The *Real World* Web Search Problem:

Bridging The Gap Between Academic and Commercial Understanding of Issues and Methods

Dr. Eric Glover - SearchMe.com

Overview and Objectives

- The world of academic search engine research has a large disconnect from the ‘commercial search world’
- This talk aims to reduce this gap:
  - Explain how a commercial search engine works (components, problems, objectives)
  - Describe five classes of problems academics face in this space
  - Share real stories and personal experiences
  - Propose approaches to reduce the effects of these problems
- Talk outline:
  - Part 1: The “theoretical search engine”
  - Part 2: The disconnect between the theory and reality
  - Part 3: Five classes of problems which face researchers in this area and suggestions to move forward

About the Speaker: Dr. Eric Glover

- Years of academic and commercial experience in Web Search
- Numerous publications and PC memberships
- PhD from University of Michigan (completed 2001)
  - Title: Using Extra-Topical User Preferences to Improve Web-Based Metasearch
- Worked at NEC Laboratories America (Web group)
- Worked at Ask.com - variety of projects (several are live)
- Now at SearchMe.com as Classification Architect
  - http://www.searchme.com/ - jobs@searchme.com
  - We are hiring, hint hint...

A True Story

- At a commercial search engine, from an employee with a relevant PhD

Started with a conference paper (won awards) about a particular classification application relevant to their needs.

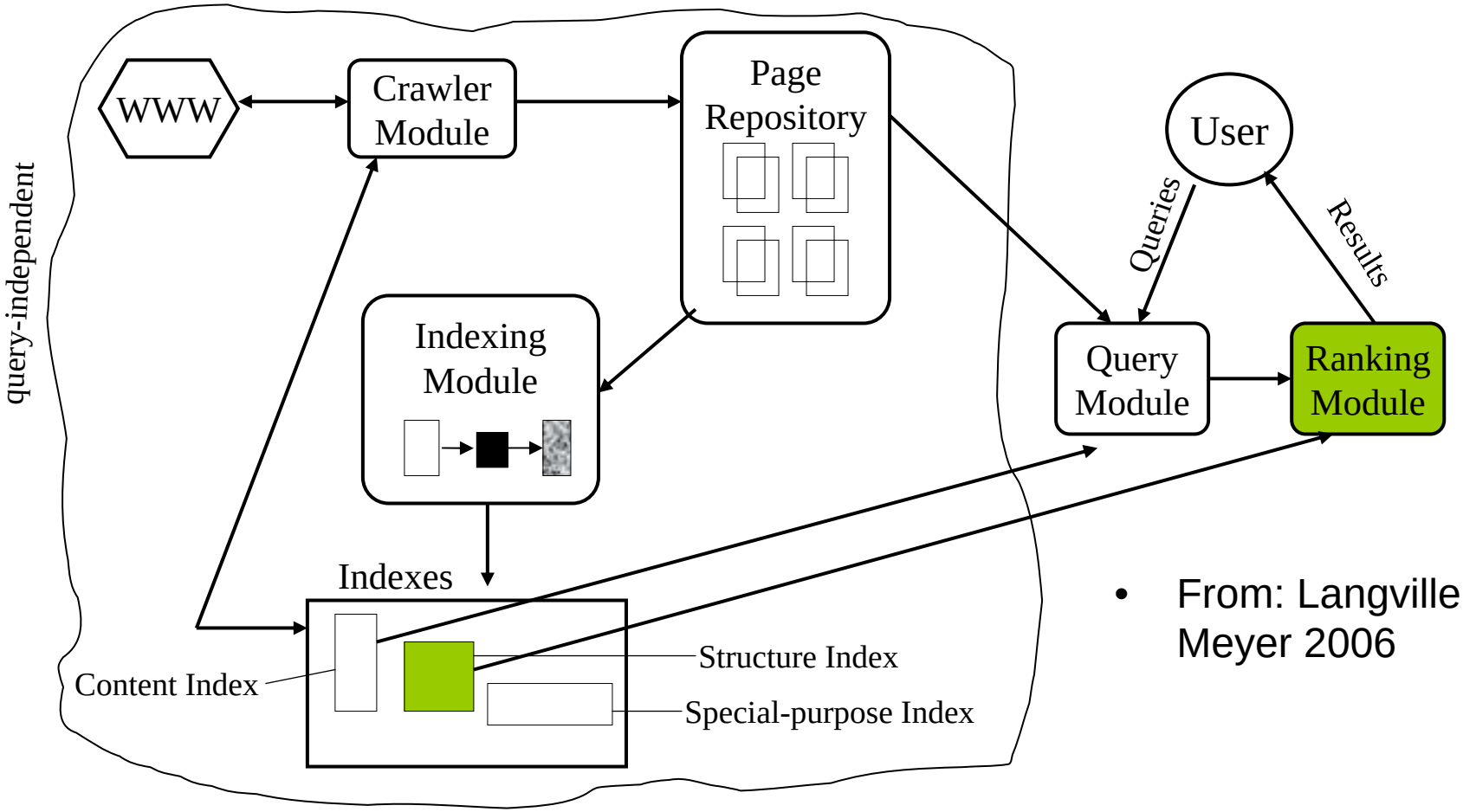
Implemented the paper, and confirmed it performed very well - on the same data described in the paper

