

cuil

Search 127 billion pages

[About Cuil](#) | [Preferences](#) | [Add Cuil to Firefox](#)

[Privacy Policy](#) | © 2010 Cuil, Inc.

Charting SearchLand: search quality for beginners

Valeria de Paiva
Cuil, Inc.
Aug 2010

Check <http://www.parc.com/event/934/adventures-in-searchland.html>



Charting SearchLand

SearchLand? codename for
search in start-up culture

Academic research is a
landscape I understand and
can travel in (if you change the field
you do research, you keep the tools and
the methods)

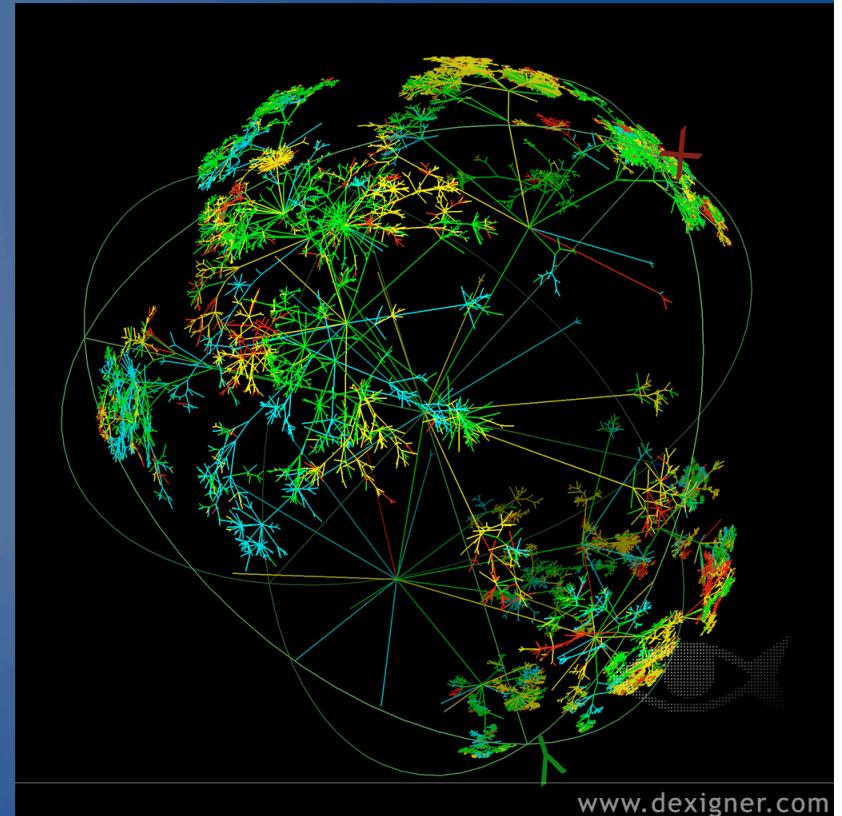
Searchland feels very foreign
and it needs charting

Even a primitive map is better
than no map at all...



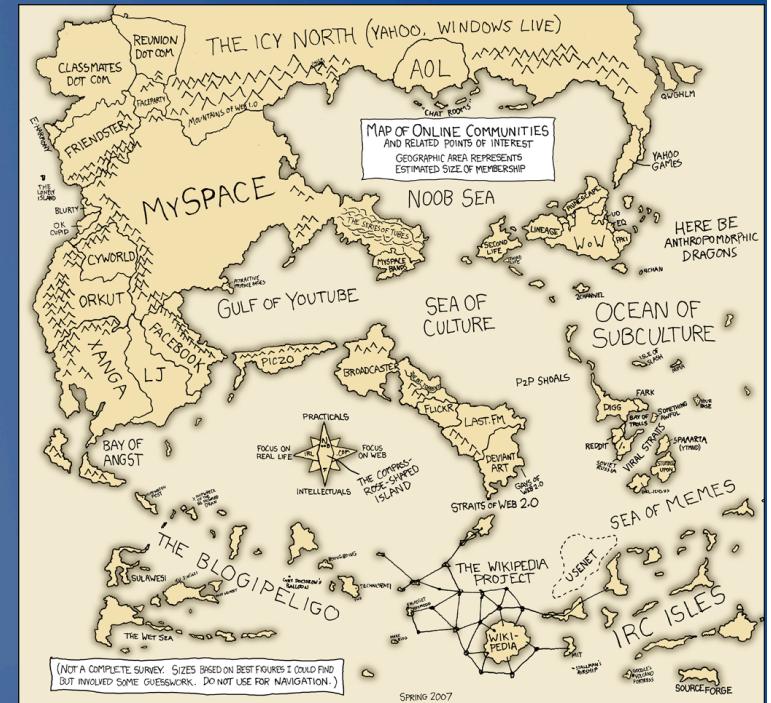
Outline

- SearchLand
- Search engine basics
- Measuring IR systems
- Measuring search quality
- Conclusions...
- and Opportunities



SearchLand?

“Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a **black art** and to be advertising oriented.” Brin and Page, “The anatomy of a search engine”, 1998

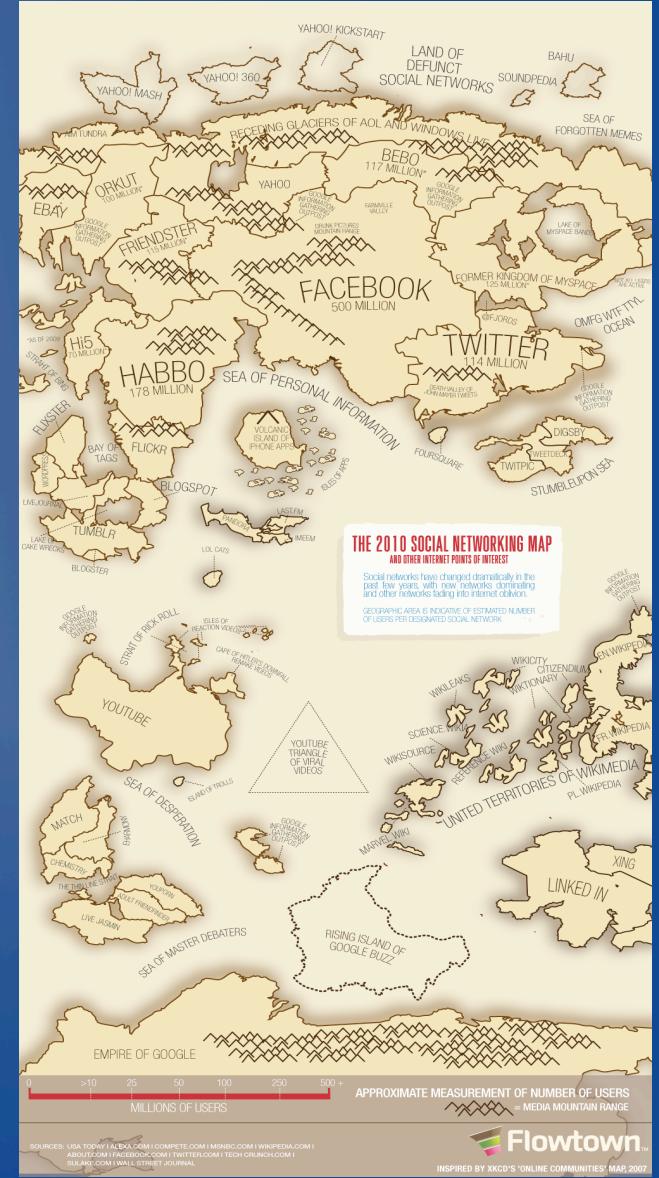


**Disclaimer: This talk presents the guesswork of the author.
It does not reflect the views of my employers or practices at work.**

Thanks kxdc2007!

