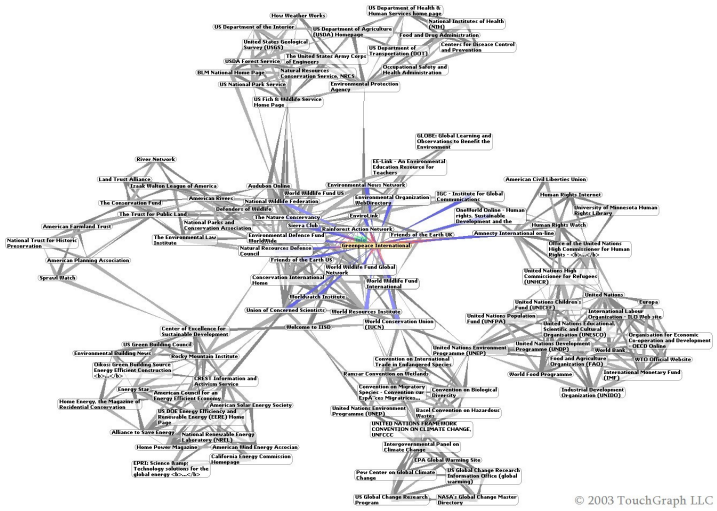
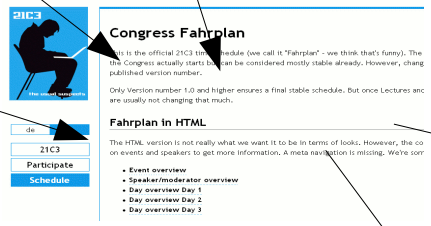
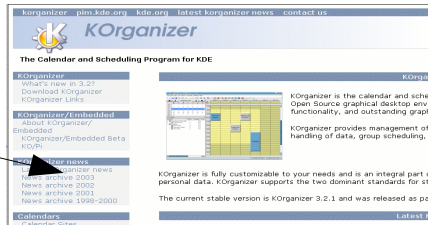
Developing Intelligent Search Engines

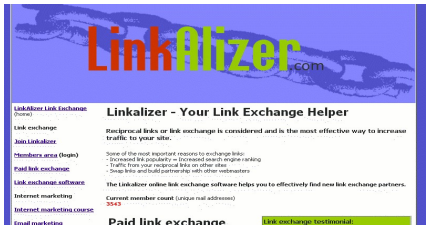
Isabel 'MaineC.' Drost

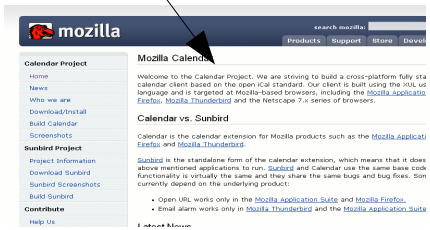


The world wide web





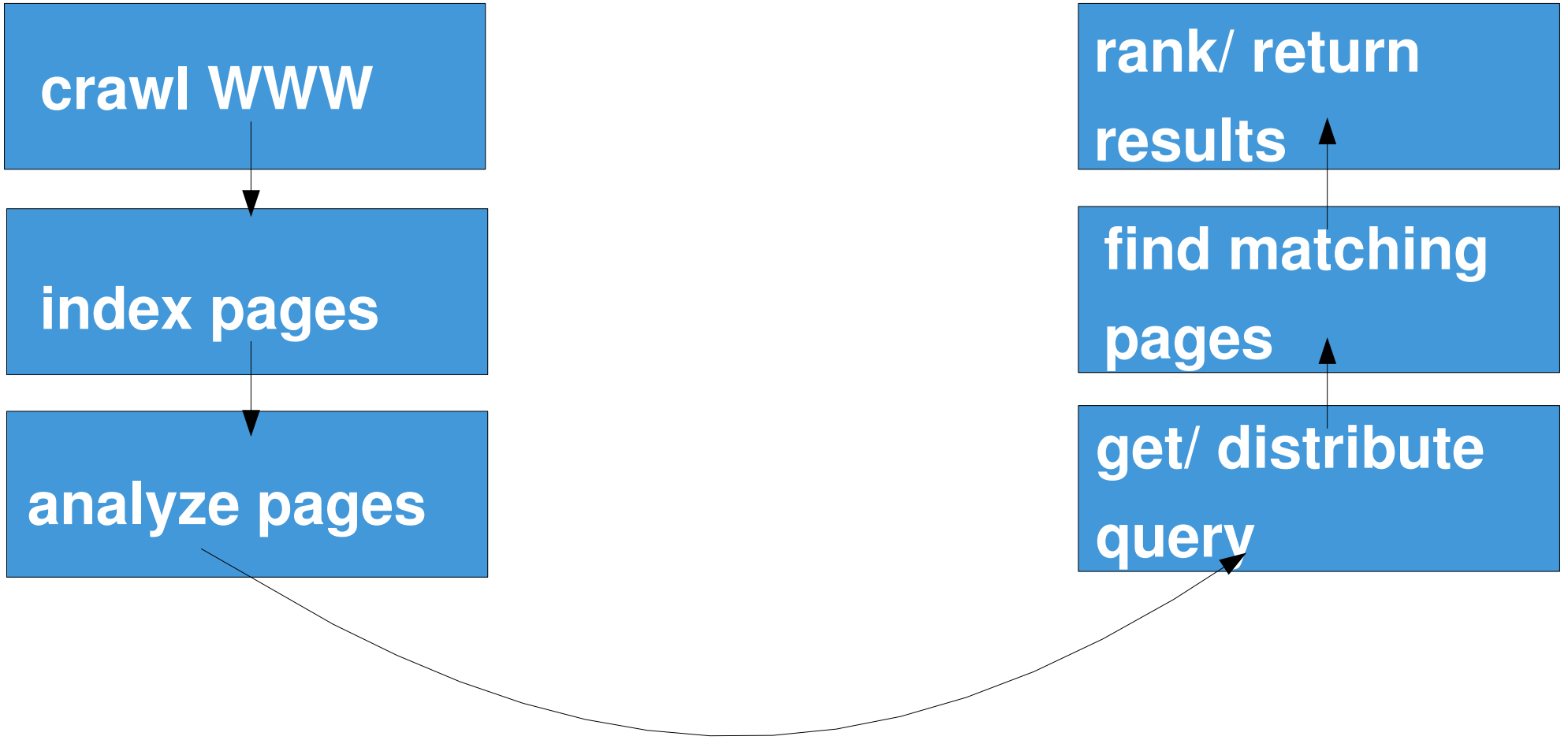




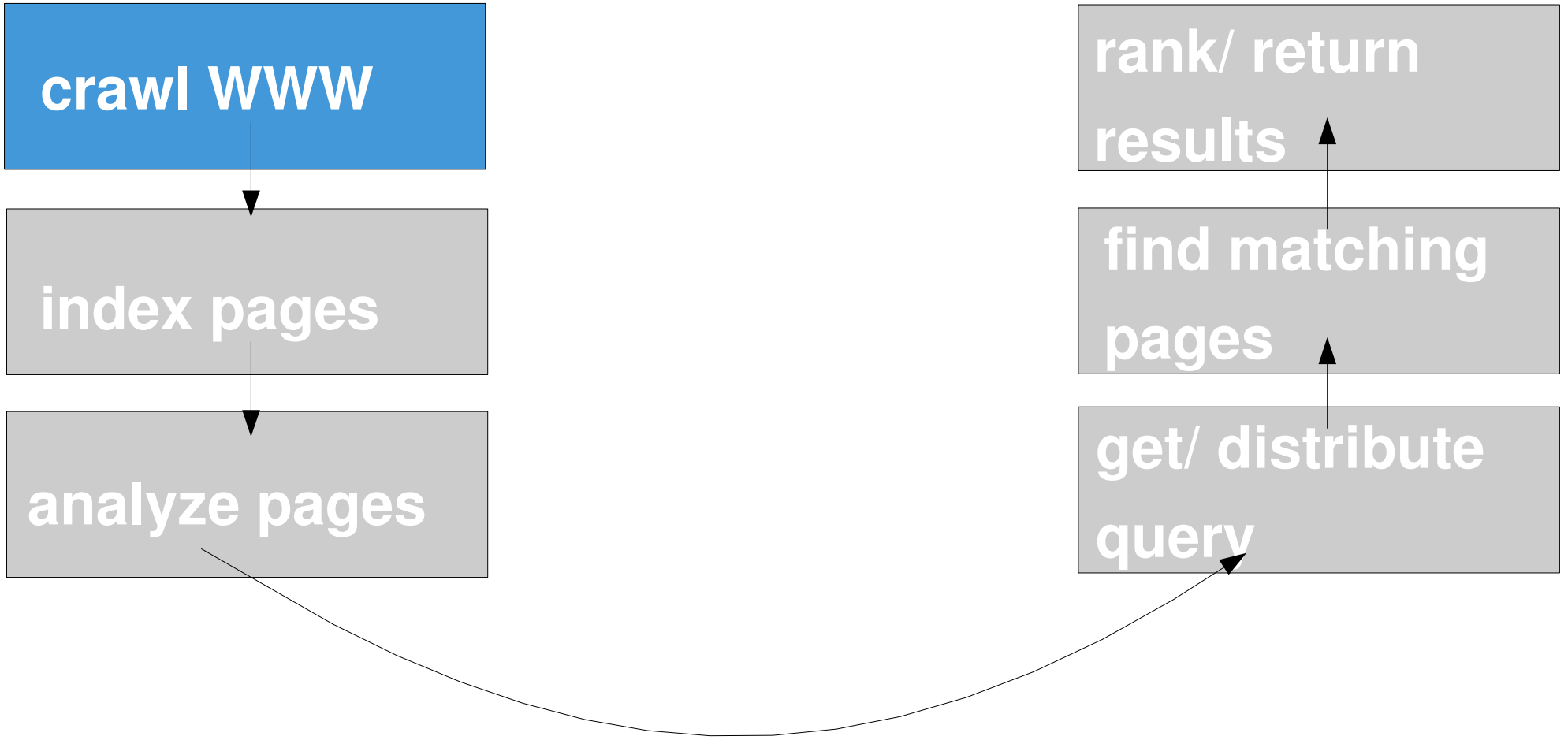
Agenda

- Basic architecture of search engines.
- Properties of the WWW
- One machine learning task in search engines.
- How can YOU contribute?