Applied to real-data and the result was not statistically significant

Talk Flow - Part 1 - theoretical search

- Introduction to Web Search Engines (theory)
  - Architecture
  - Theoretical model
  - Fundamental component details

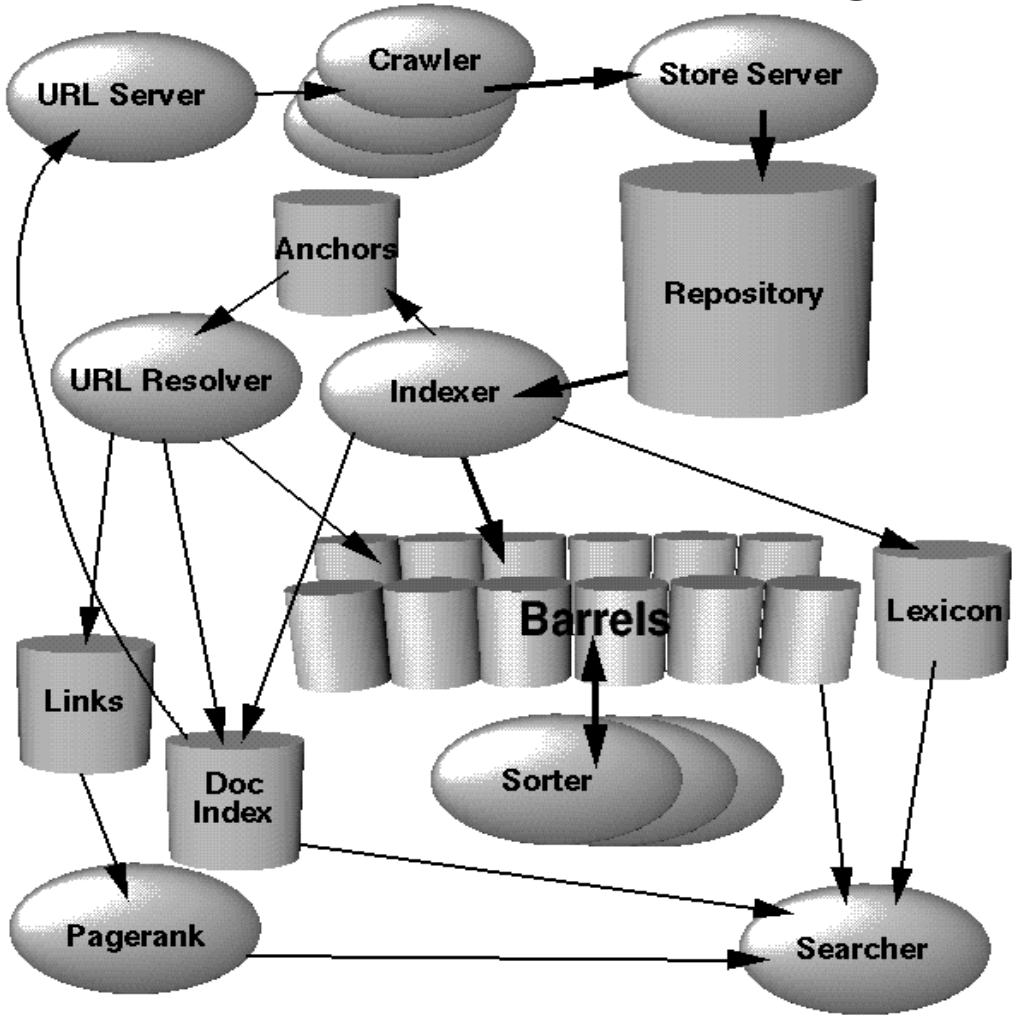
What is a Web Search Engine?



