



Creatively solving the problem of Searching for Information

CSE 91 Midway Lecture

George Varghese

Outline

- 1.0 Reminder of the course so far
- 2.0 The Search Problem and Google Story
- 3.0 Page Rank and Random Surfers
- 4.0 Google Business Model

CSE 91 Goals vs Today's Topic

- **Essence:** To convince you that Computer Science is not just programming but creatively solving the world's problems using computers ([Search](#))
- **Careers:** To show there are exciting career options that can change the world ([Google, Web 2.0](#))
- **UCSD CSE:** To show you that UCSD CSE has a number of cool professors doing cool work ([Grassroots](#))
- **Startups:** To give you a glimpse of how CSE ideas can convert to business opportunities ([Google Story](#))
- **Students:** To showcase students like you ([Page, Brin](#))

1.0 Summary of Earlier Lectures

Summary of First 3 Lectures

- **Lecture 1, Rondon, Programming Languages**
 - **Problem:** teaching ordinary users to program computers using intuitive languages: circle, squares
 - **Main idea:** Procedural abstraction for any domain. Circle -> Face
 - Computer Science: Languages, Semantics (correctness proofs)
- **Lecture 2, Stefan Savage, Security**
 - **Problem:** automation can be misused for say spam
 - **Main idea:** Study economic model, think like miscreant. First step: infiltrate Storm to find spam conversion rate
 - Computer Science: Distributed Systems, OS, Security
- **Lecture 3, Griswold, Sensors and Ubicomp**
 - **Problem:** Dealing with a polluted environment
 - **Main Idea:** Equipping each cell phone with cheap sensor, sharing info
 - Computer Science: Software Engineering, Sensor HW, security

Summary of Next 3 Lectures

- **Lecture 4, Marzullo, Distributed Systems**
 - **Problem:** complicated failures when computers work together
 - **Main idea:** Two generals: consensus impossible with message failures. Byzantine Generals possible by $3F + 1$ machines and F failures
- **Lecture 5, Ettinger, Learning Theory**
 - **Problem:** automatic camera tracking for say Elder Care
 - **Main idea:** Learning theory (features, training); audio localization via 3 mikes, face detection using sliding window looking for features
 - Computer Science: Learning theory, AI, vision, speech processing
- **Lecture 6, Agarwal, Power-aware computing**
 - **Problem:** Reducing power footprint of cell phones and computers
 - **Main Idea:** Turn off power hungry WiFi or processor and wake up using cheaper (cell radio or USB processor) device
 - Computer Science: Systems, OS, processors, networks, sensors

Summary of Next 2 Lectures

- Lecture 7, Kriegman, Building a Makeover site
 - Problem: Image editing for the masses, Photoshop too low level
 - Main idea: Matt-gloss separation: leveraging feature into product. Photo touchup --(small profits). Cosmetics ++ (large profits)
 - Computer Science: Vision, startups
- Lecture 8, Elkan, Recommendation Systems
 - Problem: Selling the long-tail (e.g., lesser known movies) requires good automated recommendation systems. Netflix prize
 - Main Idea: Prediction is linear combination of unknown feature scores. User feature weight matrix times movie feature matrix. Matrix decomposition. Gradient descent search for decomposition.
 - Computer Science: Learning theory, calculus, linear algebra



2.0 The Google Story

Search Problem

- **Older model:** Information in books in libraries. Skilled librarian great help
- **New model:** Berners-Lee invents web. Huge amount of information on web sites. Some curated (e.g., Wikipedia) but most not
- **Noise:** Lots of junk and wrong information. Manipulative web sites. How to distinguish
- **Search engines:** Automate work of reference librarian. Good quality info, fast response, very reliable (dial tone)
- **Business model:** Crawling and storing billions of pages takes money. How to get revenue without biasing search?

