

# Foundations and Future of Information Search

Andrea LaPaugh  
Department of Computer Science

Lunch 'n Learn

March 4, 2009

# Historic Goals

“A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.”

Vannevar Bush, *As we may think*, *Atlantic Monthly*, July 1945.

“Google's mission is to organize the world's information and make it universally accessible and useful” [Google's mission statement](#), ~ 1998.

# Vannevar Bush's 1945 vision

- Director of the Office of Scientific Research and Development (1941-1947)
- End of WW2 - what next big challenge for scientists?



Vannevar Bush, 1890-1974

"This is a much larger matter than merely the extraction of data for the purposes of scientific research; it involves the entire process by which man profits by his inheritance of acquired knowledge"

# Prophetic: Hypertext

❄ "associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing."

# How have we achieved search capability?

- Vannevar Bush envisioned personal index
- General open collections
  - keyword/subject-based search
  - 👉 full-text search
  - 👉 hypertext enhanced search

# Full-text search: beginnings

- Gerald Salton, founding father of **information retrieval**
  - SMART retrieval system
- major idea: score documents by frequency of words of query occur
  - take into account
    - document length
    - frequency of words in collection
- one of many major contributions



Gerald Salton  
(1927-1995)

# Frequency Model example

**Doc 1:** “Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies... Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Frequencies:

science 1; knowledge 2; principles 0; engineering 0

**Doc 2:** “An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science ...” (cos 126 description)

Frequencies:

science 2; knowledge 0; principles 1; engineering 1

# Scoring documents (vector model)

	frequency-based		adjusted for word value	
	Doc 1	Doc 2	Doc 1	Doc 2
“science”	1	2	.51	1.02
“engineering”		1		1.6
“principles”		1		1.6
“knowledge”	2		3.2	
<b>Combined SCORE</b>	<b>3</b>	<b>4</b>	<b>3.71</b>	<b>4.22</b>

# Using word occurrence and features

- word **frequency** in documents
- **positions** of words in documents
- appearance in **special parts of documents**
  - title
  - abstract
  - section header
  - ...
- special **features of word**
  - bold font
  - larger font
  - ...

plain text

marked-up text

# Interpreting queries

- sequence of words - what look for
  - vector score requires returned documents **contain only some of words**
    - weight of query word in document
    - number of query words in document
  - Web search engines require returned documents **contain all words**
    - sort of

# Enhancing queries

- Special features in queries
  - word OR word: **cat** OR **dog**
  - NOT word: **cat**, NOT **dog**
  - phrase: “**peanut butter**”
  - ...
- retrieval system modify queries
  - remove “noise words” (stop words): **a, an, the ...**
  - find stems of words: **cats** → **cat**
  - add synonyms

Search engines’  
“Advanced Search”

# The Index

- Retrieval systems record all info will use about document in an **index**
- index **organized by word**
  - all words in all documents = lexicon
- for each word, index records list of:
  - **documents** in which it appears
    - **positions** at which it occurs in each doc.
    - **attributes** for each occurrence
- record summary information for documents
- record summary information for words

# Along came the Web

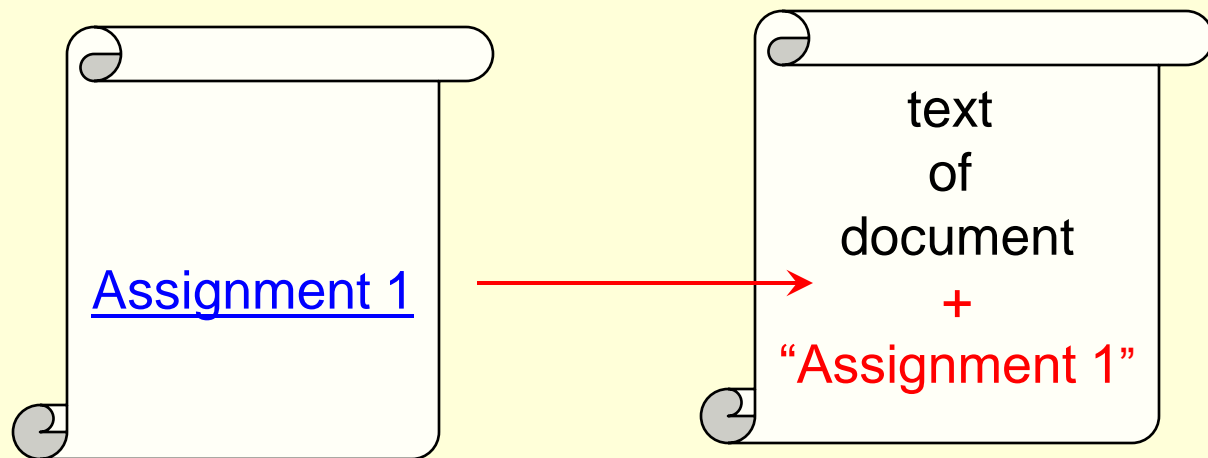
- Major new element: **links**
  - hypertext
- How use links?
  - **expand documents** to rank
  - **anchor text**
  - **link analysis**

# Anchor text

- the words used when making a link to another document:

“[Assignment 1](#) is now available.”