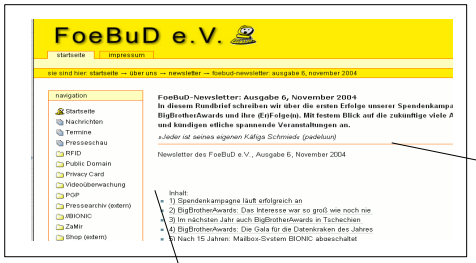
Search engines from scratch

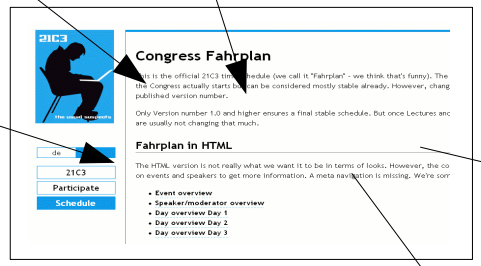


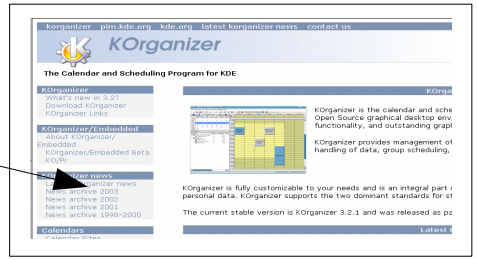
Search engines from scratch

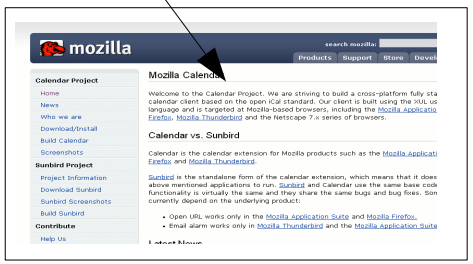


Crawling the WWW





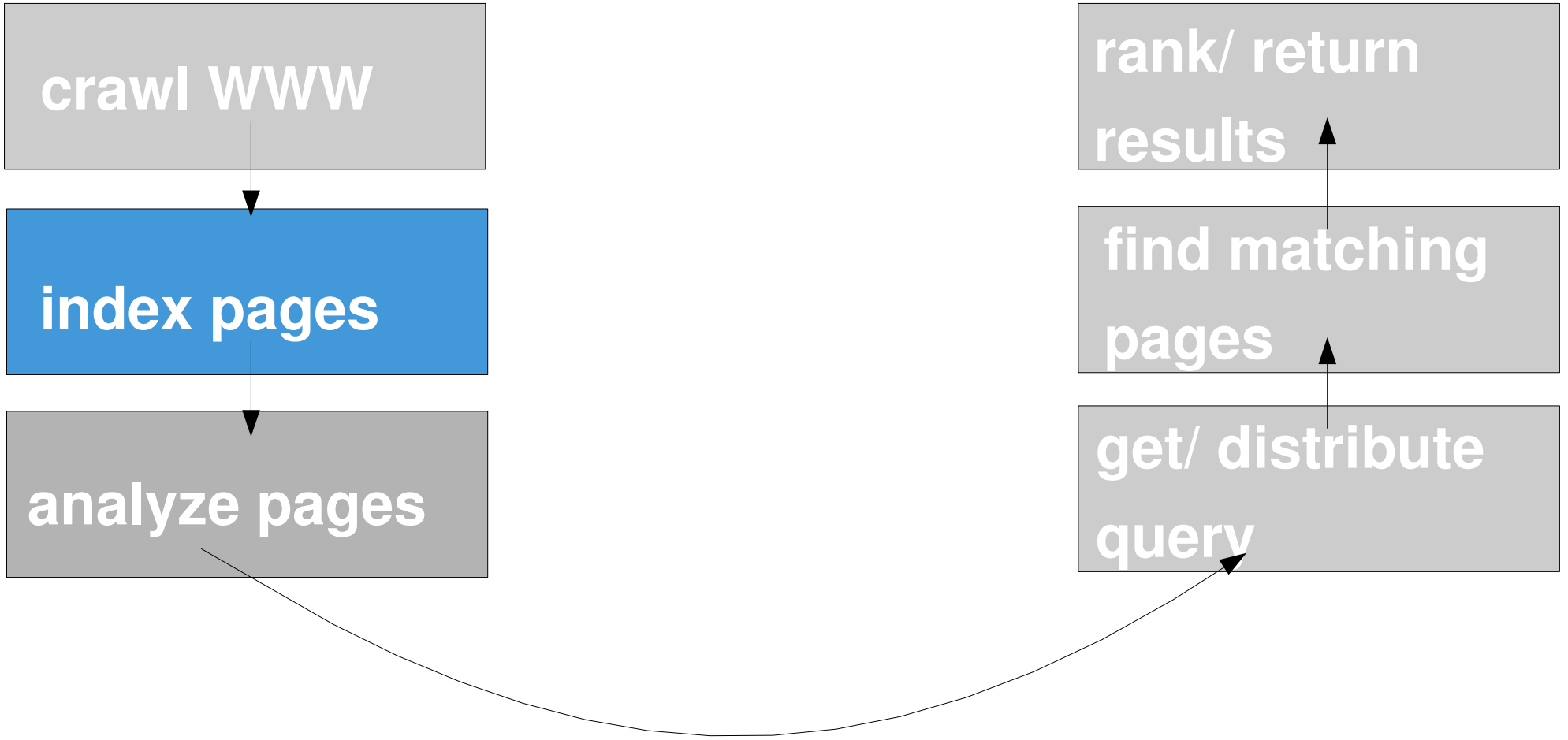




Crawling the WWW

- Choose starting page (e.g. DMOZ).
- Get and save page content.
- Extract and follow links recursively.
  - Parallelize crawling.
  - Prioritize links to follow.
  - Avoid gathering too much content at once – DOS.

Search engines from scratch



Indexing web pages

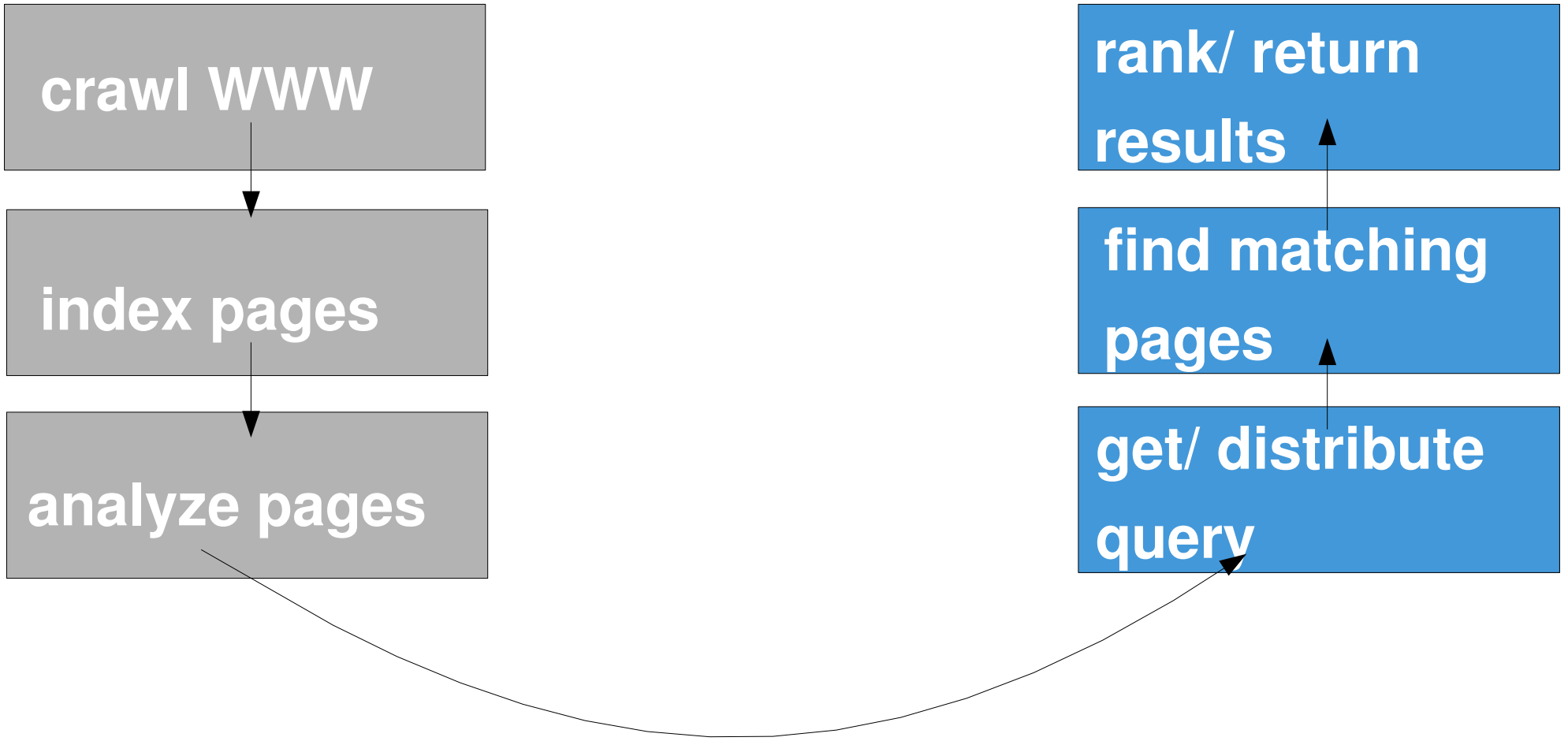
- Parse web pages.
  - Extract keywords.
  - Extract keywords from anchor texts of inlinks.
- Index content of pages.
  - Each keyword part of a very long list.
  - URLs of matching pages associated with list entries.

Indexing web pages

This year the 22C3 will take place in Berlin.
The 27th Chaos Communication Congress will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.
This year the 22C3 will take place in Berlin.

2005	http://www.ccc.de/congress/2005
22C3	http://www.fukami.de http://www.ccc.de/congress/2005
27th	http://www.ccc.de/congress/2005
at	http://www.ccc.de/congress/2005
Berlin	http://www.berlin.de http://www.ccc.de/congress/2005
december	http://www.ccc.de/congress/2005
in	http://www.ccc.de/congress/2005
place	http://www.ccc.de/congress/2005
take	http://www.ccc.de/congress/2005
the	http://www.ccc.de/congress/2005
will	http://www.ccc.de/congress/2005

Search engines from scratch



Handling user queries

- Retrieve queries from user.
  - Parallelize query retrieval.
- Retrieve matching pages from index.
- Rank found pages.
- Create easy to understand representation.

Handling user queries

22C3 | go

retrieve

2005 http://www.ccc.de/congress/2005
22C3 http://www.fukami.de http://www.ccc.de/congress/2005
27th http://www.ccc.de/congress/2005
at http://www.ccc.de/congress/2005

rank

http://www.ccc.de/congress/2005
http://www.fukami.de

generate webpage

Handling user queries

22C3: Private Investigations
The 22nd Chaos Communication Congress (**22C3**) is a four-day conference on ...
More information on **22C3** is coming up. For now, study and distribute our Call ...
www.ccc.de/congress/2005/ - 3k - Cached - Similar pages

CCC | Chaos Communication Congress - [Translate this page]
22C3: Private Investigations. Der 22. Chaos Communication Congress findet vom 27. bis 30. ... All Informationen zum **22C3** finden sich unter [Interner Link] ...
www.ccc.de/congress/ - 13k - Cached - Similar pages
[More results from www.ccc.de]

Upcoming.org: **22C3**: Private Investigations at Berliner Congress ...
The 21nd Chaos Communication Congress (**22C3**) is a conference on technology, society and utopia ... Chaos Communication Congress **22C3**: Private Investigations ...
upcoming.org/event/23579/ - 9k - Cached - Similar pages

22c3 h07 - NerdPedia
Darum: META-REFRESH http://hackerhippie.org/newiki/**22c3**. [old content deleted...use diff feature of this wiki] ...
www.nerdpedia.org/index.php/**22c3**_h07 - 8k - Cached - Similar pages

22C3 (Private Investigations) - Call for Papers | Uwe Hermann
The first signs of the upcoming **22C3** congress ("Private Investigations") are ...
The 22nd Chaos Communication Congress (**22C3**) is a four-day conference on ...
www.hermann-uwe.de/blog/ **22c3**-private-investigations-call-for-papers - 44k - Cached - Similar pages

The Lunatic Fringe » Blog Archive » **22C3** Updates
22C3 Updates. Sunday October 23rd 2005, 18:41 Filed under: General. The last months have been quite exhausting in terms of getting people on track for the ...
tim.geekheim.de/2005/10/23/**22c3**-updates/ - 17k - Cached - Similar pages

22C3 (Plan 9 wiki)
22C3. Chaos Communication Congress Where: Berlin, Germany When: 27-30/12/2005 URL: http://www.ccc.de/congress/. Who will be there? 20h; MTG; garbeam; uriel ...
www.cs.bell-labs.com/wiki/plan9/22C3/ - 2k - Cached - Similar pages

maha's blog » **22c3**
22c3. Sunday, July 10th, 2005. Gestern fand die erste Sitzung der ... You are currently browsing the archives for the **22c3** category. ...
www.maha-online.de/blog/category/ccc/**22c3**/ - 14k - Cached - Similar pages

symlink.ch | **22C3**: Fnord-Jahresrückblick oder nicht? - [Translate this page]
'**22C3**: Fnord-Jahresrückblick oder nicht?' | Einloggen/Account erstellen | 9 Kommentar(e) ... Re: Bin gespannt auf **22c3** by Anonymer Feigling Wednesday 19. ...