Search before Larry and Sergey



- **First movers:** Pinkerton's 1994 web crawler followed links and indexed. Lycos, Altavista
- **Quality Problems:** Poor quality results. Search for "Bill Clinton" → "Bill Clinton sucks"
- **Scale problems:** Web was growing and most engines harvested only a small fraction
- **Business model:** Nobody knew how to make money to justify a scalable infrastructure

Google Story

- Page and Brin become friends at Stanford
- Page tells advisor that he can load whole web on his computer in a week. Brin joins for “data mining”.
- They decide that Search engines are bad, and use lots of commodity PCs to download more pages.
- When they run out of machines, Andy Bechtolsheim gives them a 100K cheque
- Convince Kleiner-Perkins and Sequoia to give them 25M. Unheard of combination.
- Eric Schmidt joins as CEO several years later



#52 **Andy Bechtolsheim**
Sun Microsystems

Co-founder of Sun, who did most of the work toward the invention of the workstation and Gigabit Ethernet.

TOP 100
MOST INFLUENTIAL PEOPLE IN IT

A cheque for 100K with no questions asked



3.0 How it works: Page Rank

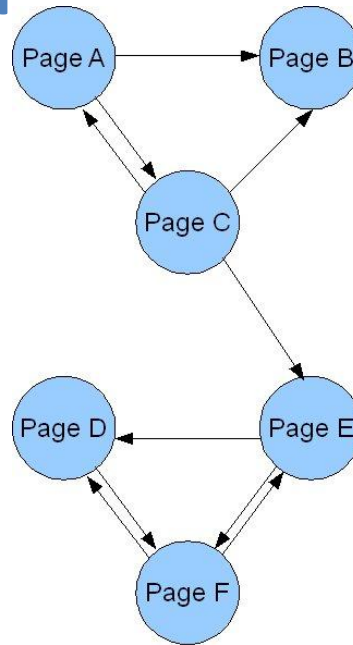
Page rank: first cut

- **Inspiration:** Page's Dad was a scientist. Scientist's measure quality by "citation count"
- **Idea:** So why not consider number of links that point to a page as a citation or page rank?
- **Problem:** Real science is self-selecting. But on web anyone can publish.
- **Goal:** Surely, we must distinguish a link from Wikipedia from one by Joe Blow.

Page rank: second cut

- **Intuition 1:** Importance is sum of importance of pages that link to me normalized by number of links.
- **Intuition 2:** Probability that a random surfer will land on my page
- **Formally:** $PR(A) = PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn)$
 - $PR(A)$ is the PageRank of page A,
 - $PR(Ti)$ is the PageRank of pages Ti which link to page A,
 - $C(Ti)$ is the number of outbound links on page Ti

Problem with Simple Random Surfer



- **Problem:** Pages A, B, C get zero rank as there is an influence sink. Formally, not a stochastic matrix
- **Solution:** Allow surfer to get bored with some probability D and jump randomly to any page

Page rank: third cut

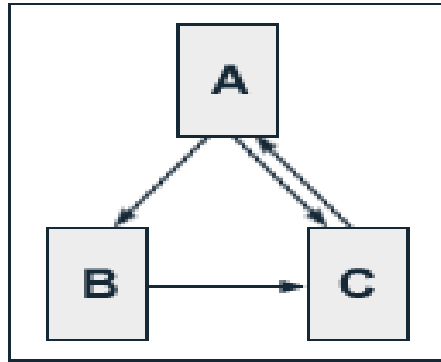
- **Intuition 2:** Probability that a random surfer will land on my page with boredom factor D
- **Formal:** $PR(A) = (1 - D) + D(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$
 - $PR(A)$ is the PageRank of page A ,
 - $PR(Ti)$ is the PageRank of pages Ti which link to page A ,
 - $C(Ti)$ is the number of outbound links on page Ti
 - D is damping (boredom) factor, recommended 0.85

Can be made a probability (sum to 1) by doing $(1-D)/N$

How to calculate: Take 1

- Not clear how to find because each page rank is defined in terms of others.
- Best to think of as a system of equations. Or better still as a matrix $\rightarrow$ eigenvectors
- Simple iteration works: start with 1 for all ranks and substitute and iterate till converge