- Add words of anchor text to document pointed to



# Example

7<sup>th</sup> result for search on Google:  
“cat dog garden turtle night cake”

This is Google's cache of <http://www.sugarplumdreams.com/nl-blue-camo-turtle.html>. It is a snapshot of the page as it appeared on Feb 24, 2009 23:22:49 GMT. The [current page](#) could have changed in the meantime. [Learn more](#)

These search terms are highlighted: **cat dog garden turtle night** These terms only appear in links pointing [Text-only version](#) to this page: **cake**

Welcome! [Sign-In](#) or [Create An Account](#)

[View Cart](#)

Call: 888-747-7586

Shop By Category

Shop by Theme

Shop By Brand

Shop By Gender

Search



[Home](#) > [Children's Room Decor](#) > [Lamps and Lighting](#) > [Nightlights](#) > [Night Light - "Blue Camo Cooter Turtle" Straight Border](#)

**Night Light - "Blue Camo Cooter Turtle" Straight Border**

# Example

12<sup>th</sup> result for search on Google: toxin

in URL

This is Google's cache of <http://www.toxin.org/>. It is a snapshot of the page as it appeared on Feb 9, 2009 17:32:35 GMT. The [current page](#) could have changed in the meantime. [Learn more](#)

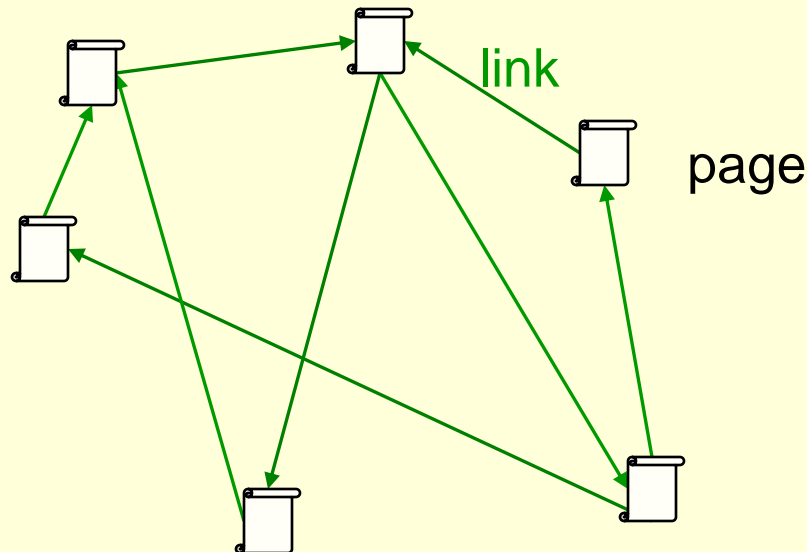
These terms only appear in links pointing to this page: toxin

[Text-only version](#)

I upgraded the system. Cutover was on 6/2/2008. Everything should still work, but if it doesn't, email [andr00@earthlink.net](mailto:andr00@earthlink.net)

# Link analysis

- Intuition:  
when Web page **points** to another Web page,  
**confers status/authority/popularity** to that page
- Find a scoring of pages that **captures intuition**