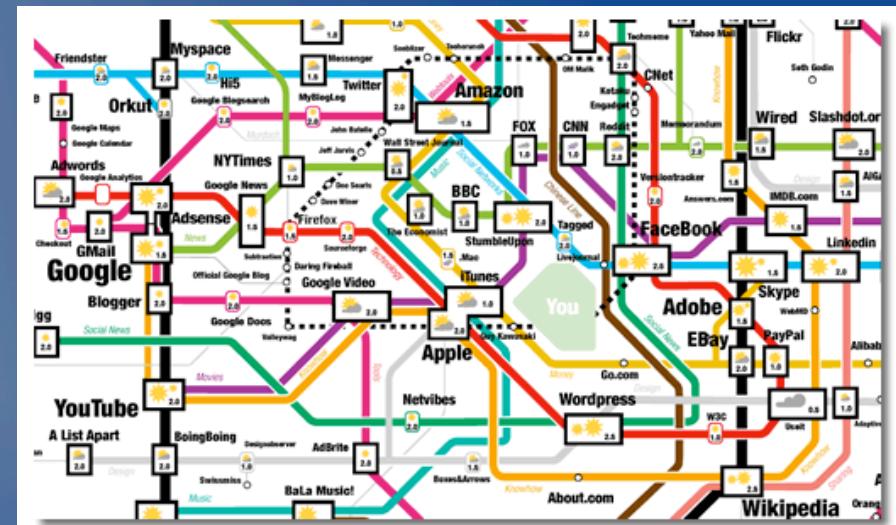
SearchLand

- Twelve years later the complaint remains...
- Gap between research and practice widened
- Measuring SE quality is `adversarial computing'
- Many dimensions of quality: pictures, snippets, categories, suggestions, etc



SearchLand: Draft Map

Based on slides for Croft, Metzler and Strohman's "Search Engines: Information Retrieval in Practice", 2009 the tutorial '**Web Search Engine Metrics**' by Dasdan, Tsioutsiouliklis and Velipasaoglu for WWW09/10 and Hugh Williams slides!



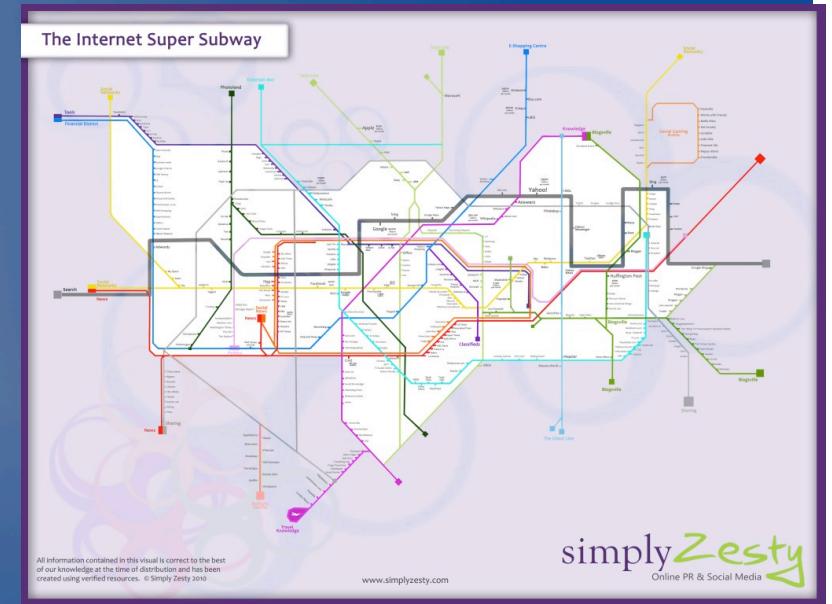
THANKS GUYS!!...

Search engines...

My understanding of search engines is very basic: I know when I like results and when I don't. (yes, you'll get the user perspective here...)

Assuming you're like this too, some basics, from folks that know more.

Basic metaphor: search engine as a librarian in super-hyped library...



Search Engine Basics...

Web search engines don't search the web:

They search a *copy* of the web

They *crawl* documents from the web

They *index* the documents, and provide a search interface based on that index

They present short *snippets* that allow users to judge relevance

Users click on links to visit the actual web document

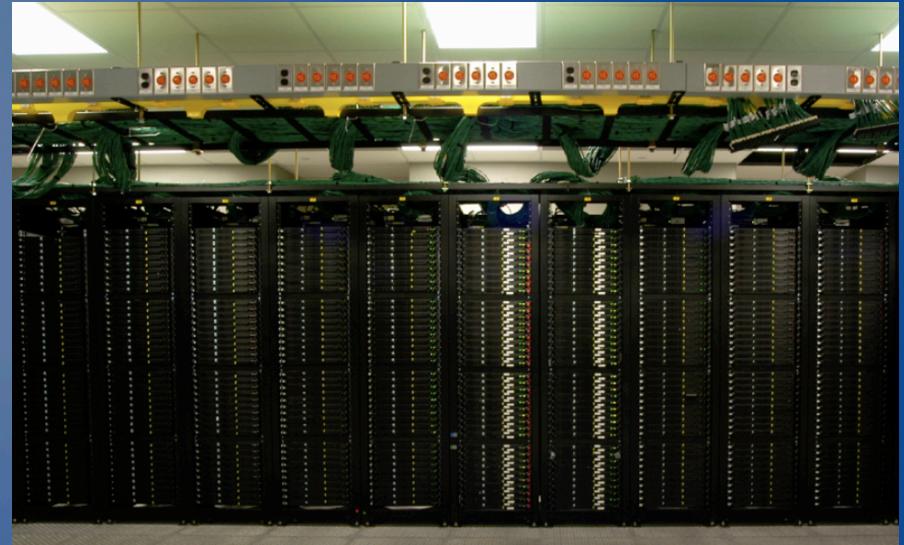
(thanks ACMSIG May 2010)



Search Engine Basics

Modules: Crawler

- Indexer
- [content mining]*
- Query analyzer
- Query server
- Index server
- Snippeter
- Ranking
- Webserver, etc...



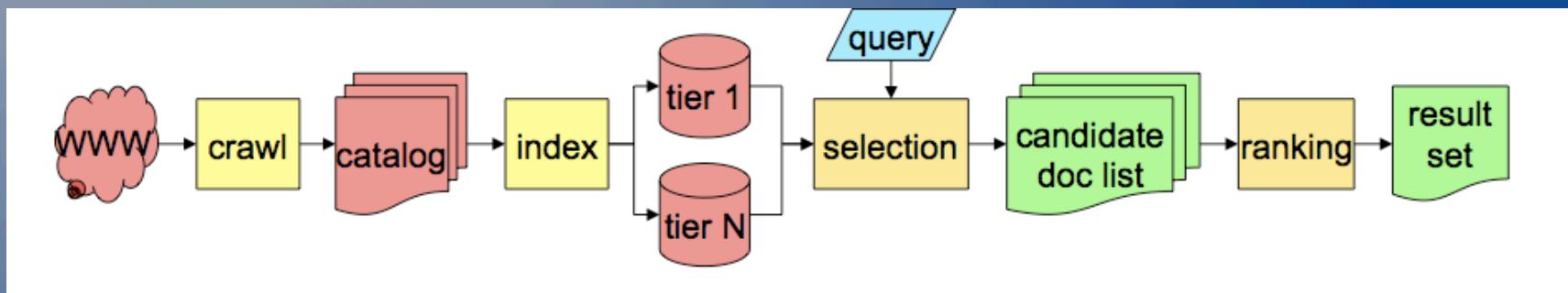
Why writing your own search engine is hard, Patterson, ACM Q, 2004

Search Engines: Information Retrieval in Practice, Croft, Metzler and Strohman, Addison Wesley, 2009

picture: Trevor S.

Search Engines

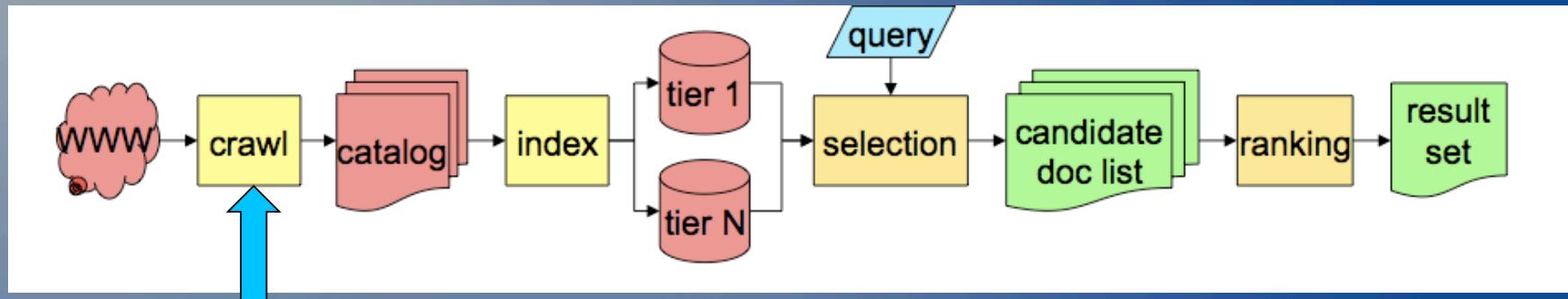
Basic architecture



From **Web Search Engine Metrics for Measuring User Satisfaction**
Tutorial at WWW conference, by Dasdan et al 2009,2010

Crawlers

First one needs to crawl the pages.