- From: Langville and Meyer 2006

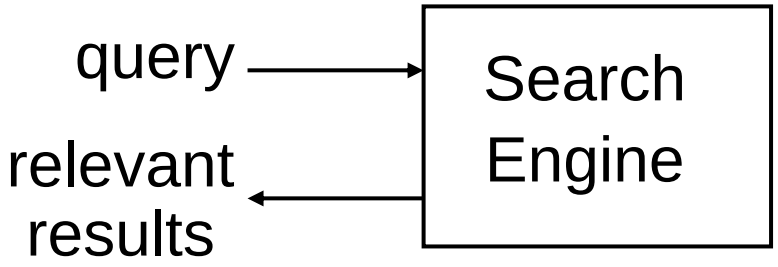
What is a Web Search Engine?

From:
Brin and Page 1998



What is a “Web Search Engine”?

- Conceptually: input *queries*, output *relevant* results



- What is the source of the data that becomes results?
- How is a query mapped to this set of relevant results (technically)?
- What *defines* “relevant”? How is the calculated?
- What constraints and requirements are there affecting system design?

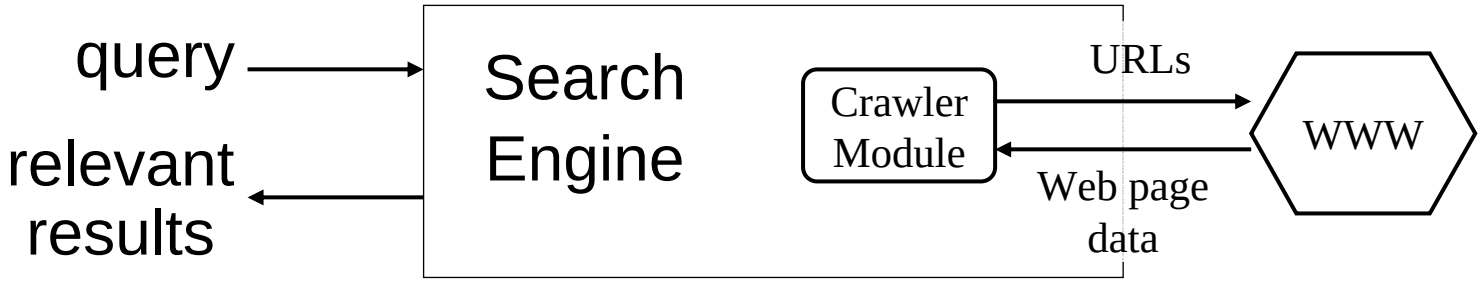
What is a “Web Search Engine”?

Fundamental tasks of the ideal/theoretical web search engine

- OFFLINE
  - Obtaining contents - Crawling
  - Structuring contents - Indexing
  - Analyze the “contents”
- ONLINE
  - Processing a query - Obtain ‘relevant set’
  - Ranking results - Order (rank) ‘relevant set’

Search Engine Theory - Crawler

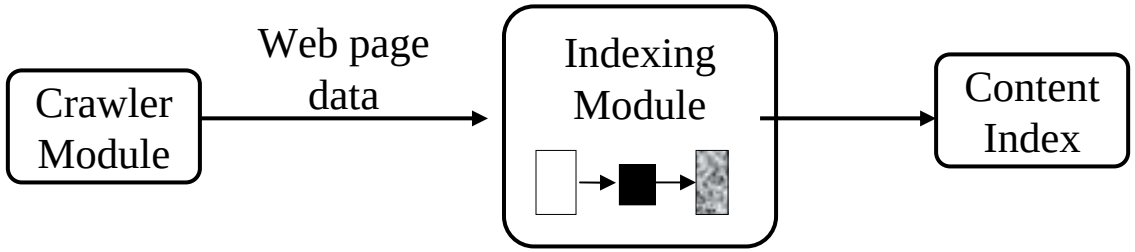
- What is the source of the data that becomes results?