Handling user queries in depth

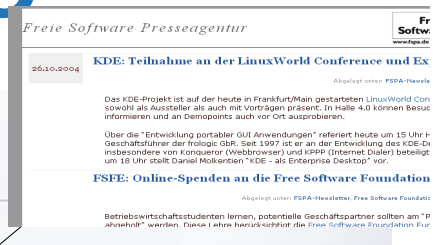
- What makes ranking difficult?

- Enhanced query handling.
  - Presentation of search results.
  - Enriching queries by further information.
  - Building search engines that answer questions.

Ranking search results

- Unknown what the user actually is looking for.
- Rank pages according to:
  - Text based relevance for query.
  - Overall importance.
- Measure for page importance: Page Rank.
- Problems when ranking search results:
  - Users tweaking pages to spam the ranking.
  - Evaluation of rankings.

Influencing the ranking of pages



http://naivewebpage.net

Search engines

Congress

Chaos Communication Congress and Search Engines

This is a rather content free text to be published on a web page. It deals with the chaos communication congress in general and specifically addresses the topic of the design of search engines.

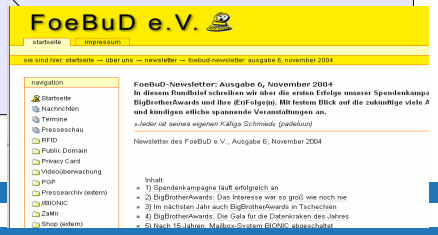
This is a rather content free text to be published on a web page. It deals with the **chaos communication congress** in general and specifically addresses the topic of the design of **search engines**.

Search engine chaos communication congress search engine chaos congress communicate search engine

Chaos Communication

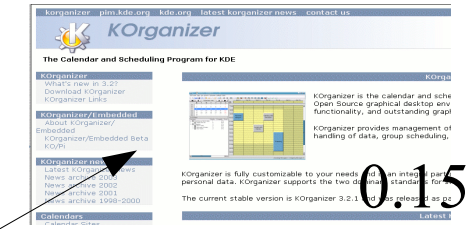
Search engines

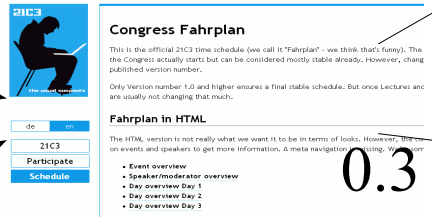
Search engines

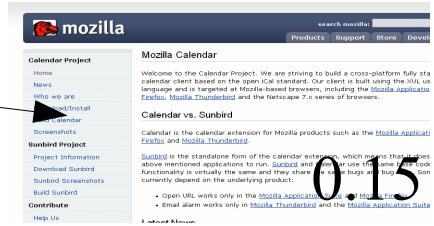


Ranking by page rank

- Page rank expresses popularity.
- Based on in-links.

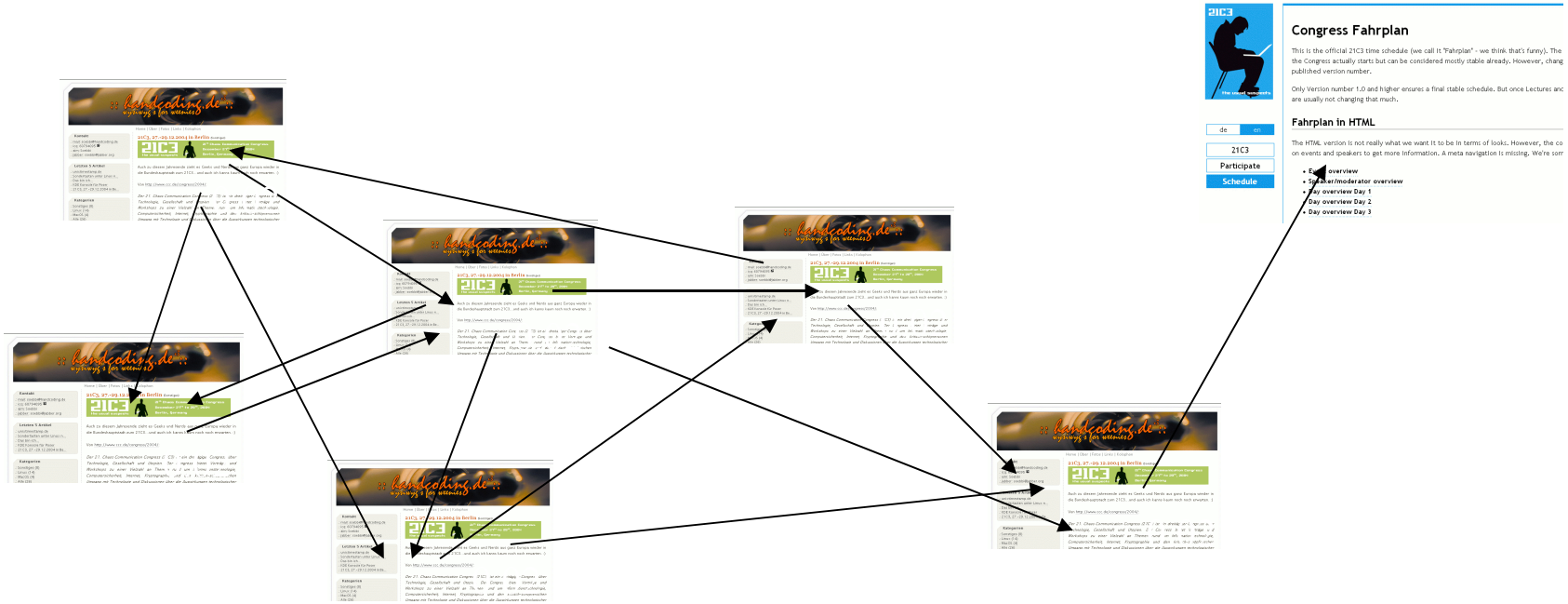






$$P_{x_i} = \frac{\alpha}{|X|} + (1-\alpha) \cdot \sum_{x_j; e_{j,i} \in E} \frac{P_{x_j}}{|x_j|}$$

HowTo inflate page rank?



A Google bomb

http://www.google.de/search?hs=Bot&hl=de&client=opera&rls=de&q=%22miserable+failure%22&btnG=Suche&meta=

Google

Web Bilder Groups Verzeichnis News Froogle[Neu!] **Mehr »**

"miserable failure" Suche Erweiterte Suche Einstellungen

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web Ergebnis

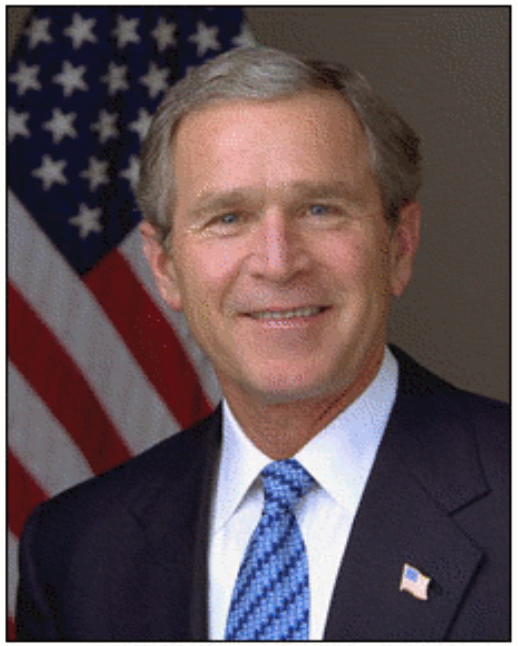
Tipp: Suchen nur nach Ergebnissen auf **Deutsch** . Sie können Ihre bevorzugten Spracheinstellungen in Einstellungen angeben.

Biography of President George W. Bush - [Diese Seite übersetzen]
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 33k - 15. Sept. 2005 - Im Cache - Ähnliche Seiten
Past Presidents - Kids Only - Current News - President
Weitere Ergebnisse von www.whitehouse.gov »

BBC NEWS | Americas | '**Miserable failure**' links to Bush - [Diese Seite übersetzen]
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - 14. Sept. 2005 - Im Cache - Ähnliche Seiten

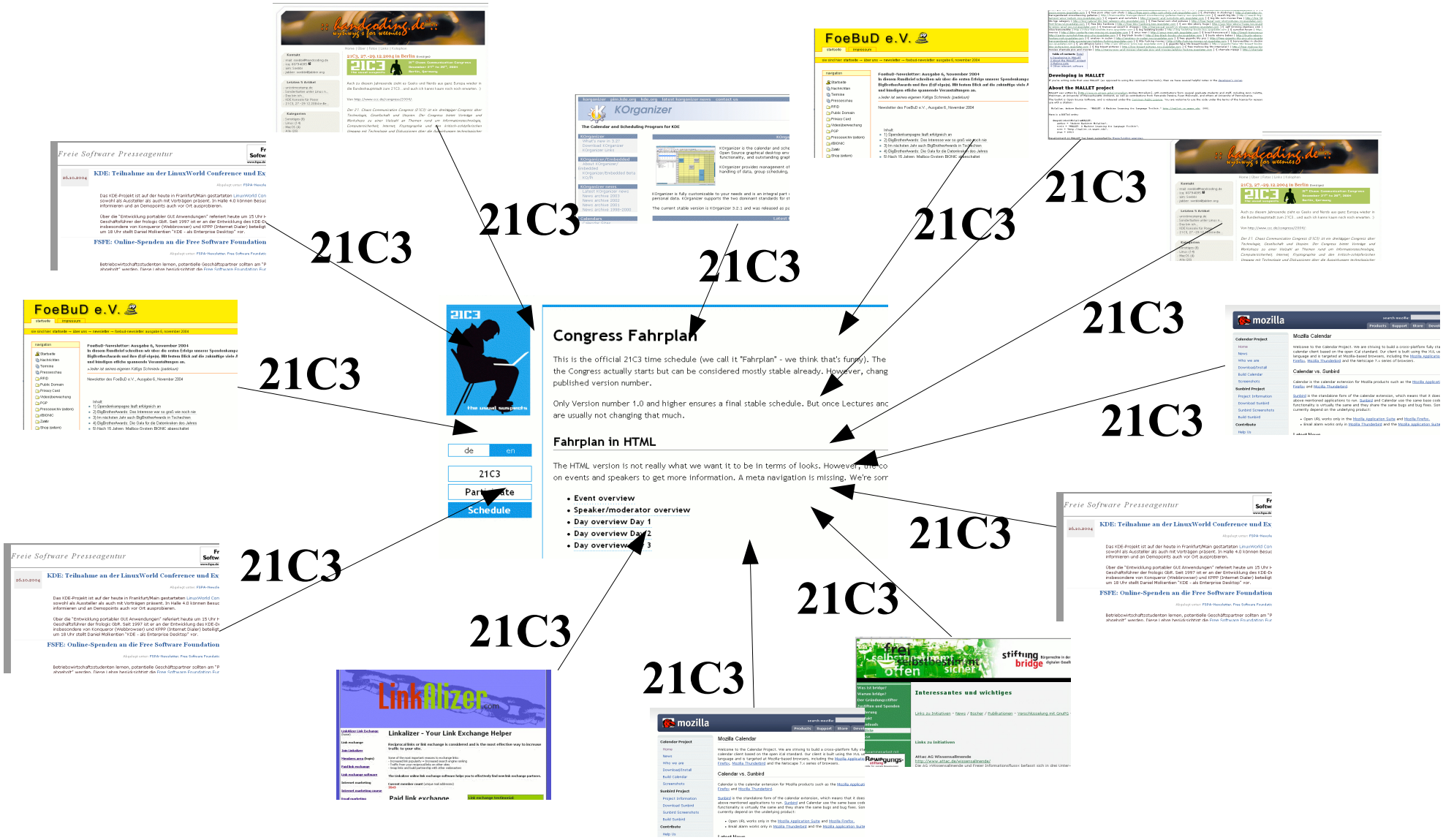
Welcome to MichaelMoore.com! - [Diese Seite übersetzen]
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 39k - 14. Sept. 2005 - Im Cache - Ähnliche Seiten

:: ringfahndung :: Google verbindet Bush mit '**Miserable failure** ...
:: ringfahndung :: ist ein netzine fuer freunde des schraegen humors, der satire, der kunst und des neuen digitalen stils ;). kurz :: wir schreiben hier all ...
www.ringfahndung.de/index.shtml - 11k - 14. Sept. 2005 - Im Cache - Ähnliche Seiten



White House photo by Eric Draper

How was the bomb created?