“you don’t need a lot of thinking to do crawling; you need bandwidth, so any old CPU will do”.
Patterson, 2004

BUT Writing a crawler isn’t straightforward...

Crawlers

writing/running a crawler
isn't straightforward...



Must respect robots.txt exclusion standard to limit which pages should be retrieved

Crawler shouldn't overload or overvisit sites

Many URLs exist for the same resource

URLs redirect to other resources (often)

Dynamic pages can generate loops, unending lists, and other traps

URLs are difficult to harvest: some are embedded in JavaScript scripts, hidden behind forms, and so on

Crawlers

Crawlers actually need to

Fetch new resources from new domains or pages

Fetch new resources from existing domains or
pages

Re-fetch existing resources that have changed

Crawl **prioritization** is essential:

There are far more URLs than available fetching
bandwidth

For large sites, it's difficult to fetch all resources

Essential to balance re-fetch and discovery

Essential to balance new site exploration with old
site exploration

Snapshot or incremental? How broad? How
seeded?

Brin and Page again..



Crawler Challenges

Not Found pages often return ok HTTP

codes

DNS failures

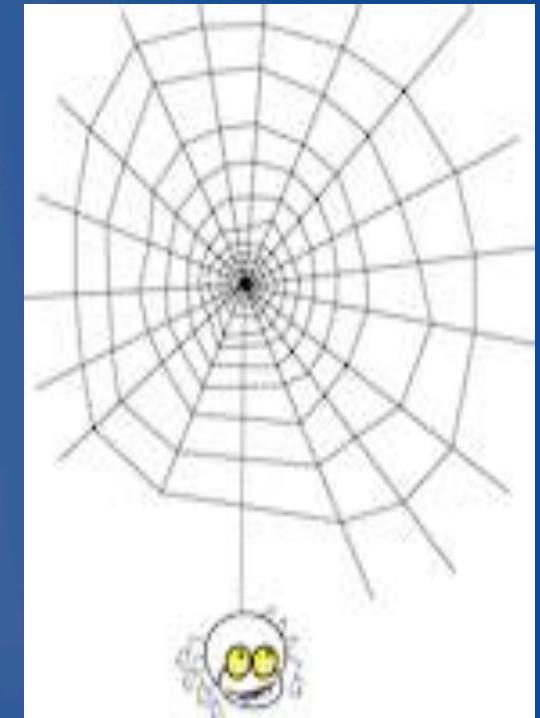
A pages can redirect to itself, or into a cycle

Pages can look different to end-user browsers and crawlers

Pages can require JavaScript processing

Pages can require cookies

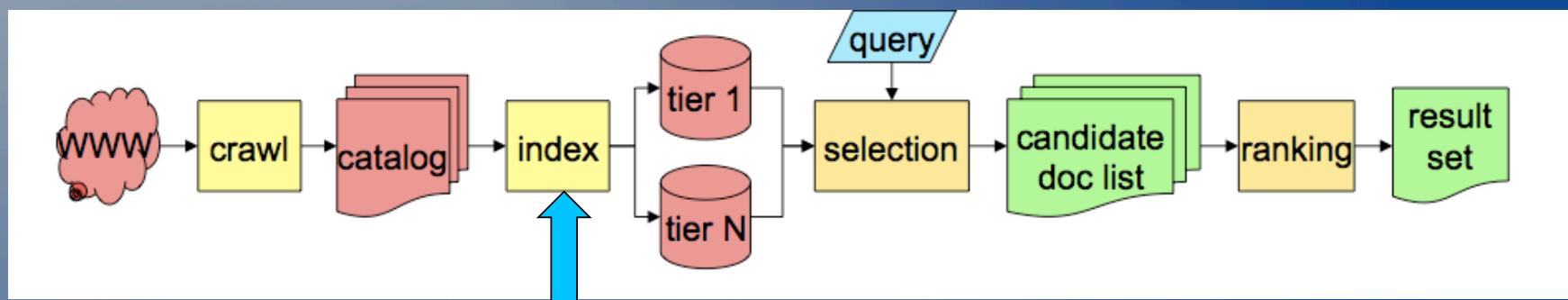
Pages can be built in non-HTML environments



Open source solutions exist: Heretrix, Nutch, UbiCrawler, etc..

Processing pages...

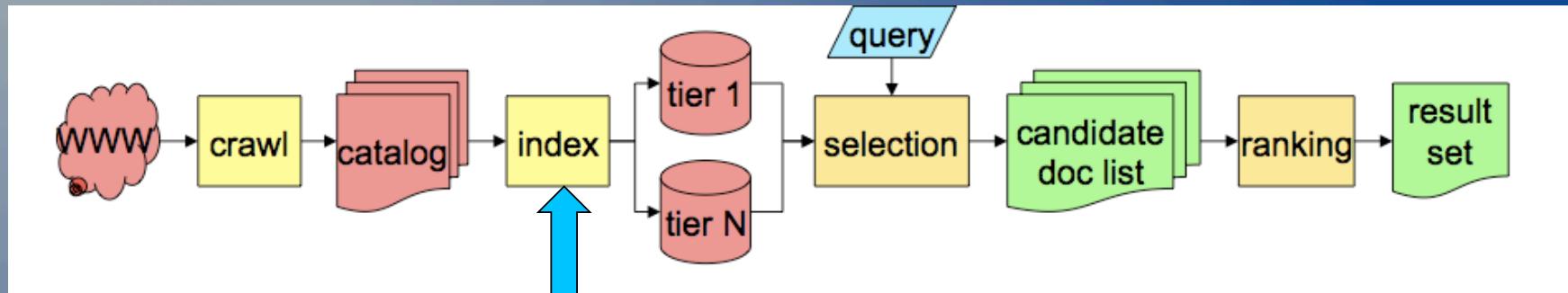
Second one needs to convert documents into index terms... to create an index.



“For indexing, you are doing a lot of I/O and a lot of thinking/analyzing the page, so the bigger (CPU) the better.” Patterson, 2004

To index or not to index...

Writing an indexer isn't easy...



You need to decide how much of the page you index and which 'features or signals' you care about.

Also which kinds of file to index? Text, sure.

Pdfs, jpegs, audio, torrents, videos, Flash???

Indexing...

There are hundreds of billions of web pages

It is neither practical nor desirable to index all:

- Should remove spam pages
- Should remove illegal pages
- Should remove malware
- Should remove repetitive or duplicate pages
- Should remove crawler traps
- Should remove pages that no longer exist
- Should remove pages that have substantially changed

Most search engines index in the range of 20 to 50 billion documents, says Williams

How many pages each engine indexes, and how many pages are on the web are hard research problems



Indexing: which are the right pages?

- pages that users want (duh...)
- pages that are popular in the web link graph
- pages that match queries
- pages from popular sites
- pages that are clicked on in search results
- pages shown by competitors
- pages in the language or market of the users
- pages that are distinct from other pages
- pages that change at a moderate rate
- the head is stable, The tail consists of billions of candidate pages with similar scores



Indexing...