- "Results" start as "pages" on the web
- The web can be viewed as a tool to obtain a mapping between URL and "content" (URL, content) tuples
  - A "crawler" or "spider" requests web pages and 'saves them' to later become results
  - In effect the 'crawler' simulates a human browser to capture the page content

Search Engine Theory - Indexer

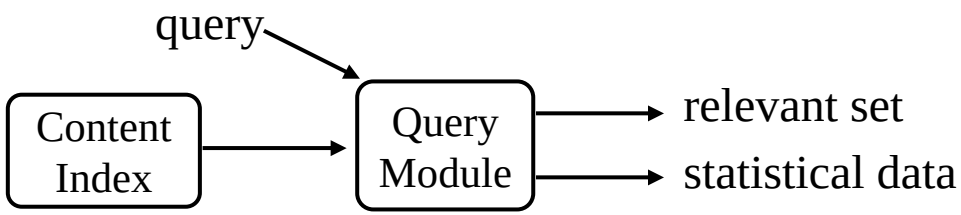
- How is the ‘content’ ‘processed’



- Web pages are parsed (not shown) and the ‘text’ is “indexed”
  - Parsing converts the HTML to the ‘text’
- An index, often an ‘inverted index’, is a tool to provide a map from “word” to URLs
- A theoretical indexing module inverts the set of [URL->words] to a set of [word->URLs]

Search Engine Theory - Relevant Set

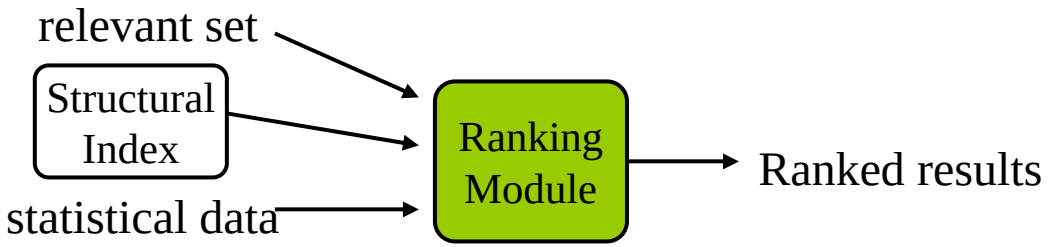
- How to obtain a ‘relevant set’ from a query (theory)



- Input query is “processed” against the ‘content index’ producing the ‘relevant set’, and ‘statistical data’
- The ‘theoretical process’ maps query to words `q->[w1,w2, w3,...,wn]`
  - Each word is applied to the inverted (content) index
  - The combined set of pages forms the ‘relevant set’ or `q->[U1, U3,...,Un]`
- The ‘relevant set’ is supposed to be the set of documents considered for ranking, ideally covering all ‘possibly useful results’.

Search Engine Theory - Ranking (Theory)

- How to order the ‘relevant set’