What is link spam?

Alternative **Medicine** Links - Add Your URL
Alternative **Medicine** Links section of the health directory at PillWatch.com. **... Order** All Heralife Weight Loss products here. www.katsherbs.com/Herbalife **...**
www.pillwatch.com/directory/alternative-**medicine**/ - 65k - Supplemental Result - Cached - Similar pages

Express-Scripts.com
Track Your **Order** · Get Helpful Answers. Get details about your coverage, **...** Follow our 7 Steps to Safety to avoid common **medicine** mistakes. **...**
www.express-scripts.com/ - 23k - 24 Sep 2005 - Cached - Similar pages

Entrez PubMed
National Library of **Medicine's** search service provides access to over 10 million citations in Medline, PreMedline, and other related databases, **...**
www.ncbi.nlm.nih.gov/entrez/query.fcgi - Similar pages

Experimental Biology and **Medicine Order** Form
Experimental Biology and **Medicine Order** Form. Please print this form, complete and send via fax to: 201-291-2988 or by mail to: Society for Experimental **...**
www.sebm.org/**order**.htm - 4k - Cached - Similar pages


PillWatch.com
Featured Drugs
Impotence
Viagra
Cialis
Levitra
Hair Loss
Propecia
Weight Loss
Xenical
Phentermine
Adipex
Meridia
Ionamin
Tenuate
Skin Care
Vaniqa
Renova
Retin-A
Pain Relief
Celebrex
Ultram
Vioxx
Birth Control
Ortho Tri-Cyclen
Herpes
Valtrex
Acyclovir
Find the Lowest Price for Viagra Online
Product:
Viagra (100 mg)
Choose the amount:
10 Pills
Preferred shipping method:
Normal Delivery
Find the lowest price!
With our price calculator above, you will find the lowest internet price on Viagra, exactly for your order.
Our calculator calculates the final price for your Viagra order which includes all fees and charges in all internet pharmacies. Say what you buy, how much and how to ship - and let the calculator above list the lowest internet
More Information
Buying Tips
1. Some pharmacies charge hidden fees for your Viagra order. Use our price calculator to see the final price for your order. The difference can be huge.
2. You can save a lot of money by buying Viagra in bulk. This requires a larger up-front investment but the saving is enormous. Or get your friends to buy Viagra together!
3. Some pharmacies heavily discount the cost for Viagra refill orders. If you want to make an order regularly, save money by considering how much you will save in future refills.
About This Drug
Viagra was approved by the Food and Drug Administration in 1998 as the first truly effective oral medication for the treatment of erectile dysfunction.
Story of the Month
Does any man, having satisfied a woman in bed with the help of Viagra,

Dieting Websites linking to weight loss for all.com
www.PillWatch.com. An up to date price calculator that finds
web site **links** are listed as a convenience to our visitors. ...
www.weightlossforall.com/dieting**links**.htm - 25k - Cached - Si

Links ab
Website - N
Reciprocal
www.about

Family &
http://www
Noni Juice
www.edene

HgH Human Growth Hormone
... Drug Information by RxList TOP 200 DRUGS OF 2002 Drug Information for the most prescribed products with **links** to Side Effects, Drug ... http://**www.pillwatch.com**/. ...
www.mdhealthline.com/**links**/online-pharmacies.htm - 26k - Supplemental Result - Cached - Similar pages

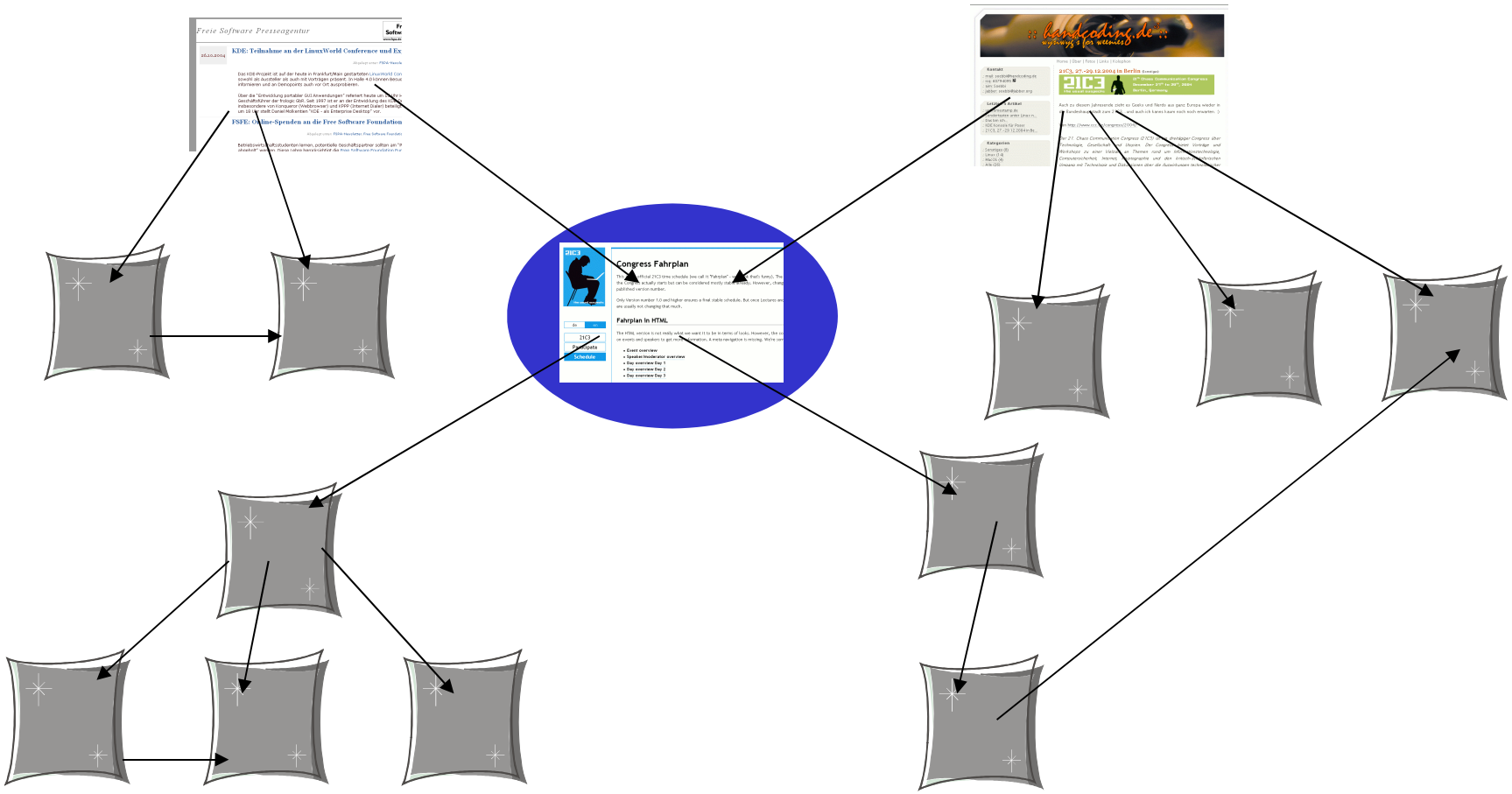
Transportation **Links**
links, bonuses section, search engine and directory. ... http://**www.pillwatch.com**
DESCRIPTION : An up to date price calculator that finds the lowest price ...
www.rosshealthcareclinic.com/**links**.html - 27k - Cached - Similar pages

- High rank for medical queries.
- Page itself looks trustworthy.
- Receives unrelated links.
- Spammed Page!

Classifying link spam

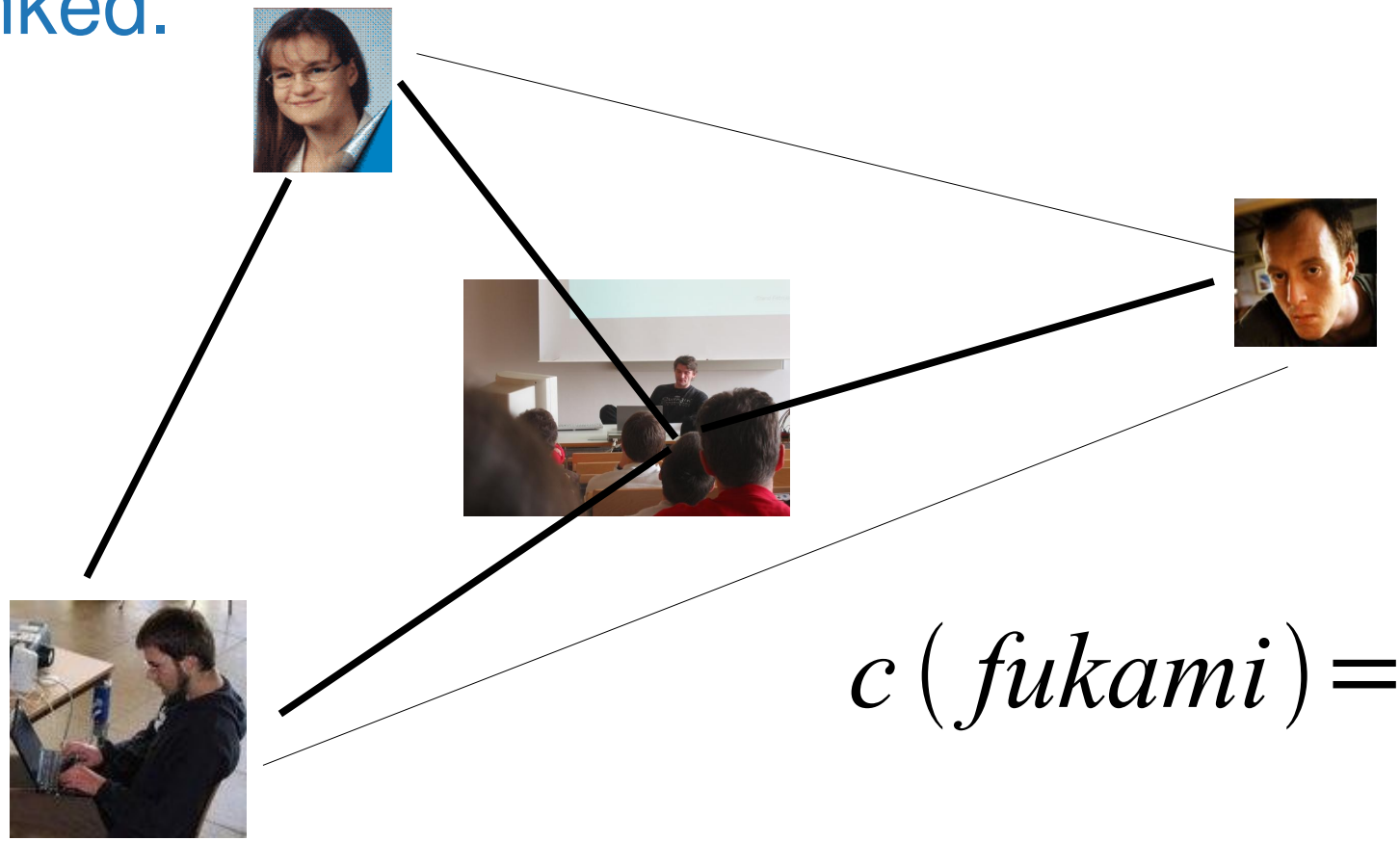
- Page rank assigns weight to pages.
- Spam disturbs coupling of links/ relevance.
- Identification of two spam types:
  - Link exchange/ link farms.
  - Guestbook spam.
- Evaluation of performance of classifier.