Store (multiple copies of?) the web in a document store

Iterate over the document store to choose documents

Create an index, and ship it to the index serving nodes

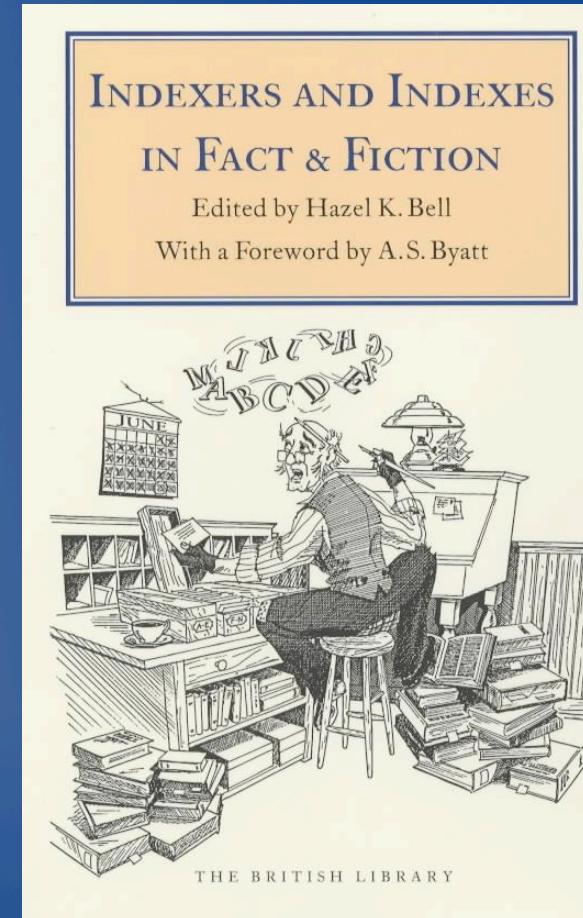
Repeat...

Sounds easy? It isn't!

[..] build an index. This is tricky. Just don't do anything wrong, as the saying goes. One false step and those billions of pages are going to take too long to process...

Patterson, 2004

Three words: scale, parallelism, time



Indexing: features

How to index pages so that we can find them?

A **feature (signal)** is an attribute of a document that a computer can detect, and that we think may indicate relevance.

Some features are **obvious**, some are **secret sauce** and how to use them is definitely a trade **secret** for each engine.



Features: the obvious...

Term matching

The system should prefer documents that contain the query terms.

Web Images Videos Maps News Shopping Gmail more ▾

tree

About 364,000,000 results (0.29 seconds)

Search Advanced search

Images for tree - Report images



Tree - Wikipedia, the free encyclopedia  A tree is a perennial woody plant. It is most often defined as a woody plant that has many secondary branches supported clear of the ground on a single main ... en.wikipedia.org/wiki/Tree - Cached - Similar

Stanford Tree - Wikipedia, the free encyclopedia  The Stanford Tree is the unofficial mascot of Stanford University. Stanford's team name is "The Cardinal," referring to the vivid red color (not the common ... en.wikipedia.org/wiki/Stanford_Tree - Cached - Similar)

Show more results from en.wikipedia.org

Yoga Journal - Tree Pose  Yoga article: Vrksasana clarifies just how challenging it can be to stand on one leg. www.yogajournal.com/poses/496 - Cached - Similar

Trees at arborday.org.  This official site of the Arbor Day Foundation provides information about planting and caring for trees, our Rain Forest Rescue and Tree City USA programs, ... www.arborday.org/ - Cached - Similar

Sponsored links

tree
Order now for Fall Shipping.
Free Shipping on Orders over \$3
www.arborday.org

Buy Bonsai Trees Online
Selling Bonsai & Supplies since
Free Guide & FAQ on Bonsai -a
www.bonsaiboy.com

Save On Trees
Buy Shade & Ornamental Trees
at Michigan Bulb - Free \$20
www.MichiganBulb.com

See your ad here »

Features: the obvious...

Term frequency

The system should prefer documents that contain the query terms many times.

Article Discussion Read View source View history Search

Tree

From Wikipedia, the free encyclopedia

For other uses, see [Tree \(disambiguation\)](#).

A tree is a [perennial woody plant](#). It is most often defined as a woody plant that has many secondary branches supported clear of the ground on a single main stem or [trunk](#) with clear [apical dominance](#).^[1] A minimum height specification at maturity is cited by some authors, varying from 3 m^[2] to 6 m;^[3] some authors set a minimum of 10 cm trunk diameter (30 cm girth).^[4] Woody plants that do not meet these definitions by having multiple stems and/or small size are called [shrubs](#). Compared with most other plants, trees are long-lived, some reaching several thousand years old and growing to up to 115 m (379 ft) high.^[5]

Trees are an important component of the [natural landscape](#) because of their prevention of [erosion](#) and the provision of a weather-sheltered [ecosystem](#) in and under their [foliage](#). They also play an important role in producing [oxygen](#) and reducing [carbon dioxide](#) in the atmosphere, as well as moderating ground temperatures. They are also elements in [landscaping](#) and [agriculture](#), both for their [aesthetic](#) appeal and their [orchard](#) crops (such as [apples](#)). [Wood](#) from trees is a [building material](#), as well as a primary energy source in many developing countries. Trees also play a role in many of the world's [mythologies](#) (see [trees in mythology](#)).^[6]

[Contents \[hide\]](#)
1 Classification



Trees on a mountain in northern Canada during early autumn.



tree.com

Home Our Businesses Management

WHERE SMART DECISIONS START

Whether you're buying a home, making a career change, or getting the right insurance, we provide you with everything you need to make smart decisions.

About Tree.com

Making decisions can be tough. Whether you're buying a house, switching careers, or getting the right insurance, there are literally thousands of opinions and answers online. So how do you get to the good stuff? Tree.com helps you make smart decisions in vital areas of your life.

Features: the obvious... Inverse document frequency

Rare words are more important than frequent words

Article Discussion Read Edit View history Search

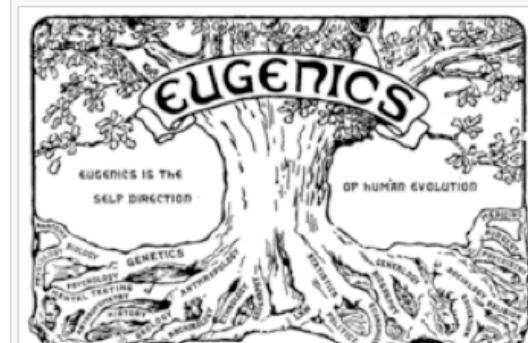
Eugenics

From Wikipedia, the free encyclopedia

Eugenics is the study and practice of [selective breeding](#) applied to humans, with the aim of improving the species. In a historical and broader sense, eugenics can also be a study of "improving human genetic qualities." Eugenics was widely popular in the early decades of the 20th century, but has fallen into disrepute after having become associated with [Nazi Germany](#). Since the postwar period, both the public and the scientific communities have associated eugenics with [Nazi](#) abuses, such as enforced [racial hygiene](#), [human experimentation](#), and the [extermination](#) of "undesired" population groups. However, developments in [genetic](#), [genomic](#), and [reproductive technologies](#) at the end of the 20th century have raised many new questions and concerns about the meaning of *eugenics* and its ethical and moral status in the modern era.

Contents [hide]

- 1 Overview
- 2 Meanings and types
 - 2.1 Implementation methods
- 3 Notable proponents
- 4 History



"Eugenics is the self-direction of human evolution": [Logo from the Second International Eugenics Conference, 1921, depicting it as a tree which unites a variety of different fields.](#) [1]

Indexing: More features

Term proximity

Words that are close together in the query should be close together in relevant documents.

the WHITE HOUSE PRESIDENT BARACK OBAMA

Get Email Updates | Contact Us

BLOG PHOTOS & VIDEO BRIEFING ROOM ISSUES the ADMINISTRATION the WHITE HOUSE our GOVERNMENT

Your
WEEKLY ADDRESS

On the 75th anniversary of Social Security, President Obama promises to protect it from Republican leaders in Congress who have made privatization a key part of their agenda.

Watch the Video

1 2 3 4



White House Photo, Lawrence Jackson, 8/13/10

FEATURED TOPICS

SEARCH the SITE

Indexing: more features

Term location

Prefer documents that contain query words in the title or headings.