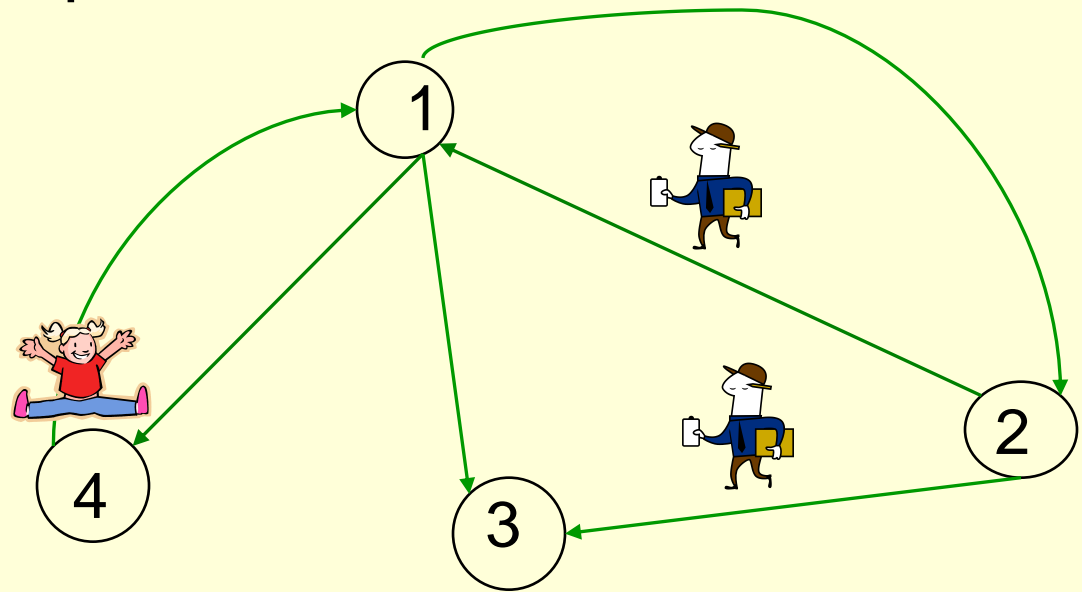


# PageRank

- Algorithm that gave Google its “killer” performance

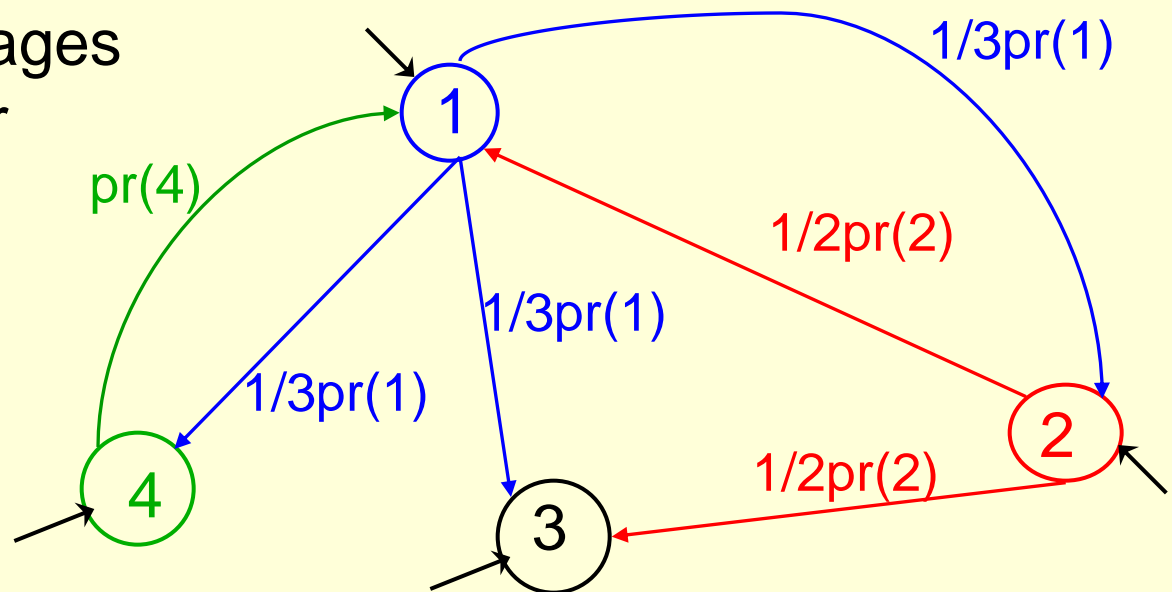
Larry Page, Sergey Brin, R. Motwani, T. Winograd  
(1998)

