number	**Total Weight (%)**
Quizes	3	10
Programming Assignments	2	10
Project	1	10
Midterm	2	25
Final Exam	1	45

Plagiarism Policy

You are not allowed to copy code for programming assignments from internet or any other student. Penalty of plagiarism in programming assignments will be from one of the following depending on severity of case:

- -1 absolute from final grade
- Final grade is lowered
- F in course

Slide Credits

- Dr. ChengXiang Zhai
- Lecture Notes, Text Retrieval and Mining by Christopher Manning and Prabhakar Raghavan, Stanford University