Crawling examples with nutch



Context similarity features

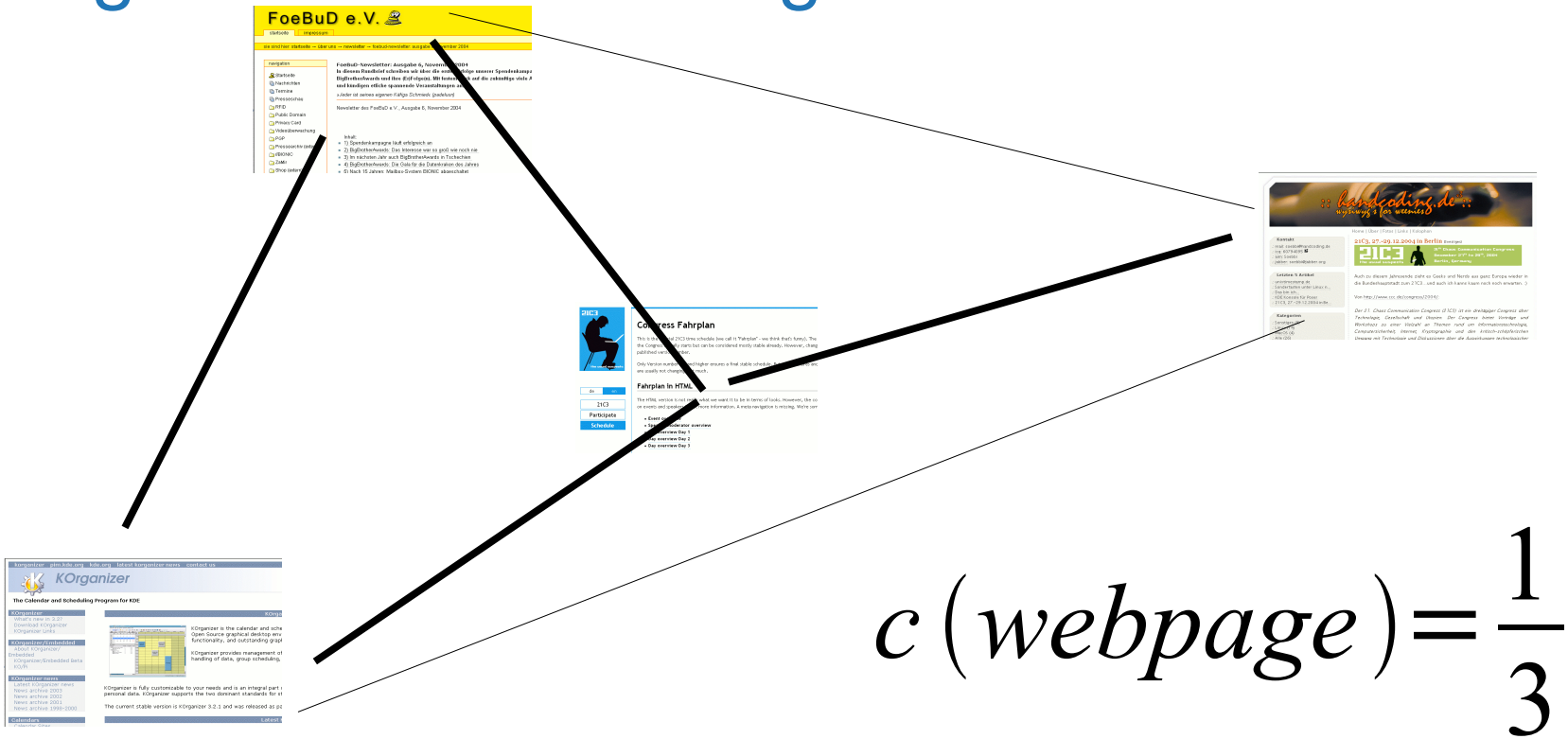
- In natural networks neighboring nodes tend to be linked.



$$c(fukami) = \frac{1}{3}$$

Context similarity features

- Clustering coefficient measures ratio of existing links between neighbors.



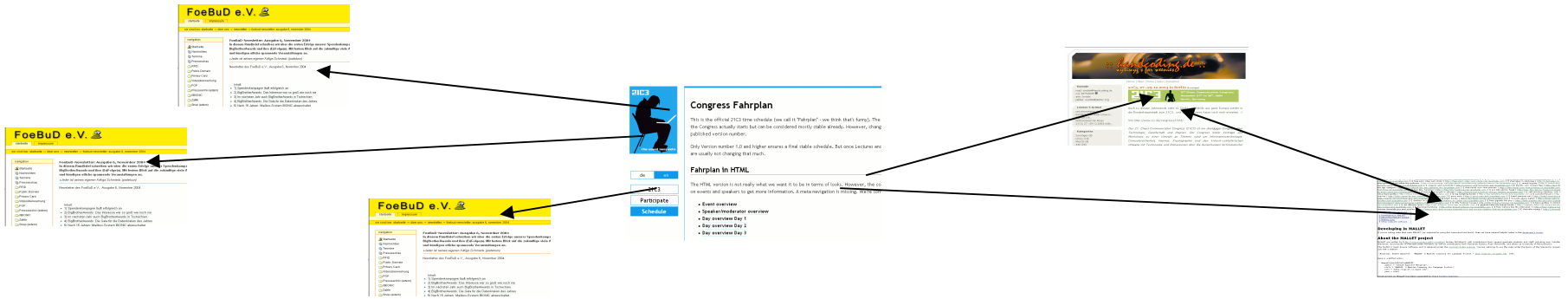
$$c(webpage) = \frac{1}{3}$$

Common features in linked pages

- Link farms tend to be automatically created.
- Linking pages are mostly similar.
- We look for common length, IP, MD5 hash of content, link target and link sources.

Common features in linked pages

- Lets look at the average number of pages linking to our example with common length:



$$common(webpage) = \frac{3}{5}$$

Similarity of example and links

- Some trivial link farms tend to recite on the same server as their target.
- We look for linked/-ing pages with the same IP, length, MD5 hash as our example.

Similarity of example and links

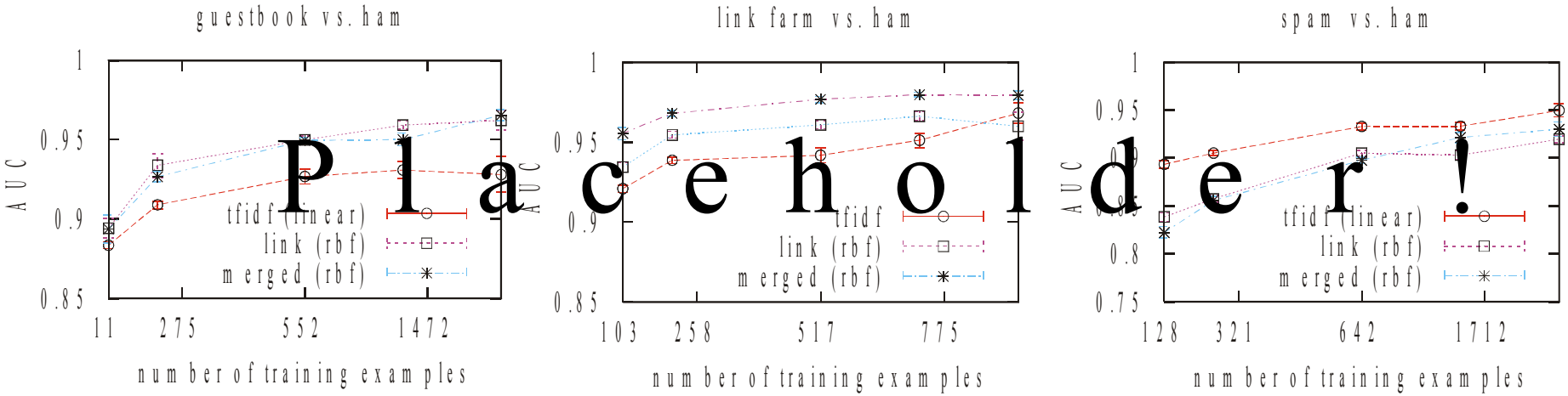
- Lets look at the average number of inlinking with same length as our example:

$$similar(webpage) = \frac{2}{5}$$

Evaluating spam detection

- Compare to gold standard – manual labeling.
- Use implicit ranking evaluation measures:
  - Compare the average rank of the link a user clicked on before and after changing the ranking.
  - In case you are a provider: Compare the ranking of a page vs. the actual usage of this site.

Performance of classifier



Placeholder!

- Best features:
  - Link based features.
  - Clustering Coefficient.
- Content useless for adversarial classification.

Game Theory

- Ranking search results as adversarial game.

- Spammers will react on any classifier.
  - Is there any kind of equilibrium in this game?
  - Are there “unspammable” ranking criteria?
  - Can internet usage data help win this game?