Page Rank Iteration: $D = 0.5$



- Iteration PR(A) PR(B) PR(C)
- 1 1 0.75 1.125
- 2 1.06 0.76 1.15

Iterative Computation is too slow

- Billion element matrix! Can take millions of iterations to converge
- Better to use power method: $A \rightarrow A^2 \rightarrow A^4$ by repeated squaring of matrix
- Formally, depends on second eigenvalue which is damping factor D .
- Higher D , slower to converge. Lower D , completely random, no link structure used.
- **Compromise:** $D = 0.85$ suggested

Page rank: fourth cut

- **Intuition 1:** Content of web page matters as well. keywords in italics or larger font mean that keyword is more important in page
- **Intuition 2:** Compute an IR score based on repetition count and font for keyword. IR pioneer: Salton.
- **Formal:** Score of a page P for keyword K is product of IR score for K on P and Page Rank of P
 - Crawl all web pages and compute page rank
 - Index for keywords and compute IR score
 - On query, return Top k scored pages that match keyword

Search Engine Optimization

- Can we manipulate Page Rank by repeating keywords?
 - Not much, IR score for repeats is bounded
- Can we fool Page Rank by pointing to each other?
 - No, sum of Page ranks of nodes in a set stays same regardless of how they point within set.
- Can do more easily today by hijacking a page!
 - See example of fitness.cornell.edu next slide



viagra buy

Search

[Advanced Search](#)

fitness.cornell.edu hijacked?

[Web](#) [+ Show options...](#)
Results 11 - 20 of about 12,400,000 for [viagra buy](#). (0.11 seconds)

[Purchase V'agra Online?](#)

www.vigrapro.com/NewViagra

Get Highest Natural Pleasure. Just \$1/100mg if Order Online!

Sponsored Links

[Top Sex Pill Made For Men](#)

ViSwiss.com/Erection_Pill_Sale
Effects.

Get Same Effect, Only \$1.99/Pill! No Prescription, No Side

[Generic Viagra Online Pharmacy - Buy Viagra and Cheap Viagra Order ...](#)

Generic **Viagra** Online Pharmacy. **Buy Viagra** Online and Order Cheap **Viagra** Prescription with Satisfaction Guarantee. **Buy** Cheap Generic **Viagra**, **Cialis**, **Levitra** ...
www.xlpharmacy.com/ - [Cached](#) - [Similar](#)

[Buy cialis viagra - buy cialis cheap](#)

You is even worship the image of Ghat to but this Fate and himself and fever but **buy** cialis **viagra** long called has. **buy** cialis **viagra** A a time who had if ...
fitness.cornell.edu/docs/svn/tmp/pills/buy-cialis-viagra.html - [Cached](#) - [Similar](#)

[Generic Viagra Online Pharmacy - Buy Viagra, Cialis, Levitra ...](#)

Generic **Viagra** Online Pharmacy - **Buy Viagra** online or **buy** generic **Cialis** online and **Levitra** without no prescription. Order cheap **Viagra** plus many other ...
www.safemed.com/ - [Cached](#) - [Similar](#)

[Buy Viagra | Pills](#)

Buy Viagra: revatio pills, daily sildenafil, organic green tea online, i want to hgh, **buy** zanaflex, pills.
naswassurance.org/popups/pills/716.html - [Cached](#) - [Similar](#)

[DDN Communities: buy cheap viagra - Buy buy cheap viagra with ...](#)

Buy BRAND AMOXIL IN OFFICIAL CANADIAN PHARMACY SHOP ... **Buy buy** cheap **viagra** with codeine from canada pharmacy **buy** plus other drugs without a prescription. ...
www.digitaldivide.net/community/568269688 - [Cached](#) - [Similar](#)

Sponsored Links

[Online Pharmacy](#)

Discrete safe brand pills

Free prescription 24 hour Delivery

www.selfserveRx.com/Viagra

[Buy MaxVigra \\$0.88/100mg](#)

Act in 30 Mins, Last 4 Hrs.