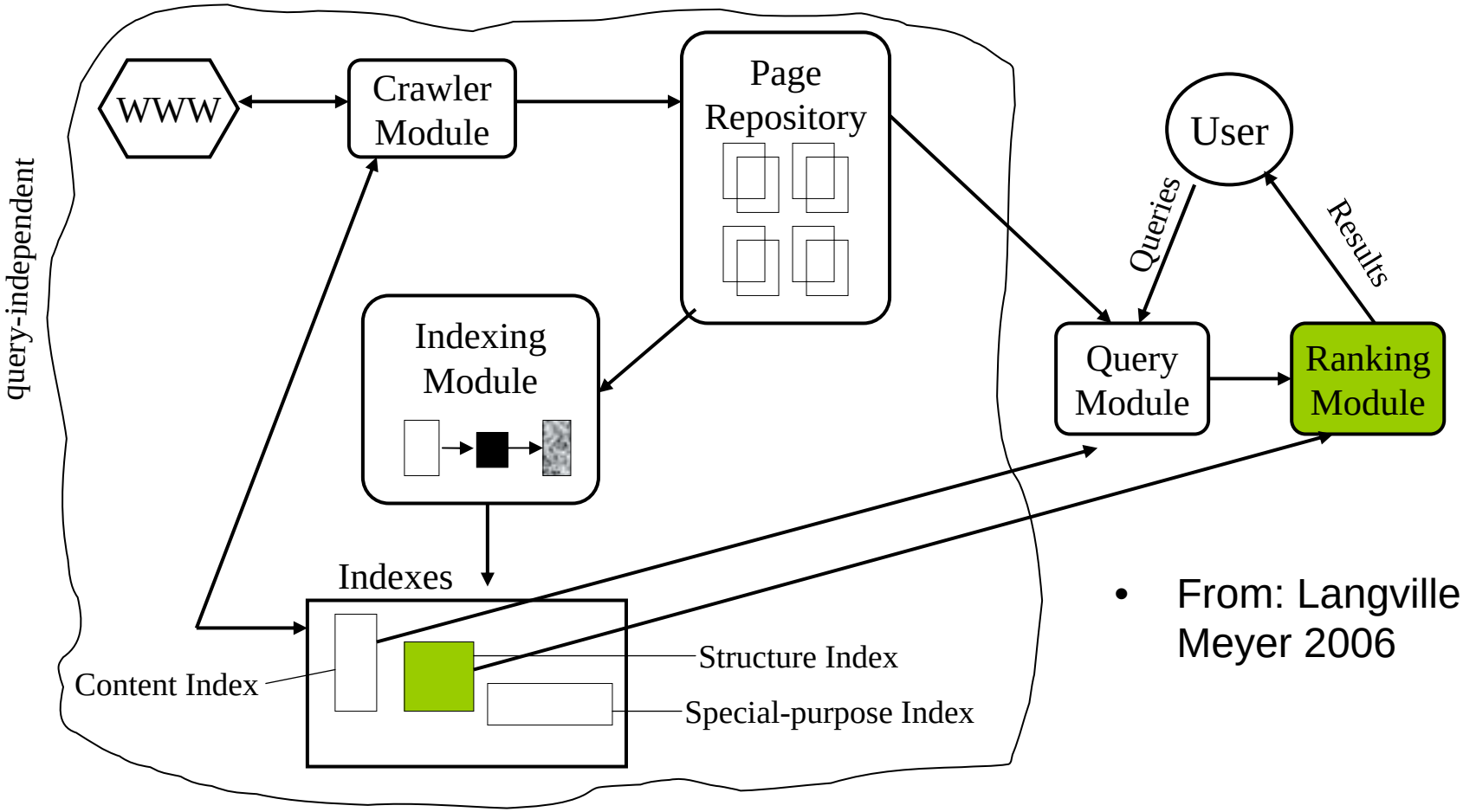


- Ranking module scores results based on information about each result in the relevant set from the indexes (and query module)
- Each result is given a score based on the ‘data’ about it
  - Simple model can include f(content score) + f(popularity-score)
  - Content score is a function of the word statistics, and popularity-score is a function of the ‘structural index’ (PageRank may be used here)

What is missing (theory)?

- How does a crawler work?
  - Crawler - needs a source of URLs
    - Indexer/indexes can provide both a set to crawl as well as priority
- “Other indexes”
  - content index maps words to URLs (need more than ‘words’)
  - structure index maps URLs to in/out-links (a web-graph)
    - How to compute/manage a ‘structural index’
- Computation of content-score
  - How do we utilize word statistics
- UI issues - i.e. descriptions

What is a Web Search Engine?