- Random walk model  
+ random leap

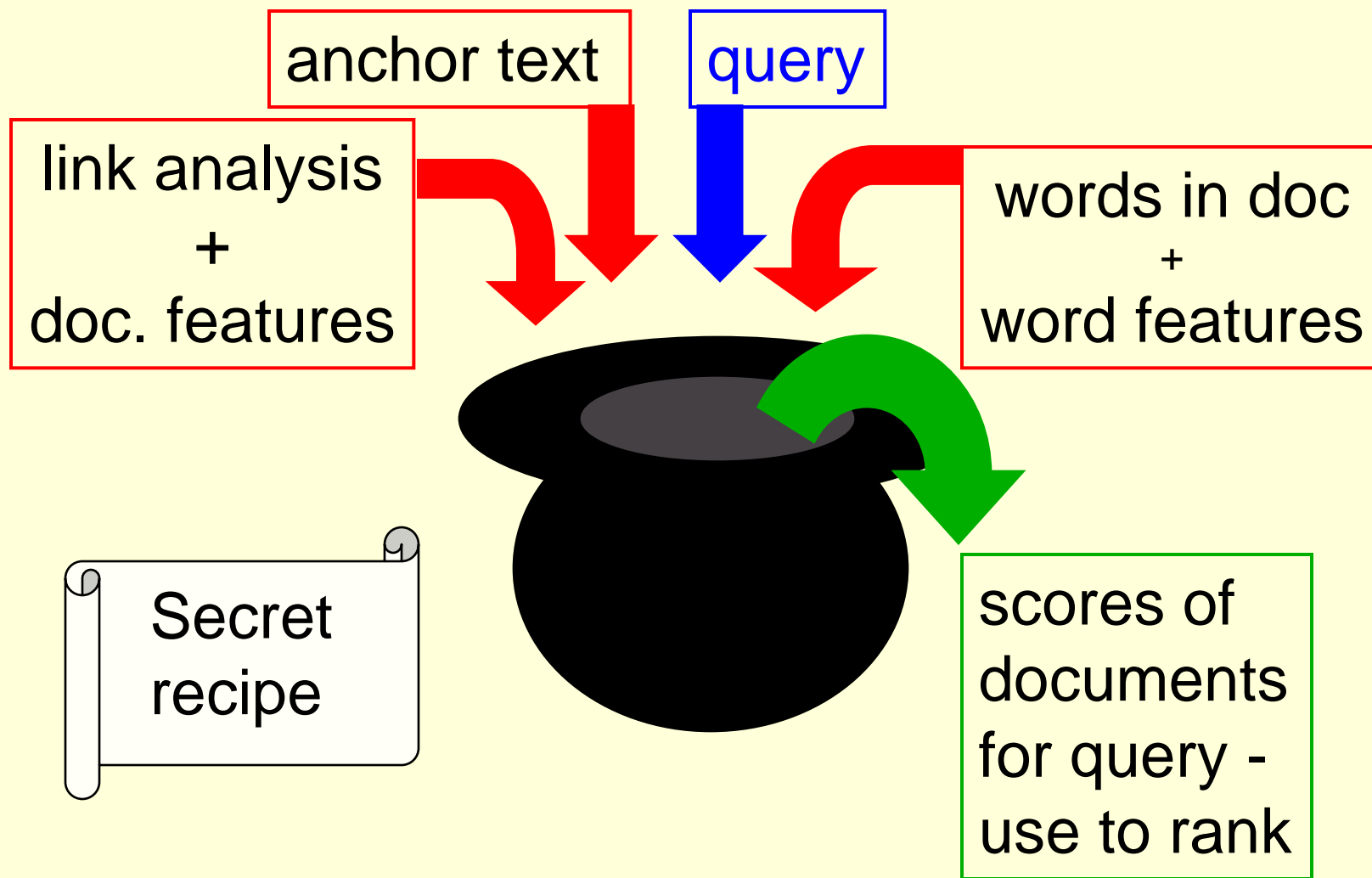


# Define score from concept

- Each page  $i$  gets a PageRank score:  $pr(i)$
- **No query involved**
- Equation  $pr_{\text{updated}}(k) = \alpha/n + (1-\alpha)\text{Sum}(k)$ 
  - $t_i$  is number of links out of page  $i$
  - $\text{Sum}(k)$  is the sum of  $(pr(i) / t_i)$  over all pages  $i$  pointing to page  $k$
  - $n$  is number pages
  - $\alpha$  is parameter



# Ranking documents w.r.t. query



Example: “peach”

# Non-text objects

- images
- music
- video
- ...