The screenshot shows a web browser window with the following details:

- Title Bar:** Valeria de Paiva
- Address Bar:** http://www.cs.bham.ac.uk/~vdp/
- Toolbar:** Standard browser controls (Back, Forward, Stop, Refresh, Home, etc.)
- Bookmark Bar:** Most Visited, QPS-AT-B, C Workqueue mining, https://www.retirem..., Cuill Trac - Trac, Bourbaki / FrontPage, D workqueue staging, weight watchers - C..., WorkqueueB, WorkqueueBY, Logic and Rational...
- Search Bar:** viewzi
- Content Area:**
 - Header:** MAKE POVERTY HISTORY
TRADE JUSTICE. DROP THE DEBT. MORE & BETTER AID.
 - Section:** Valeria de Paiva
 - Image:** A photograph of Valeria de Paiva, a woman with dark hair wearing a black jacket and a colorful scarf, holding a glass.
 - Text:** (picture by Thomas Forster, 2004)
 - Text:** Valeria de Paiva is a search analyst working for Cuill, Inc. in Menlo Park, CA. Cuill is a new search engine with tons of innovative features, check it [out](#) (or hear Anna Patterson's [presentation](#)).
 - Text:** She was until May 2008 a research scientist at the Intelligent Systems Laboratory of PARC (Palo Alto Research Center), California.
 - Text:** She received her PhD in Mathematics from Cambridge University in 1988 for work on "Dialectica Categories".
- Bottom Bar:** Find: tax number, Next, Previous, Highlight all, Match case, Reached end of page, continued from top, Done

Indexing: web features

URL text

Prefer documents that contain query words in the URL.

www.whitehouse.gov

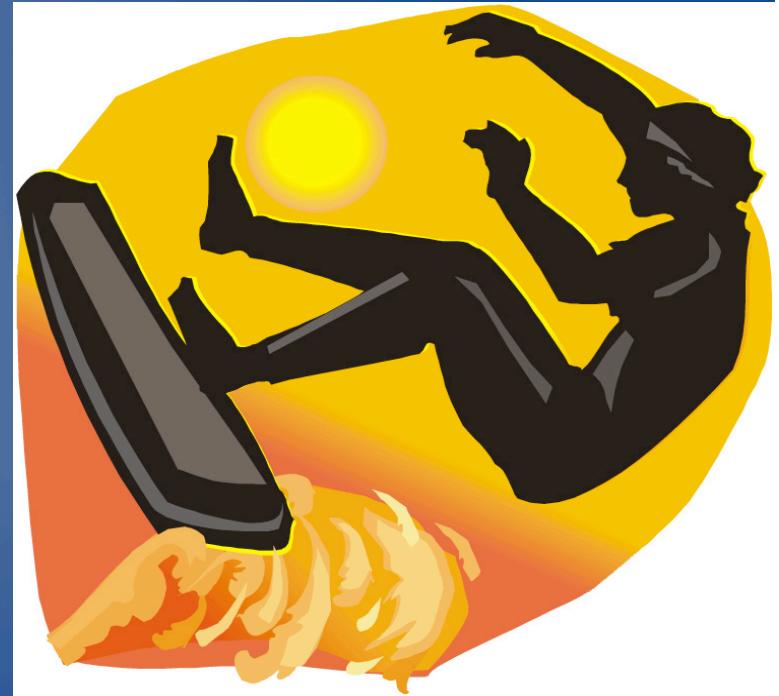
www.linkedin.com/in/valeriadepaiva

Indexing: which features?

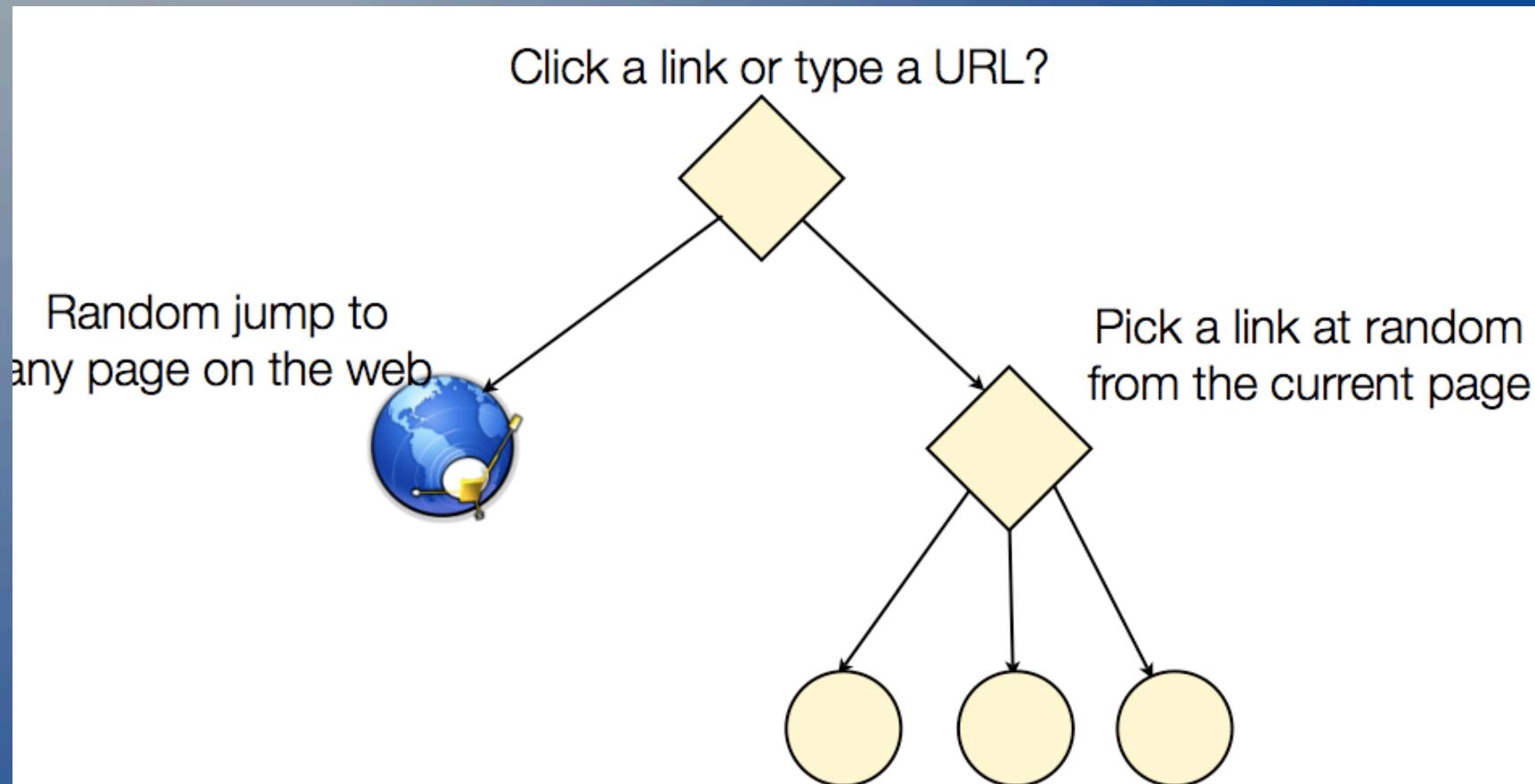
Prefer documents that are
authoritative and popular.

HOW?