- From: Langville and Meyer 2006

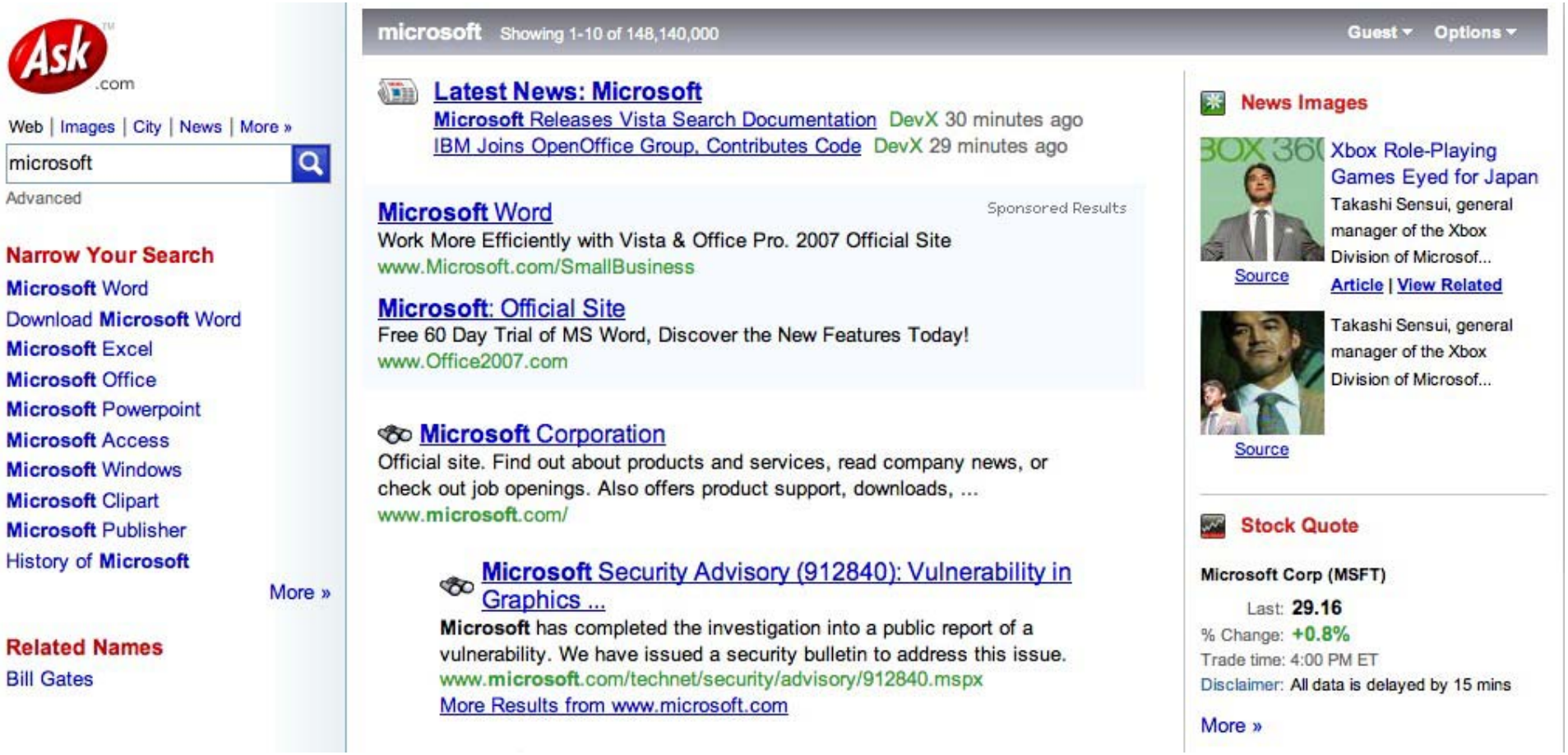
Good References

- Intro the architecture and details of graph algorithms including PageRank:
  - Amy N. Langville and Carl D. Meyer . Google's PageRank and Beyond - The Science of Search Engine Rankings. Princeton University Press, 2006
- Important paper from the 'experts' - addresses some technical issues of a 'large scale search engine
  - Sergey Brin, and Lawrence Page.The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 33:107-17, 1998
- Large scale indexing
  - I.H. Witten, A. Moffat, and T.C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, second edition, 1999.

Part II - Theory Gets Disconnected

- What we just described is a ‘theoretical’ search engine
  - Most of the components are real, but the picture is not complete, factors they utilize and how they work may vary.
  - Each piece has its own unique challenges before the system is viable commercially
- KEY QUESTION: What does a commercial search engine want?

Commercial Search Engine != 10 Blue links



Important Properties Of Commercial Web Search

Millions of heterogeneous users
Web is large (practically infinite)
No quality control on pages (quality varies)
Must consider maliciousness
Result ranking cannot assume independence
Content of web page not sufficient to imply meaning
Real-time/fast expectation
UI is extremely important
Goal is to make money

- To be successful a commercial Search Engine must address all of these issues/properties

Separating Commercial Web Search from Theory

- Real search engines are more than ‘text to results’
- Non-text searches include:
  - Images (how do you index those?)
  - Other media (audio, video, code, etc...)
  - Maps/geographic
- Other Services or Features
  - Language Translation
  - Spell correction
  - Local Search (find POI/Places)
  - “Personalized” or localized (as opposed to local search)
  - Verticalized: Papers, FAQ, Software, etc...
  - Advertisement matching - Where the $$$ come from