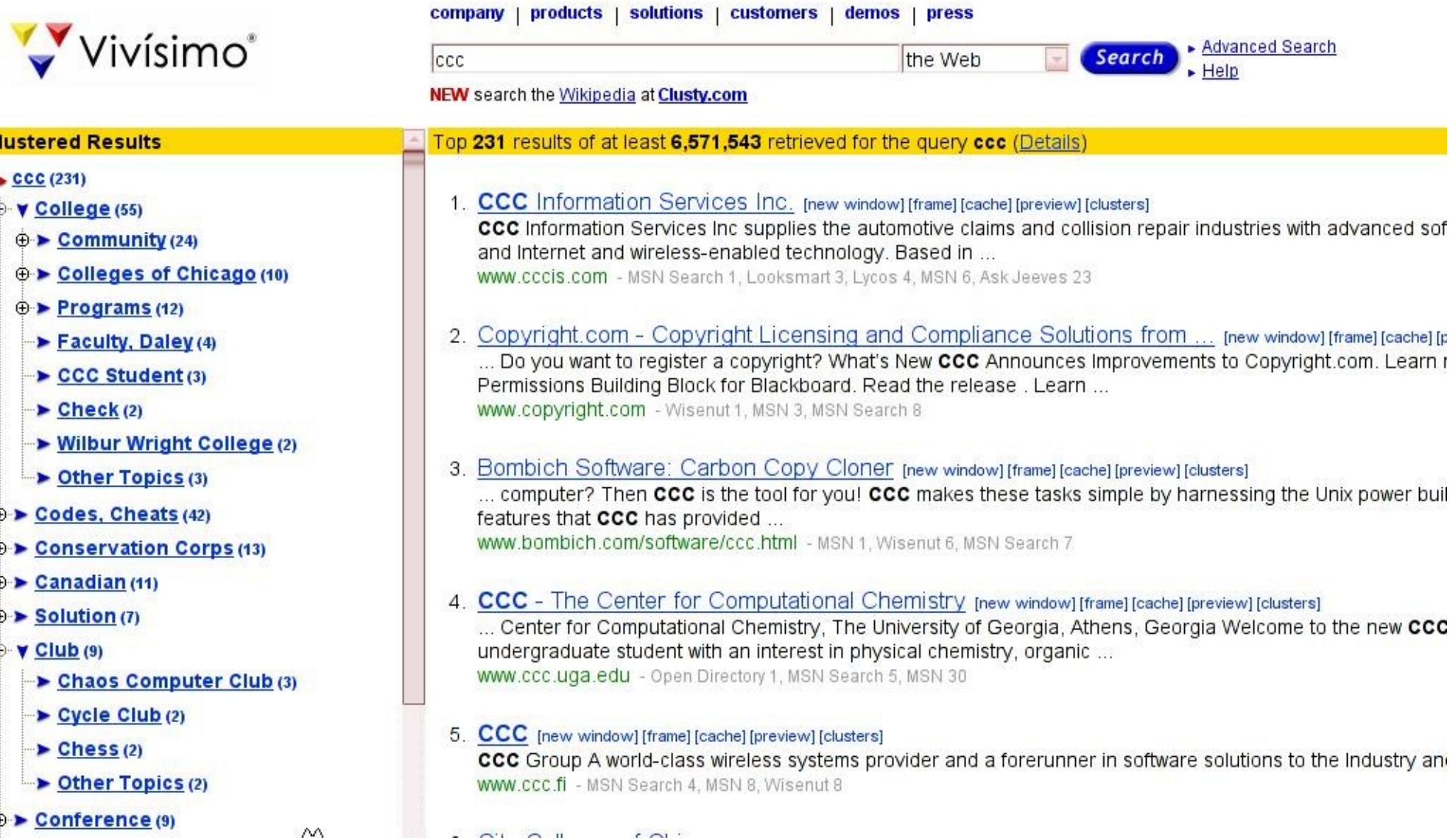
Enhanced Query Handling

- Cluster search results by topic.
- Enrich queries by further information.
- Do not ask for queries terms but for questions.

Clustering Search Results

- Most queries are ambiguous: panther, cluster.
- Handling ambiguity:
  - Cluster search results.
  - Add additional context to the query.
  - Categorize search results thematically.

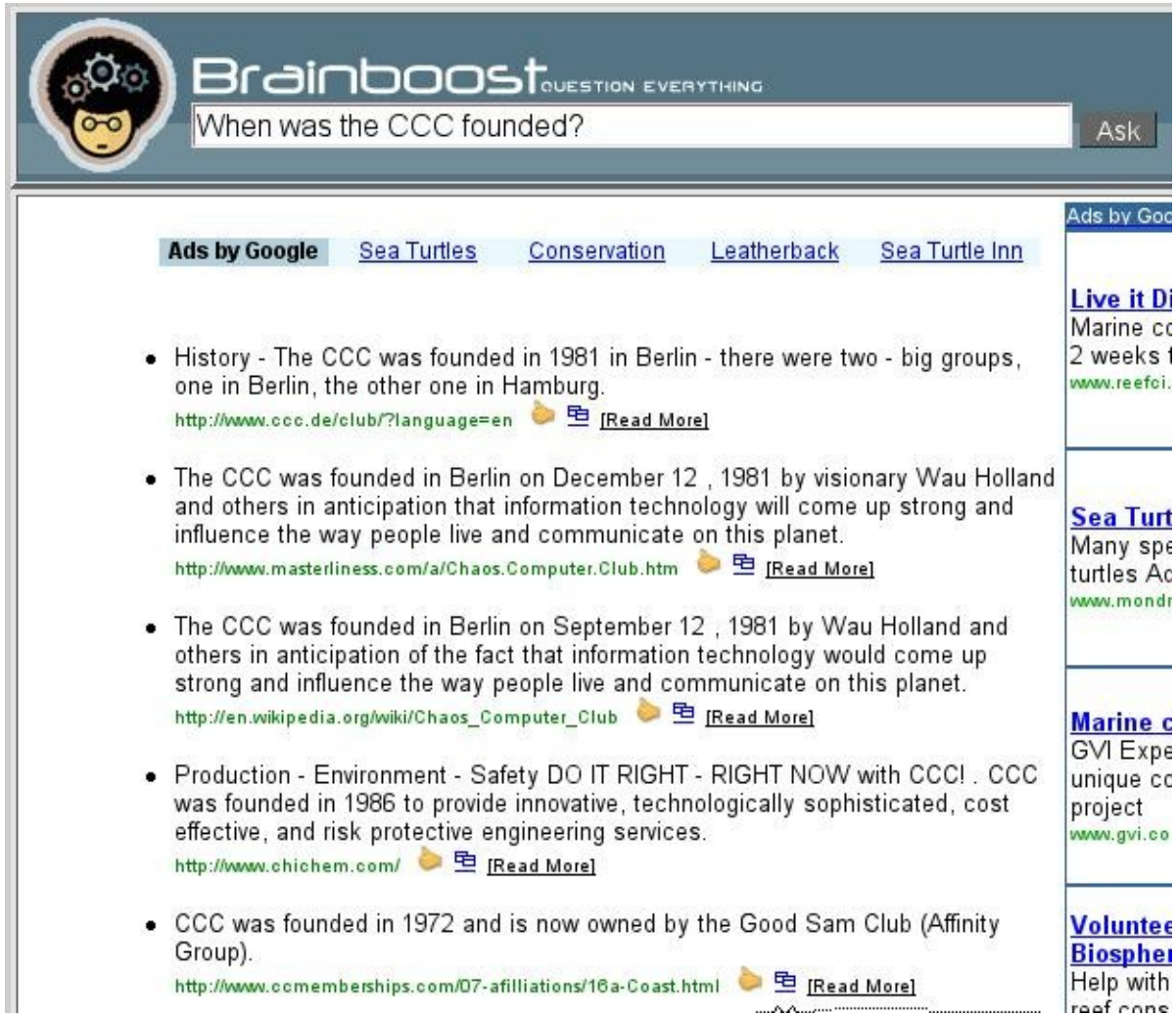
Clustering Search Results



Categorizing Queries – KDD Cup 05

- Queries in common only consist of few words.
- Categorize queries into topic hierarchies.
- KDD Cup 2005 task:
  - 1Mio queries
  - 80 categories
  - Only few examples – about 70.
  - 20 international research groups took part.
  - HU Berlin has won runner up award ;)

Answering Questions



Building your own search engine

- Don't start out to solve this problem alone.
  - Many obstacles to take.
  - Many interesting projects seeking help.
- Examples for interesting projects:
  - Google API – an API to Google search.
  - Nutch/ AspSeek – standalone search engines.
  - YaCy – Peer to peer search engine.

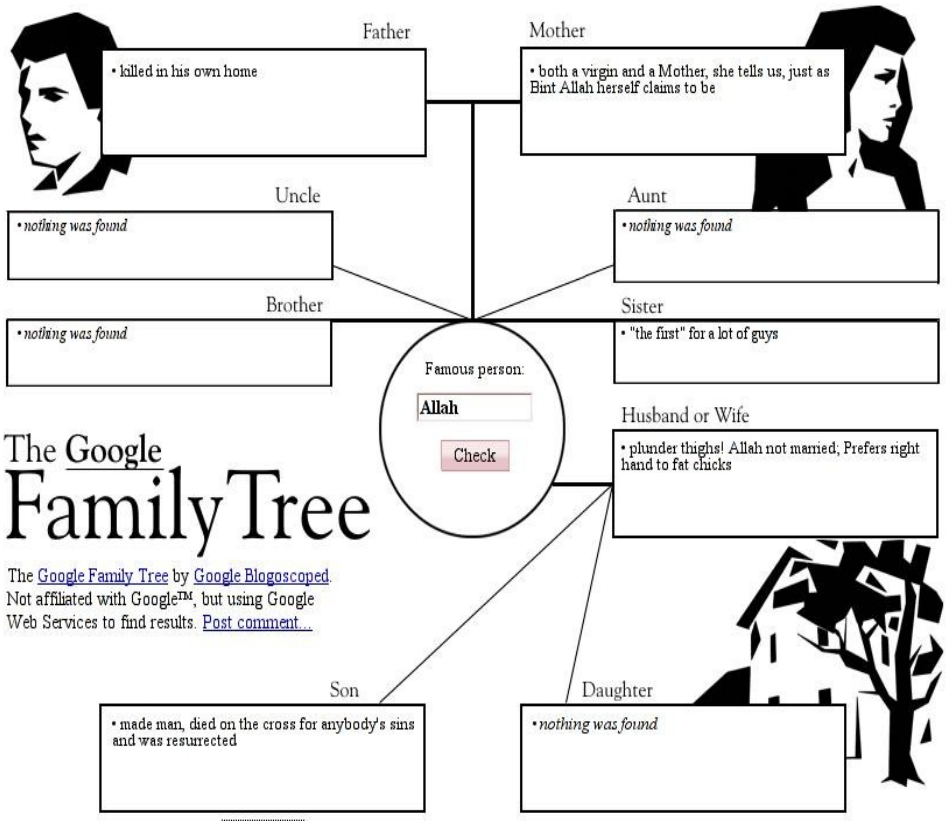
Google API

- From Google you need to download for free:
  - The example source code.
  - Your personal API key.
- Now you can write applications, that automate querying Google Web search.
- The number of daily free queries per key is restricted to keep API traffic under control.

Google API

- Advantages:
  - Use the power of Google for free.
  - Very easy to understand and use.
  - Quickly see the result of your work without having to buy expensive hardware or crawl web pages.
- Disadvantages:
  - Restricted number of queries per day.
  - Your application will rely on Google's service.

Google API



Painting Marx

Quote: *"Karl Marx and Frederick Engels had an excellent knowledge of world art and truly For this reason, he said, the appearance and poetic glorification of"*. Thus, From Hal Draper, Karl Marx's Theory of Revolution, Vol.1: State *and* Bureaucracy, He glorifies his rich Jew by painting him in pleasing colors. And, the ideas of Marx, on the contrary, took forward Blake's fourfold vision', which combined He also studied drawing and painting, combined all these. We must think of this as proven.