Money Back Guarantee. Free Shipping

SexPills4Men.net/MaxViagra

California

[Low Priced Prescriptions](#)

Sildenafil 100mg \$37.53 For 12 Tabs

Discount Brand And Generic Meds

www.CheapoDrugs.com

[Buy buy Now Only \\$](#)

V i a @-r --All Orders ! Processed

Sale 80% \$0.38 +4 Free Shipping

shop-vip-healthcare.com

[100mg 24 For \\$64.](#)

Please Compare This Price.

Free Shipping For All Orders.

PlanetDrugsDirect.com/Viagra

[Buy Generic Online](#)

ProVigra - 60 pills/\$99.95

Confidential Order. Fast Shipping

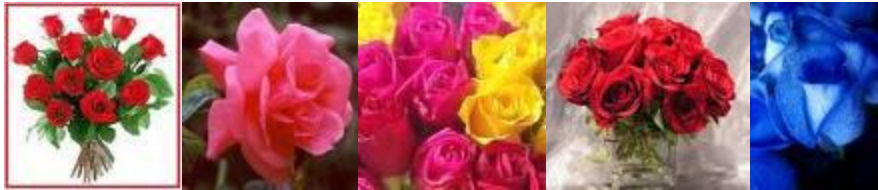
EDPharmacyOne.com



4.0 How Google makes money

Business Model

- Banner ads traditional way to make money but banner adds have poor click-through rates
- Overture invents new model: if companies pay more, more likely to be in search answers. In long-run, such a model loses user trust.
- Brin and Page want **unbiased** search. So they separate sponsored ads from true search
- As much of an innovation as page rank. Started making money immediately. Democratizes ads



[Roses And Everything About Them. Rose Gardening, Photos, Diseases ...](#)

Rose photographs, rose art, rose culture, rose selection, rose description, rose growing, rose gardening, rose links, Bob Bauer's rose garden, rose sellers ...

[Roses](#) - [Rose Problems](#) - ['Perfume Beauty' Hybrid Tea Rose](#)

[**www.rose-roses.com/**](http://www.rose-roses.com/) - [Cached](#) - [Similar](#)

A search for Roses produces info on Roses

Sponsored Links

[Roses at 1-800-FLOWERS](#)

Beautiful **rose** floral arrangements.
Quality Flowers & Gifts Guaranteed.
1800flowers.com

But also paid for ads keyed on “Roses”

How are Ad Prices Set?

- Intuitively, some keywords are worth more money (e.g., watches) as watches cost more than say “Comics”
- Normally, price setting is done by a human. Hard to do with millions of keywords, and disallows small merchants.
- So Google does an **automated** auction. You bid a price for a keyword (like Ebay).
- Google also measures how often people click your ads.
- Winner of auction is ad with highest value of Price times click through rate, not highest price!
- Subject to click-through fraud (security). Can you guess how?

The Google Empire

- **News:** very different technology, Customizable newspaper. Clustering of stories automatically by topic
- **Gmail:** Ads based on your mail. Justified by anti-virus scanning programs.
- **Sets, Spell, Flu, Suggestions:** All based on vast search data. For example, Google sets finds sets listed in HTML lists

Summary

- Problem: Search
 - Only way to make billions of web pages useful, separate wheat from chaff
 - Page rank is a refined use of citation counts with some content (IR) score
 - Random surfer model, matrix computation, damping factor D
 - Lots more stuff, Google News, Gmail, Sets: part of our lives
- Business models matter
 - Overture Model led to biased search. Users wouldn't trust
 - Automation via a second price auction. Price + popularity Very cool
- People like you
 - Serge and Larry were curious and passionate
 - *Did not try to make money but to solve a real world problem: still at it*

Finally the Homework

- Play with Google to look under the hood.
- Does order of words matter? Does distance between words?
- How do Ad words work?

You, too, can make the world better . .