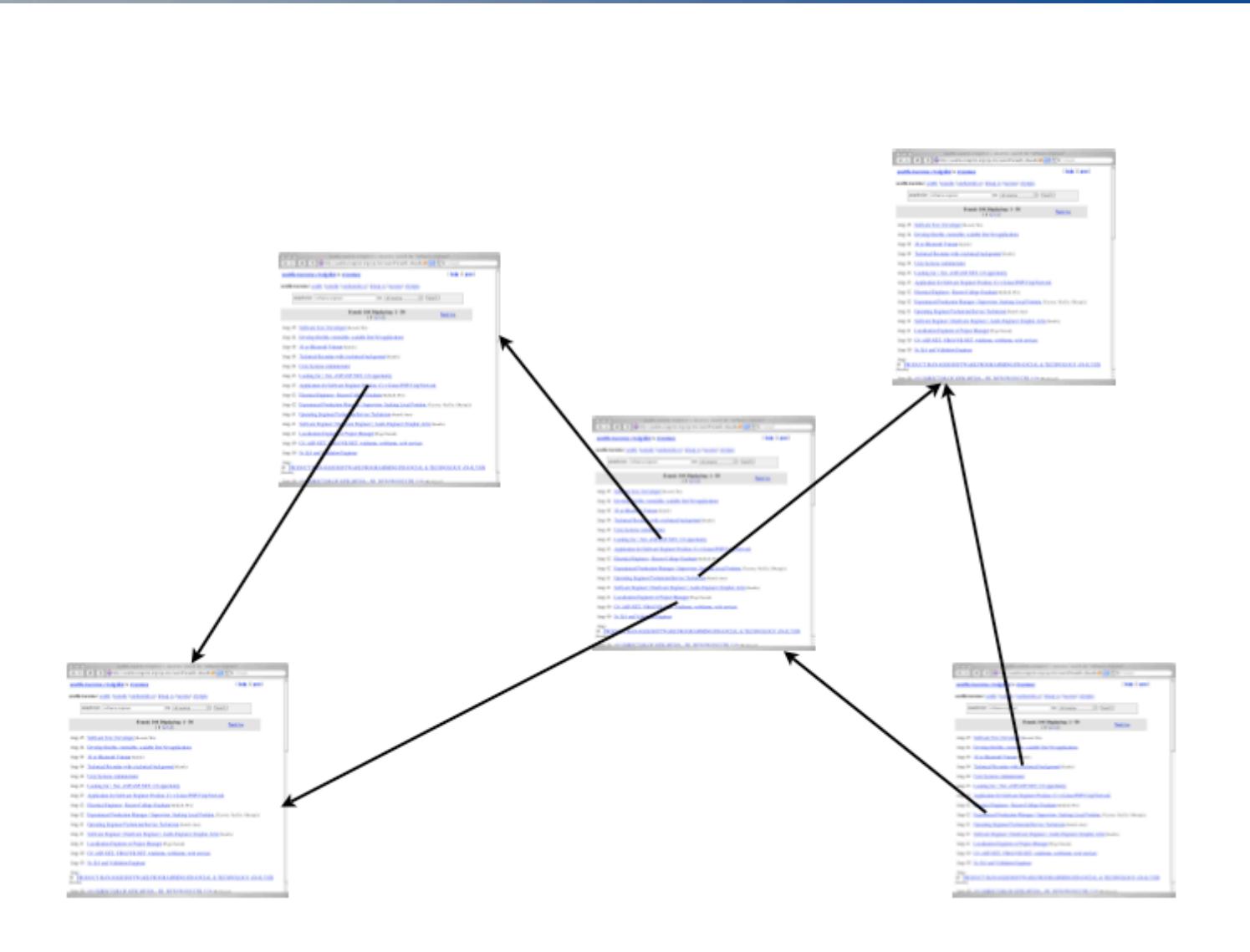
The **random surfer** is a
hypothetical user that clicks on
web links at random.



Random surfer?



PageRank: Popular Pages connect



PageRank: how can we leverage the connections?

Think of a gigantic graph and its transition matrix

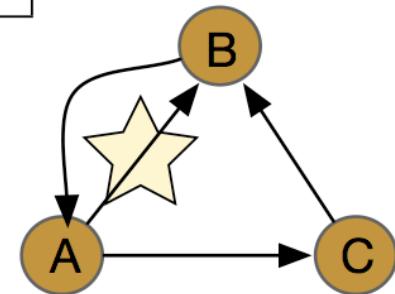
Make the matrix probabilistic

Those correspond to the random surfer choices

Find its principal eigen-vector= pagerank

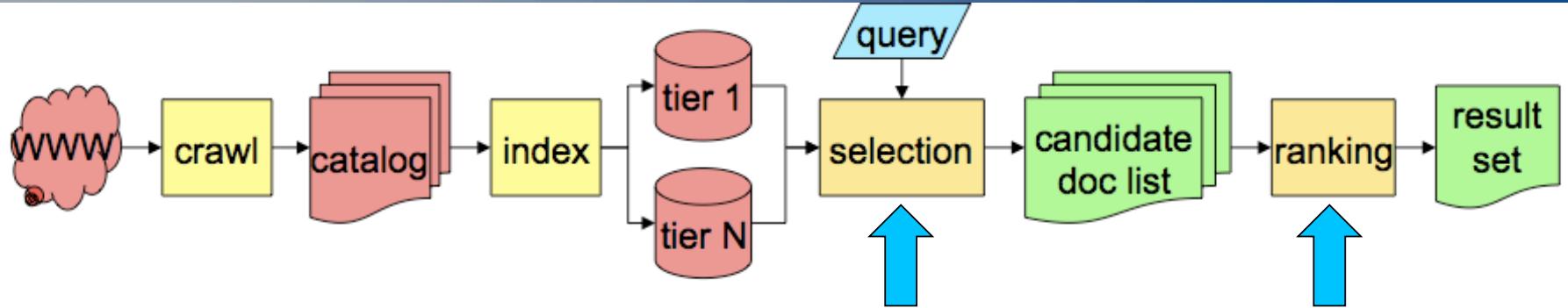
Graphs: Matrix Representation

	A	B	C
A	0	★1	1
B	1	0	0
C	0	1	0



PageRank is the proportion of time that a random surfer would jump to a particular page.

Selection and Ranking



Whatever features you will want for RANKING and SELECTION you should consider when indexing...

Patterson's advice “The hallowed ‘index format’ is not the end of the search engine, it is just the beginning. It is a tool to see results, so change it and change it often.” ACM Q, 2004

Quality: what do we want to do?

- Measure user satisfaction
- Optimize for user satisfaction in each component of the pipeline
- Automate all metrics
- Discover anomalies
- Visualize, mine, and summarize metrics data
- Debug problems automatically
- Compare different search engines?



User satisfaction?

- User issues a query to Search Engine: Receives a list of results
- How **relevant** are these results? How do we measure relevance of results? How effectively was the information need of the user met?
- How many of the results retrieved were useful?
- Were there useful pages not retrieved?



Evaluating Relevance...

- is HARD!
- Evaluation is key to *effective and efficient* search engines
- Effectiveness, efficiency and cost are related - difficult trade-off
 - Two main kinds of approach:
IR traditional evaluations,
click data evaluation & log
analysis
- Many books on IR, many
patents/trade secrets from
search engines...



Traditional IR evaluation

- TREC competitions (NIST and U.S. Department of Defense), since 1992
- Goal: provide the infrastructure necessary for large-scale evaluation of text retrieval methodologies
- several tracks, including a Web track, using ClueWeb09, one billion webpages
- TREC results are baseline, do they work for search engines?



Evaluating Relevance in IR...

if universe small you can check easily precision and recall

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})}$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})}$$

- **F-measure: Harmonic mean of precision and recall**
 - related to van Rijsbergen's effectiveness measure
 - reflects user's willingness to trade precision for recall controlled by a parameter selected by the system designer

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \alpha = \frac{1}{(\beta^2 + 1)}$$

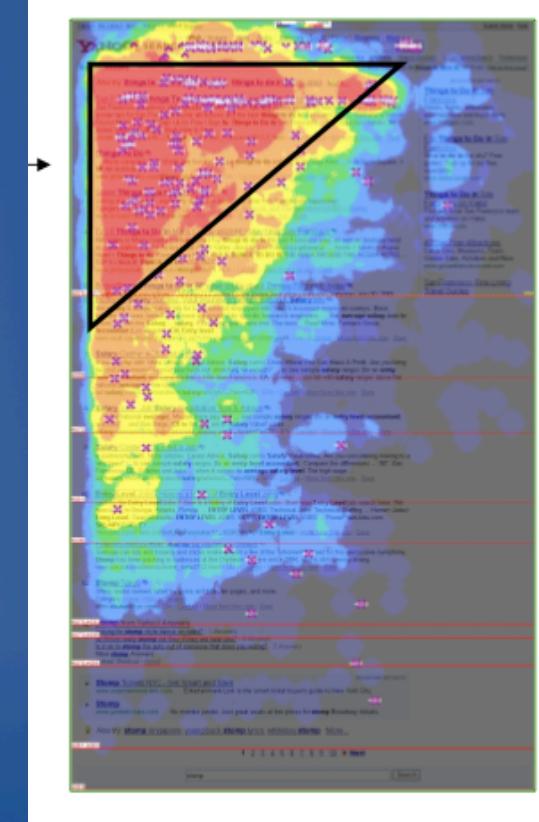
Evaluating Relevance in IR...

Many other measures in literature, e.g precision-recall curves, MAP (mean of average precision over all queries), etc..

For search engines the **first results** are much more important than later ones

A natural alternative is to report at top 5, etc

Instead of binary judgments (relevant or not) graded: very relevant, somewhat relevant, not relevant...