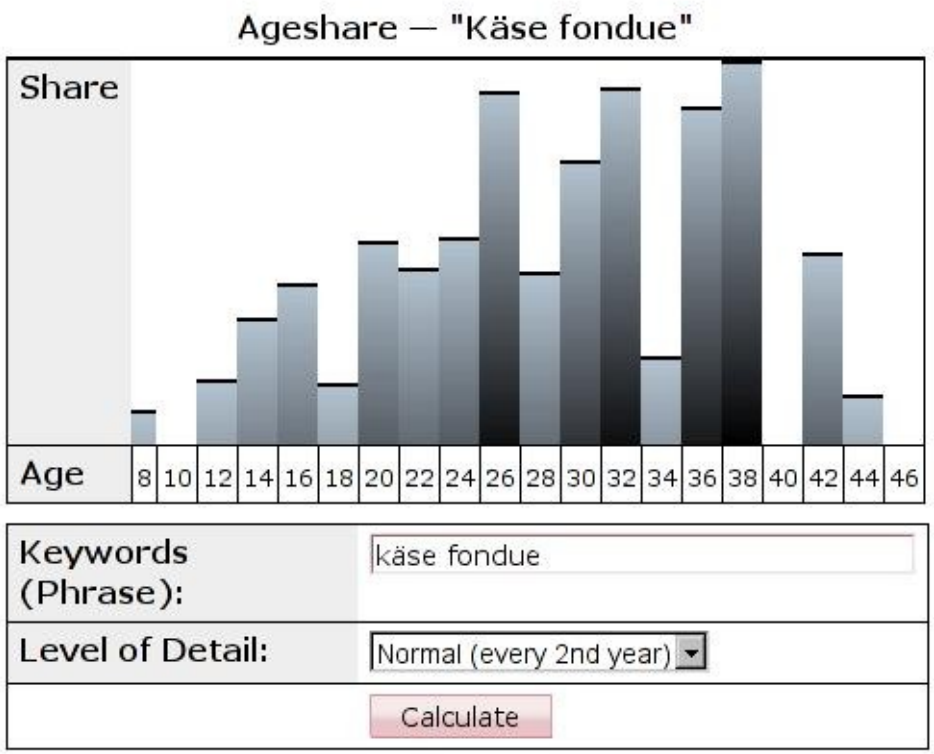
References

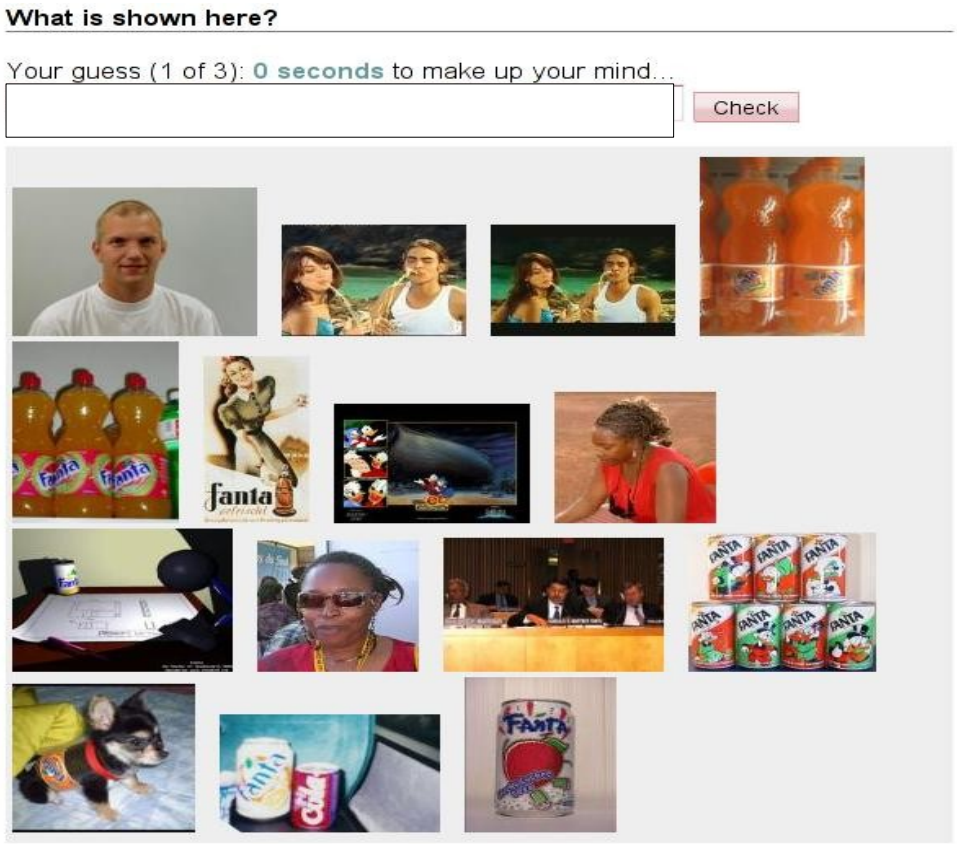
1. *Marx and Engels On Literature and Art Preface.*
2. *Hal Draper: Marx and the Economic-Jew Stereotype (1977)*
3. *Marx and the Fourfold Vision of William Blake | Libertarian*

This was one of Theorybot's 8 million constantly updated theories.

Show Random Theory

Google API


Ageshare — "Käse fondue"
Share
Age
8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46
Keywords (Phrase):
käse fondue
Level of Detail:
Normal (every 2nd year)
Calculate


What is shown here?
Your guess (1 of 3): 0 seconds to make up your mind...
Check

Nutch/ Aspseek

- Open source search engine implementations.
- For nutch online installations exist.
- From the project homepage you need to get:
  - Source code of the search engine you prefer.
  - Maybe some documentation on its internals.
- Now you can extend it according to your needs or work on tasks on the issue tracker.

Nutch/ Aspseek

- Advantages:
  - At least Nutch is open source and fun to read.
  - There are many interesting tasks in issue tracker.
  - Active project members who know the obstacles.
  - Requests on mailing lists are answered quickly.
- Disadvantages:
  - Need expensive hardware and fast internet.
  - Same problems with spam as any other SE has.

Nutch in action

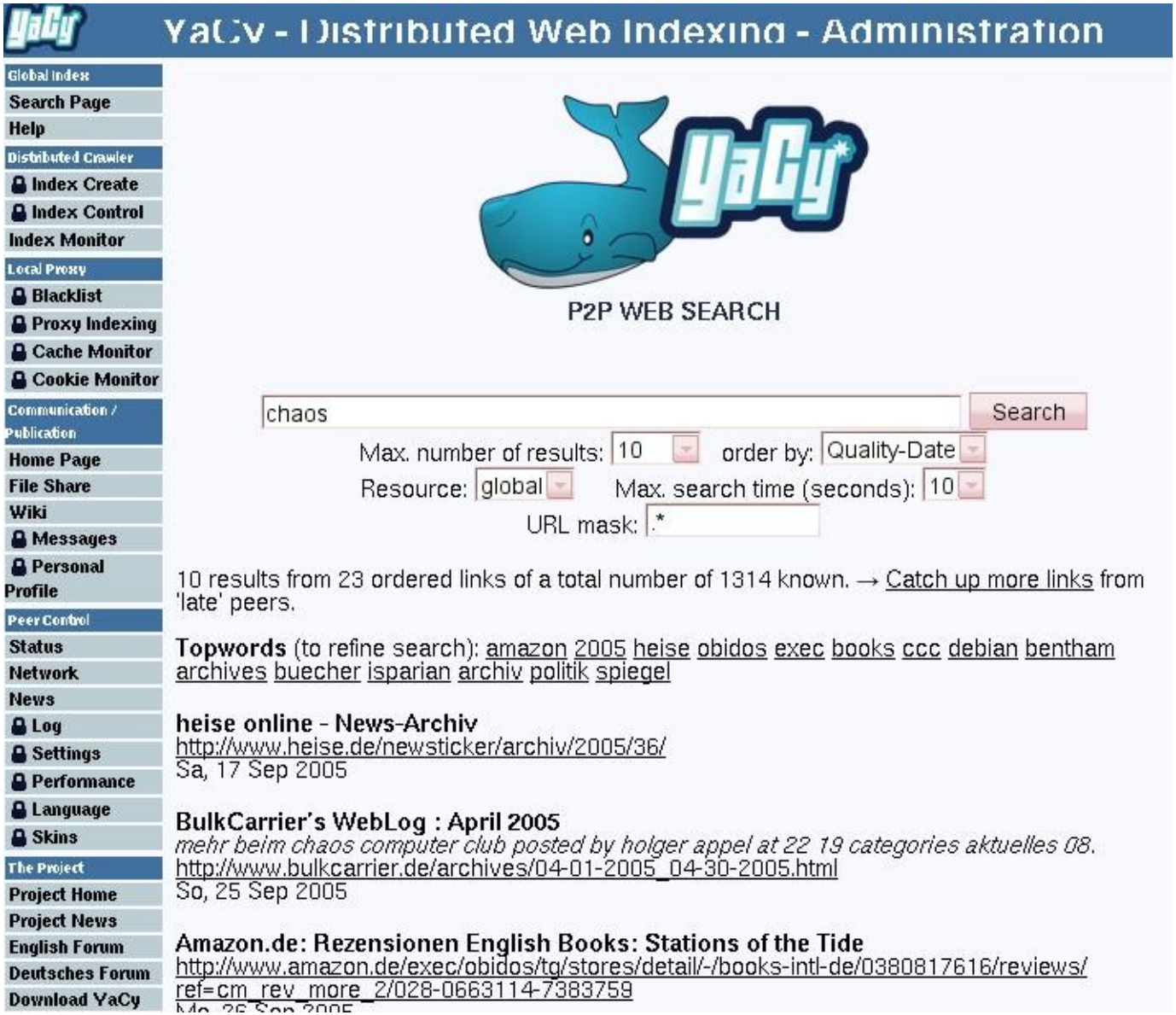
YaCy – Search going peer to peer

- Distributes crawling and indexing on clients:
  - After installation YaCy can be configured to crawl the WWW and publicly share the created index.
  - Each query answered based on distributed index.
- From the project homepage you get:
  - The

YaCy – peer to peer search

- Cool:
  - Open source search engine already running.
  - Basic idea: Distribute crawling, indexing and searching to those that want to search.
  - Again: many interesting tasks, ambitious project members :)
- Uncool:
  - Will need some time to be really useful.

YaCy – in action



Conclusion

- Search not only interesting for big companies.
- Many challenging and important tasks.
- Please take part in existing projects – its a very broad topic with many obstacles to take.

Search – only for big companies?

- To run traditional search engines we need:
  - Large bandwidth.
  - Large hard disks.
  - Fast CPUs.
  - A lot of intelligent algorithms to perform all tasks.
- Never the less, every one can contribute.
- A selection of examples:

Properties, that can be faked

- Content – easy to mask :)

  Key word stuffing, unreadable text.
- Content format – equally easy to adjust.

  Enlarge query terms, put them in head lines...
- Link popularity – once upon a time was hard.

  Link farms, spam guest books, exchange/ buy links.
- Anchor texts – remember Google bombs?

WWW – less obvious properties

- Link distribution power law governed.

  Outliers often are spam.
- Link graph has small world properties.
  - Small average shortest path.
  - Comparably small clustering coefficient.
- Degree of nodes correlated.

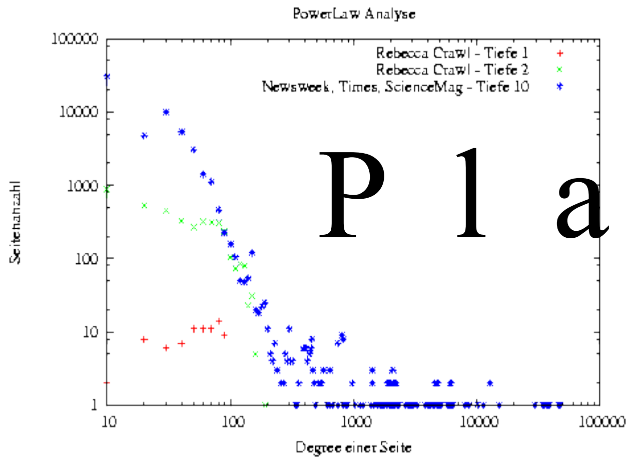
  Low in-degree nodes link high in-degree nodes and vice versa.