Simple Theory vs Cold Reality - Crawling

- Not all data originates from a crawler
- Feeds and other non-crawled sources are increasingly important
  - video, news, movies, sports scores, Wikipedia, Blogs, etc...
- Crawlers have lots of intelligence and challenges
  - Dealing with site constraints (i.e. max crawl rate per server)
  - Dealing with site errors - if you fail once is the site dead - how do you know store happened before? (for 1 Trillion page attempts)
    - Can't even store records of more than 4B URLs with 32-bit IDs
- Fundamental issue: How much of the web is 'crawlable'
  - If you follow the rules many sites say "robots get lost"
- What does 'AJAX' mean for crawlers? Dynamic content?

Simple Theory vs Cold Reality - Crawling

- Crawler challenges
  - Update frequency - should I crawl cnn.com only once every two weeks?
    - Relevance is f('indexed content')
      - Problem if 'indexed content != 'real content'
      - Old content causing false matches, or dead pages are very bad
  - Intentional attacks against search engine crawler
    - data overload "| /dev/random" or "spider traps"
    - Cloaking - send one content to crawlers another to browsers
    - Hidden redirects - hide redirects on page from parsers
  - Not all pages use HTML 1.0 http[s] URLs - this is a parsing as well as crawling problem - javascript, flash, pdf
  - Cookies, refer fields, partner relationships, other?

Simple Theory vs Cold Reality - Indexing

- Technical
  - Dealing with rapid updates, size and performance constraints
    - A search engine has fractional seconds to return all results
    - Might crawl > 100 M pages in a day (some multiple times)
  - What about ‘phrases’ - searches are not ‘bag of words’
    - Positional information? Structural (throw out case & punctuation)?
    - Positional and ‘field’ information for relevance ranking as well as retrieval i.e. in title, header, image-alt, etc...
- Deciding what to index - Should we index spammers.tld?
- Dealing with intentional manipulations
  - “invisible text” or “keyword stuffing”

Simple Theory vs Cold Reality - Indexing

- Non-text content data worth storing
  - Quality scores - a number associated with a URL or Quality=f(U,q)
  - Link graphs (could be useful for real-time ranking) (part of structural index)
  - DNS information about a site, server response times
  - Human judgments (yes Google and other search engines utilize human editorial judgments)
  - Categorization/classification data - including page-type measures

Simple Theory vs Cold Reality - Query Processing

- Fundamental problem: “words on page” not imply “purpose of page”
- Many ‘queries’ would have large relevant sets
  - According to Google: q=nato has about 57,600,000 results
- Ambiguous ‘words’
  - *Polysemy* - words with multiple meanings “train car” “train neural network” (could be indexing issue)
  - *Synonymy* - multiple words same meaning: “neural network is trained as follows” “neural network learns as follows” (could be indexing issue)
- Entities, concepts and parsing
  - “jack black” “black jack”, “a jack which is black”
  - “Michael Jordan” (NBA) “Michael Jordan” (Berkeley)
  - “pictures of the white house” “pictures of a white house”

Relevance - Theory

- What does relevance really mean?
- Imagine a ‘perfect *relevance* function’
- This function R(URL, q) is able to take some given query and a URL and produce a score.
- s1=R(U1,q), s2=R(U2,q)
- Simplification: Lets assume that the function is optimized for a specific user and has the property that:
- IFF s1 > s2 then user prefers U1 to U2

Relevance - Theory Problem 1: Duplicates

- What will the results look like for this “perfect relevance function”
- Duplicate or near duplicate content on the web is common: mirrors, spammers, same original source (AP, blog replies, etc.)
- Example (made up) - q=‘*michael jordan*’
  - R1=“Michael Jordan official home page from myspace.com”
  - R2=“Michael Jordan official home page from mirror.com”
  - ...
- Real query: ‘*"Post-Cold War" "After the September 11 attacks" "Expansion and restructuring"*’ on Yahoo
  - Returns over 100 likely Wikipedia mirrors for “NATO”
  - Several are spam sites including:

92. thinkcelebrities.com
Free female celebrity website were you can view all hot sexy girls, ... 1.5 **Post-Cold War**. 1.6 **After the September 11 attacks**. 1.7 **Expansion and restructuring** ...
www.thinkcelebrities.com/?title=NATO - 330k