# Non-text: current methods

- use words around embedded objects
- use names of embedded objects
  - Search Google Images for "only":  
top hit with image file  
"Smokey-Bear-Only-You-Posters.jpg"
  - use tagging: folksonomy flickr
- use features of object representation
  - CASS Project at Princeton Computer Science:  
Kai Li, Moses Charikar, Perry Cook, Olga Troyanskaya, Jennifer Rexford
  - One of several university or commercial projects

# Improving Web Search?

- More
  - predicting topics without word clues
  - “deep web”
- Better
  - question answering
  - natural language queries
  - more useful presentation

# Deep Web

- *Exploring a 'Deep Web' That Google Can't Grasp*, NY Times, Feb 22, 2009
- Much of info on Web behind databases
- Must query database to get info
- How search engine generate right queries on right database
- clues
  - ✧ text on front page & language analysis
    - user behavior
    - link analysis

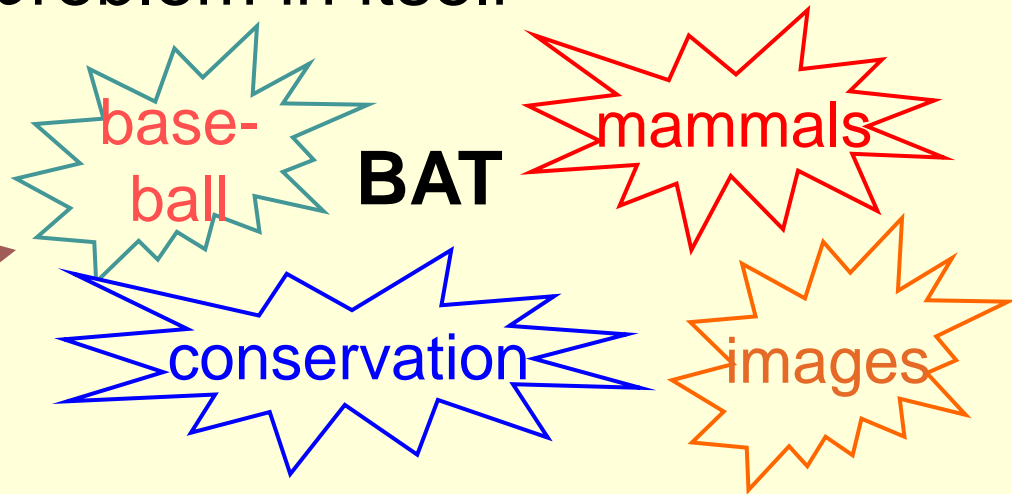
# Presentation

- Cluster similar results by labeled topic

- clustering major problem in itself

- content based
- network based

- **Clusty.com** →



- Summarization

- **Cuil.com**

# Predicting topic without text clues

- **Connectivity analysis** in Web
- **User behavior**
  - use temporally local visit behavior of user
    - know category of some pages visited
    - predict category of others
    - need “invested” visiting

# Improving search results with user behavior

- Aggregate behavior
  - what was most popular selection?
    - Ads
- Personal behavior
  - own history of preferences
    - type of sites?
    - specific sites?
  - disambiguation
- Behavior of users like you
  - compare your behavior to others
  - see what others did in new situation

# Ongoing research

- Machine learning
- network analysis
- sampling methods
- approximation methods
- architectures for large data

# Amount of information

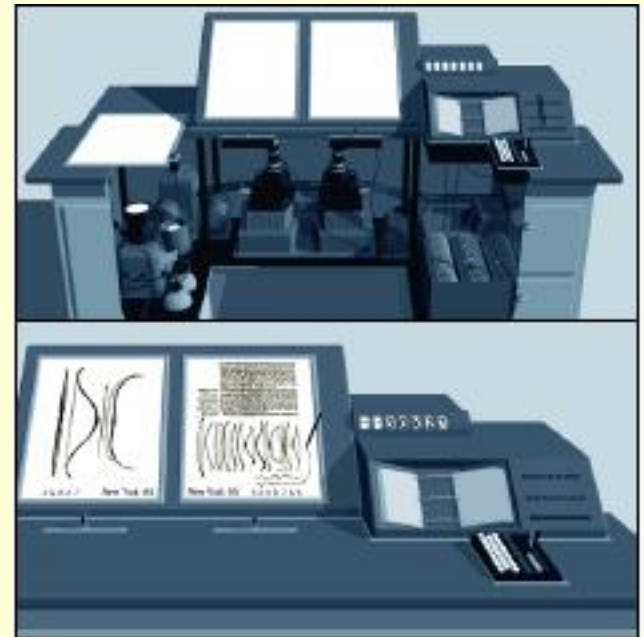
- For Web search engine
  - lexicon: million or more words and terms
  - documents: 10 - 100 billion
    - Cuil claims 124 billion Web pages

# Historic Goals

- “ an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.”
- “ to organize the world's information and make it universally accessible and useful”

# Envisioned “memex”

“A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.”



# Prophetic: [Hypertext](#)

- "associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing."
- "Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified."

# Non-text: future methods

- innovation in
  - feature definition
  - analysis
  - Sampling
  - Semantic Web
  - Architecture