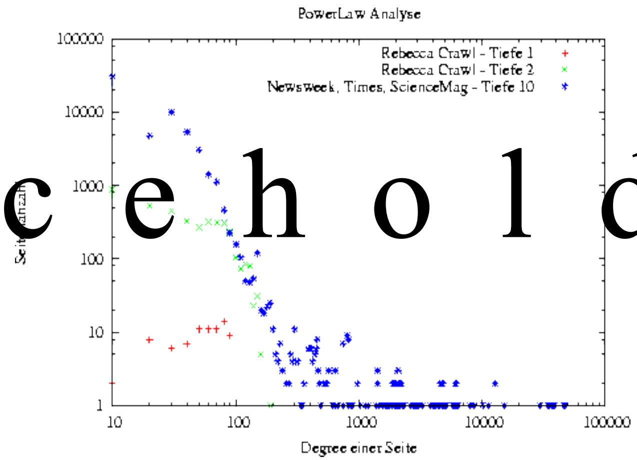
Used features – page itself

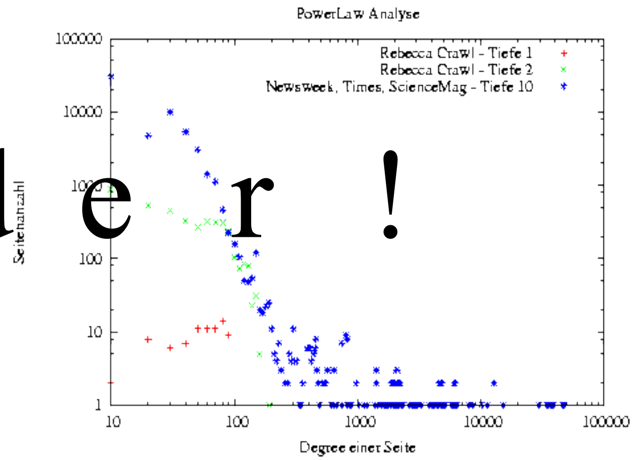
- Length of URL
- Length of domain.
- # of sub domains.
- Tilde in URL?
- TLD=edu/org?
- TLD=com/biz?
- TFIDF
- Number of inlinks.
- Number of outlinks.
- Is x a redirection?
- Size of x.
- Length of description
- Length of title.
- Number of keywords.

Power law

- Rich get richer.
  - Pages tend to link to known pages.
  - Assumption holds, if no links get removed.
- Outliers often spam pages.





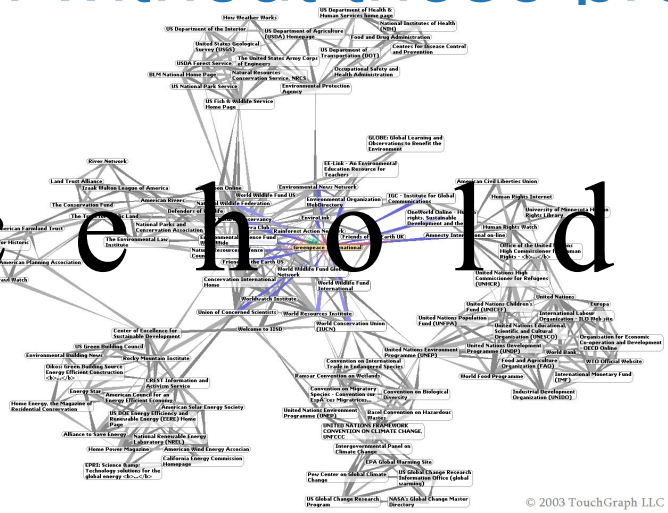


Placeholder!

We live in a small world.

- Small clustering coefficient.

  People know only comparably few people.
- Small shortest path between nodes.
- Link farms often without these properties.



P l a c e h o l d e r !

Agenda

- Properties of the WWW
- Basic architecture of search engines.
- One machine learning task in search engines.
- How can YOU contribute?

Agenda

- Properties of the WWW
- Basic architecture of search engines.
- One machine learning task in search engines.
- How can YOU contribute?

People to contact – examples

- Watch out for open source projects like:
  - Nutch.
  - Asp Seek.
  - YaCy.
- Google – Sommer of Code.
- Faculty of computer science of your choice, e.g. Knowledge Management at HU Berlin.

Research

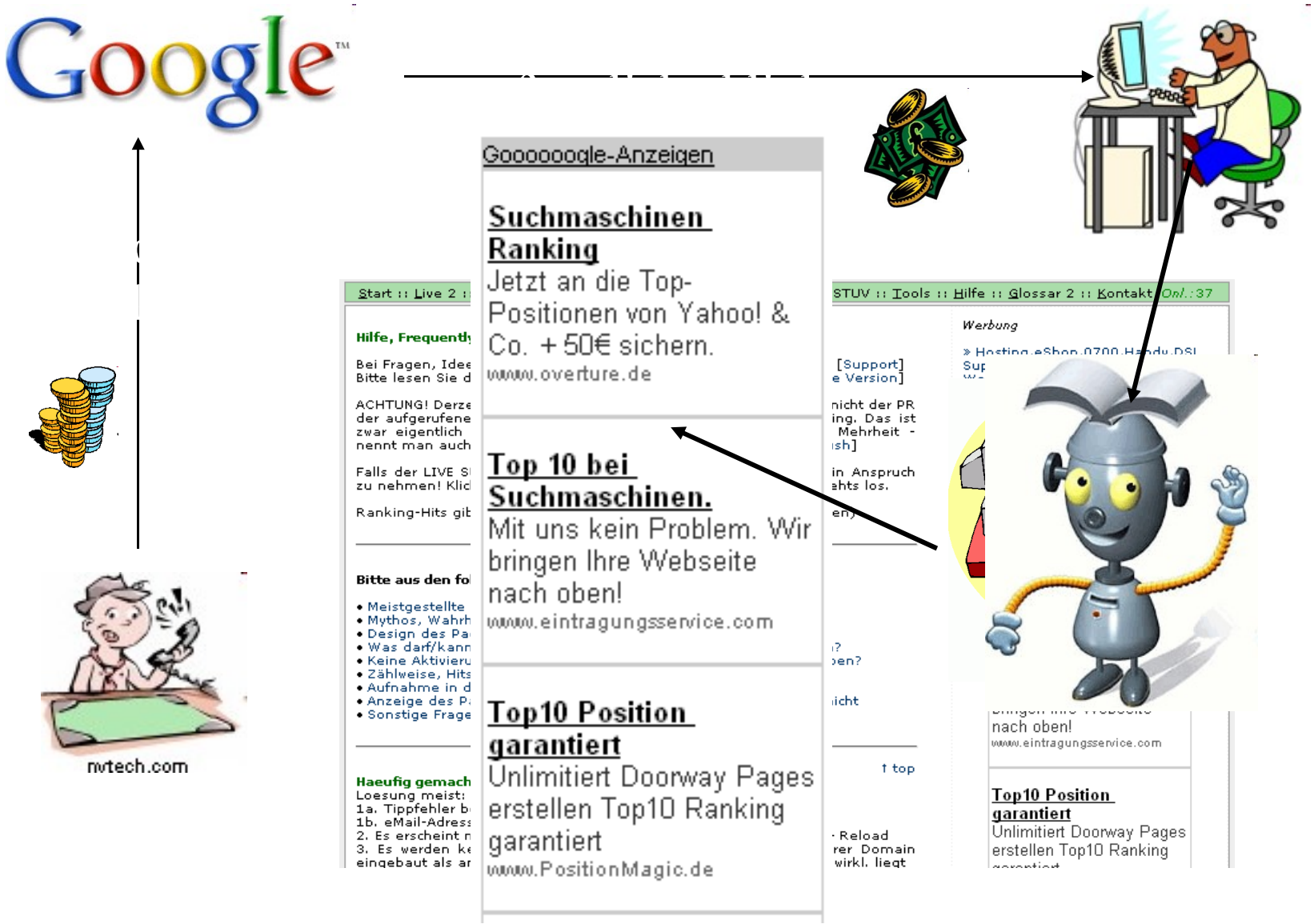
- Many open questions research topics.

- Each year:
  - Challenges e.g. KDD Cup to demonstrate state of the art technology.
  - New interesting publications on machine learning, information retrieval, link mining, ...

Rankingbewertung

- Google Proxy – User Feedback
- Explicit User Feedback
- Implicit User Feedback
- Incorporate User Location

Click spam


Google
Goooooogle-Anzeigen
Suchmaschinen Ranking
Jetzt an die Top-Positionen von Yahoo! & Co. + 50€ sichern.
www.overture.de
Top 10 bei Suchmaschinen.
Mit uns kein Problem. Wir bringen Ihre Webseite nach oben!
www.eintragungsservice.com
Top10 Position garantiert
Unlimitiert Doorway Pages erstellen Top10 Ranking garantiert
www.PositionMagic.de
Werbung
nvtech.com

Agenda

- Properties of the WWW
- Basic architecture of search engines.
- One machine learning task in search engines.
- How can YOU contribute?

Conclusion

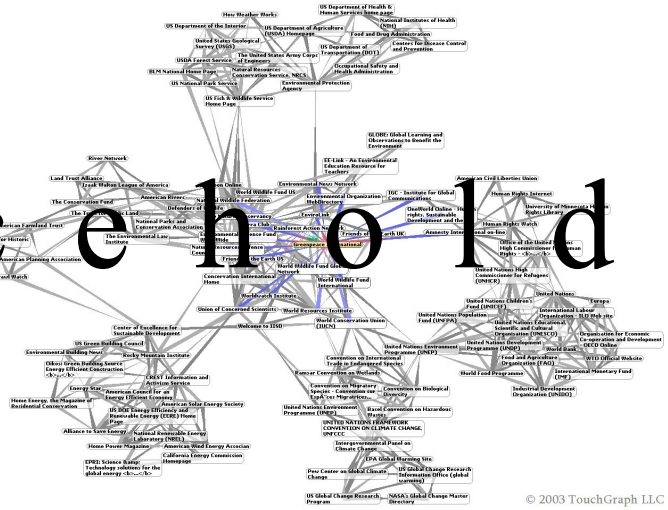
- Search not only interesting for big companies.
- Many challenging and important tasks.
- Please take part in existing projects – its a very broad topic with many obstacles to take.

Violations of observations

- Outliers in power law plot often link spam.
- Networks of pages without small world properties often artificial link farms.

Correlation of node degree

- High in-degree nodes link low in-degree ones.
- Comparable to natural networks.



P l a c e h o l d e r !

Data mining the NSA^h^h^hWWW

- What to do with all these data?
  - Search for relevant information – what's relevant?
  - Find similar web pages.
  - Track evolving topics or find new ones.
- Are there any major problems?

  Spam, unreliable or false information.

  New types of media (Blogs, Wikis).

  Presentation of huuuuge amount of search hits.

Search engines from scratch

- What we need today:
  - Laaaarge hard disk, for good coverage.
  - Looots of bandwidth, to improve crawling speed.
  - Intelligent algorithms to find most important links.
- Do we have that much money? ;)

Search engines from scratch

- What we need today:
  - Fast CPU to increase parsing/analysis speed.
  - Large hard disk to save index.
  - Intelligent algorithms for parsing pages.
  - Intelligent algorithms for scoring pages.
- Can be done with common PC.

Search engines from scratch

- What we need today:
  - Bandwidth for query retrieval/ result delivery.
  - CPU cycles for analysis.
- Intelligent algorithms for:
  - Finding similar results.
  - Categorizing web pages.
  - Ranking matching pages.
  - Disambiguation and extension of queries.
  - You can think of many more features here :)