Evaluating Relevance: DCG

Very used measure: DCG
(discounted cumulative gain)

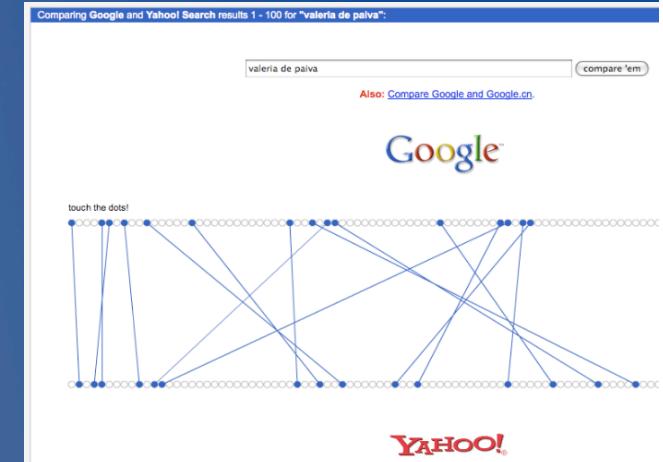
Very good=3, somewhat good=1
and not good=0 penalize good
results in bad positions by
using the discounted measure.
Discounts geometric or log..

If url1=3, url2=1 and url3=0 then

$$\text{DCG} = \frac{3}{1} + \frac{1}{2} + \frac{0}{4} = 3.5$$

same urls in opposite order

$$\text{DCG} = \frac{0}{1} + \frac{1}{2} + \frac{3}{4} = 1.25$$



Evaluating Relevance in IR

Other measures:

Kendall tau coefficient (counts
of preferences..)

Bpref (Buckley and Voorhees,
2004)

Rpref (De Beer and Moens,
2006 graded bpref)

Q-measure (Sakai, 2007)

....



Information Retrieval

Relevance

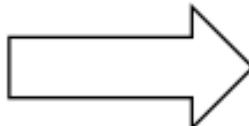
-Effective ranking

Evaluation

-Testing and measuring

Information needs

-User interaction



Search Engines

Performance

-Efficient search and indexing

Incorporating new data

-Coverage and freshness

Scalability

-Growing with data and users

Adaptability

-Tuning for applications

Specific problems

-e.g. Spam

Evaluating User Satisfaction...

Salient aspects of user satisfaction are hard for IR metrics

Relevance metrics not based on users experiences or tasks

Some attempts:

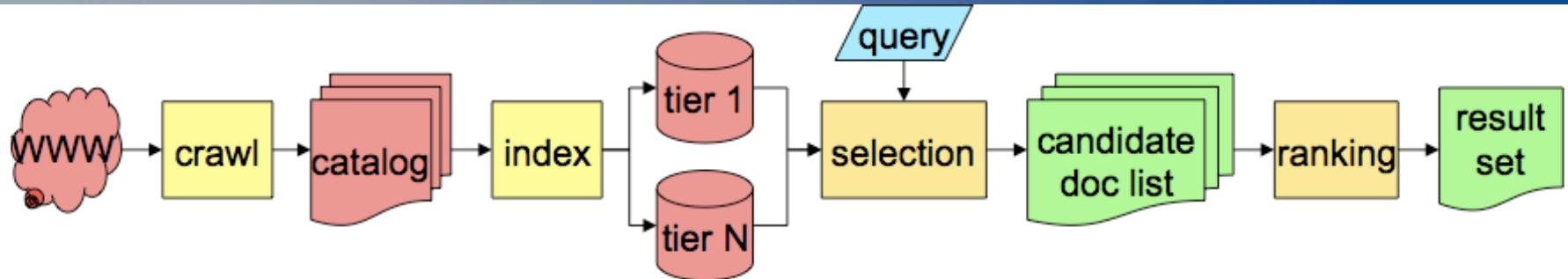
Huffman and Hochster (2007)

SIGIR 2009 Workshop on The Future of IR Evaluation (2009)

Markov models of user frustration and satisfaction (Hassan et al 2010)



Evaluating the whole system...



Coverage metrics

Query by URL

Query by content (strong queries)

Compute coverage ratio

Issues: URL normalization/ Page template

Problems:

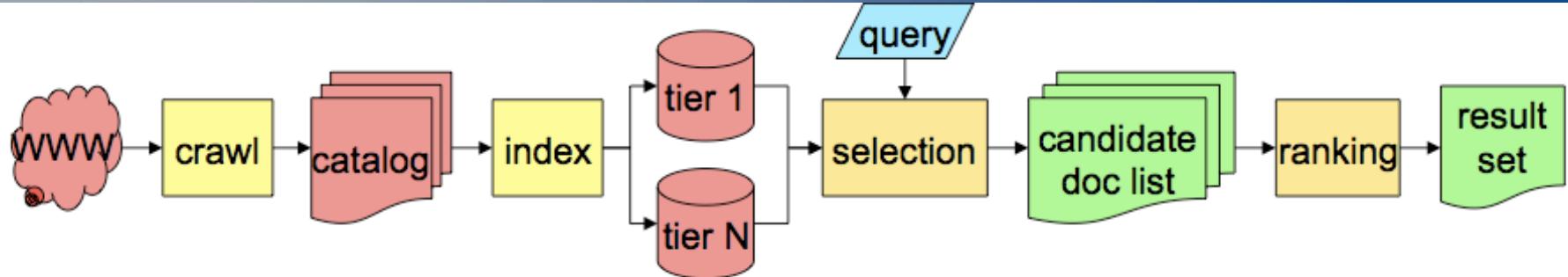
Web growth in general and in dimensions???

Improve copy detection methods quality and performance

Bar-Yossef and Gurevich: Efficient Search Engine Measurements WWW2007

Olston and Najork: Web Crawling, 2010

Evaluating the whole system...



Diversity metrics

Some queries have a single news result, reported by many

Some queries are ambiguous e.g [Stanford]:

Do you want to optimize for the dominant concept or all facets: [jaguar] ?

Implications for interfaces/visualization...

Exclusion of near-duplicates

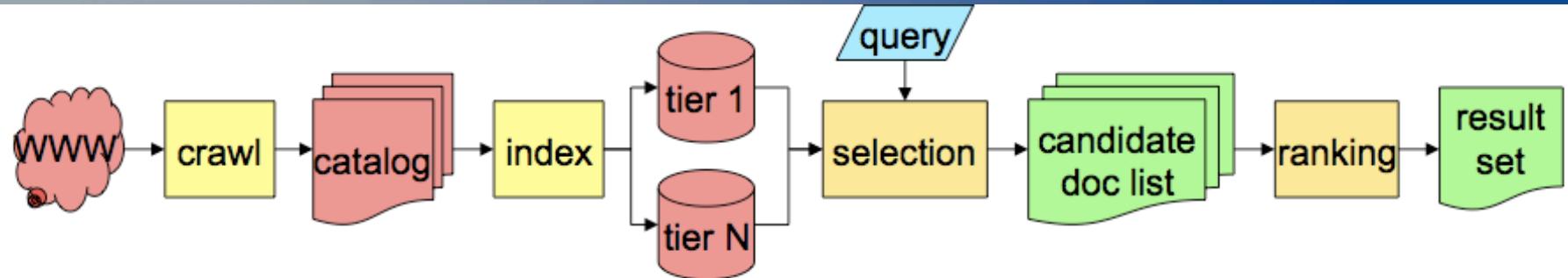
Problems: measure and summarize diversity

Measure tradeoffs between diversity and relevance [beata keller]

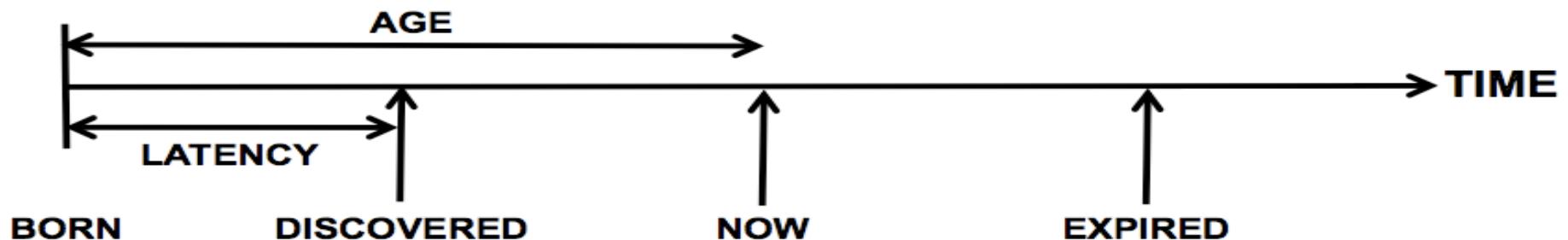
Best presentation of diversity?

Agrawal et al 2009

Evaluating the whole system...

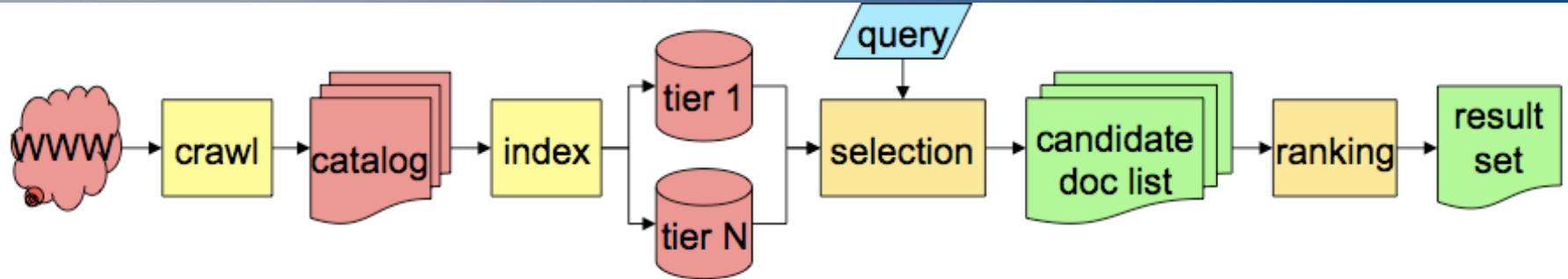


Latency and discovery metrics



Given a set of new pages, how long does it take for them to be served? How many of them become part of the results?

Evaluating the whole system...

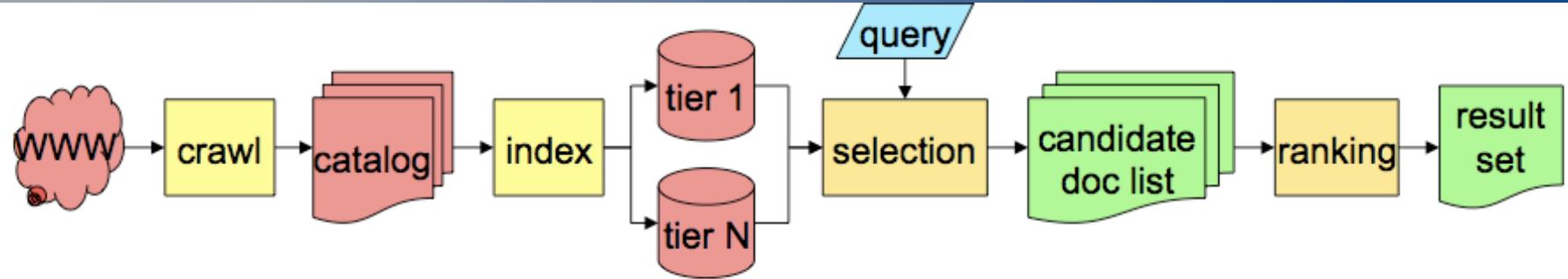


Freshness metrics

Freshness of snippets?

Measuring change of all internet?

Evaluating the system: presentation metrics



User studies. Eye-tracking studies?

Presentation modules: suggestions, spelling corrections, snippets, tabs, categories, definitions, images, videos, timelines, maplines, streaming results, social results, ...

How to measure success? How to optimize relevance? How much is too much? International differences? How to change interfaces?

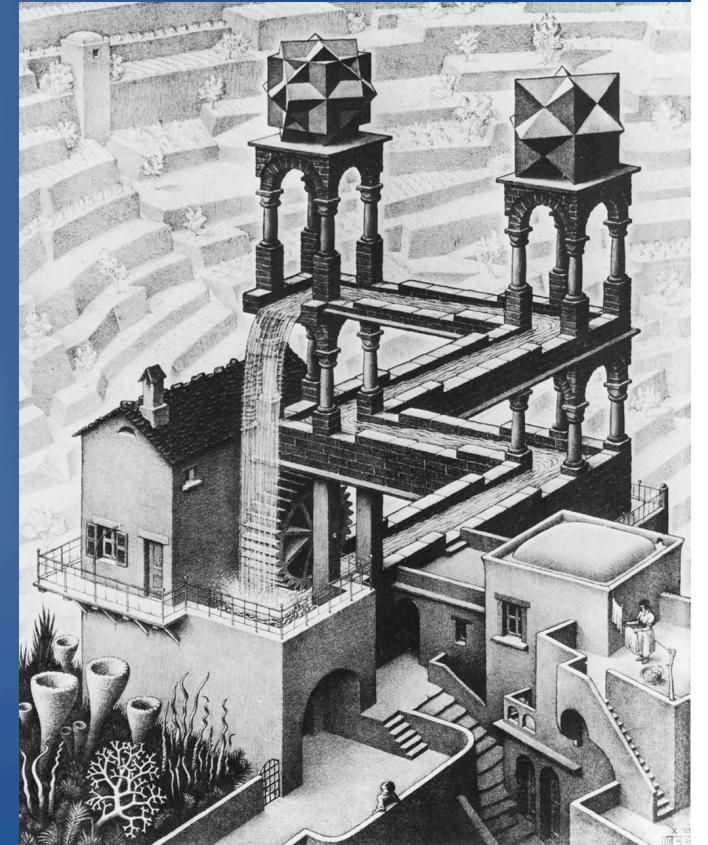
Conclusions

- Relevance is elusive, but essential.
- Improvement requires metrics and analysis, continuously
- Gone over a rough map of the issues and some proposed solutions
- Many thanks to Dasdan, Tsoutsouliklis and Velipasaoglu, Croft, Metzler and Strohman, (especially Strohman!) and Hugh Williams for slides/pictures.



Coda

- There are many search engines. Their results tend to be very similar. (how similar?)
- Are we seeing everything? Reports estimate we can see only 15% of the existing web.
- Probing the web is mostly **popularity based**. You're likely to see what others have seen before. But your seeing increases the popularity of what you saw, thereby reducing the pool of available stuff. Vicious or virtuous circle? How to measure?



Summing up

- Life in Searchland is very different. And lots of fun!
- “[...] once the search bug gets you, you'll be back. The problem isn't getting any easier, and it needs all the experience **anyone** can muster.” Patterson, 2004



Thank You!

References

Croft, Metzler and Strohman's "Search Engines: Information Retrieval in Practice", 2009, Addison-Wesley

the tutorial '**Web Search Engine Metrics**' by Dasdan, Tsoutsoulikis and Velipasaoglu for WWW09/10, available from <http://www2010.org/www/program/tutorials/>

Hugh Williams slides for ACM Data Mining May 2010

Anna Patterson:

Why Writing Your Own Search Engine Is Hard

ACM Q, 2004

A white search bar with a thin black border.

Search

Find 384,165,027 automated articles

[About Cpedia](#) | [Preferences](#) | [Add Cpedia to Firefox](#)

[Privacy Policy](#) | © 2010 Cuil, Inc.

Evaluating Relevance in IR: implicit judgments

Explicit judgments are expensive

A search engine has lots of user interaction data, like which results were viewed for a query, which received clicks..

How can we use this data?

Basic statistics from raw observations:
abandonment rate, reformulation rate,
number of queries per session, clicks per
query, mean reciprocal rank of clicked
results, time to first (last) click, etc....
(Joachims 2002, etc)

Evaluating Relevance in IR: direct and indirect clicks

Direct and indirect evaluation by clicks

(Radlinski and Joachims, Carterette and
Jones, 2007)

Model based evaluation (Dupret et al, 2007,
Dupret (2009)

ERR(Expected Reciprocal Rank) unlike DCG has non-zero probability that user stops browsing, better correlation with click metrics, etc...

So far evaluation method rather than USER satisfaction