Relevance - Theory Problem 2: Marginal Value

- The usefulness of the second ranked result is dependent on what was ranked first
- What if top 5 results are all different perspectives on the same news story? Not duplicates
- Even if the first result is the “best” the others do not add more usefulness if they do not provide new information
  - Informativeness has been studied for over 30 years
- A user wants to generally ‘learn’ about something
  - Different perspectives or types of content can be useful, even though their R(q,U) might be much lower

Relevance - Theory Problem 3: UI

- How does a real users don't "know" if a result is useful
- Which is "better", the identical result with:
  - Description 1: "Click here to play video"
  - Description 2: "A video about Michael Jordan's life"
  - Description 3: Actual images extracted from the video with the text from description 2
  - Some existing Search Engines modify the description or UI
    - Ask: binoculars, 'Ask-X mixes multiple kinds of results'
    - Google, Yahoo and Ask modify descriptions and or titles sometimes using text not contained within the page (not even in meta-description)
- The "golden triangle theory" - where do users look?
- Users avoid engines with 'bad UIs' or too many ads

Relevance - Theory != Reality

- Words on page not imply ‘intention of page’
- Not all users are average
  - Some users want a ‘non-dominant meaning’, or change goals
- “relevance” is a misnomer
  - *Usefulness* - more difficult to quantify or measure or compare - subjective
- How do ‘other factors’ come into play?
  - Not merely a function of (q,U) - other factors
  - Other factors (some in use)
    - Geographic location (IP) Google redirects based on country
    - Past history/actions Amazon.com shows different items
    - Demographic data, interactions with “other services”
    - Temporal: today vs tomorrow different ranking of same results

Relevance - Considerations

- Other factors
  - Resources/costs - if there were the ‘perfect relevance function’ how long would it take? 10 ms or 10 s
  - What is the user looking for? Content/Information need
    - Imges? Video?
    - An answer to a question implies different scoring
    - A specific site/navigational search - only show one result?
- Evaluation
  - How do you know if your approach really is better?
  - If you did have the ‘perfect relevance function’ how would you know it was better than an imperfect, but okay one?
    - Are users honest when you ask them?
    - What does session data really mean?

Relevance - Current Approximations

- What is used now? What are the assumptions?
- Popular IR measures include Precision and Recall
  - Advanced forms include: P@X, MAP, MRR
- What are the assumptions?
  - Typically assume ‘relevant’ ‘non-relevant’ OR ‘single best result’

Relevance: Academic Measures

- Precision, Precision @ X, MAP (Mean Average Precision)
- Simple summary - penalize inclusion of ‘non-relevant results’ in result set
  - Implications/assumptions
    - *Number of* relevant and not-relevant results matter
    - Typically all “errors” are equal- swap two ‘relevant’ results or two ‘non-relevant’ results produces same score
    - Results viewed as an unordered set (P@X considers top x), “score” is determined by the individual members, not as a set
  - Useful for ‘informational’ type queries where the user wants many relevant results i.e. old-style library subject search
- MRR - Mean Reciprocal Rank
  - Average of ‘reciprocal rank’ of ‘best result’

Relevance Academic Measures

- There are tons of references online
- Precision, Recall, F-1, MAP, P@X, MRR, and others I can't think of off hand.
- Some good starting points for older works include: Readings In Information Retrieval edited by Karen Sparck Jones, and Peter Willet published by Morgan Kaufmann, 1997

Relevance - How to Evaluate (data)

- “high-confidence set” - goal ensure ‘good results rank highly’
  - Full manual, or known-good results (i.e. Wikipedia/DMOZ)
  - Can include ‘popularly picked sites’
  - limitations: no ‘negative examples’, limited coverage for each query
- Full manual set - offer high coverage for a query
  - Can include both good and bad results
  - Can have multiple results labeled for a given query can include full ordering or multiple scores (i.e. more than “relevant”/”not-relevant”)
  - limitations: Set must be maintained - user judgments change over time, expensive, not easy to cover long-tail
- User behavior data - clickstream, retention data
- A-B test: Collect results from competing search engines

“Relevance” - How to Evaluate

- Absolute vs Relative
  - Am I better or worse than last month (product managers, ad-sales groups)
  - Do people like it enough to come back?
  - How do I compare to X
- Considerations:
  - Improvements or worsening “aggregate performance”
    - “best result” not ranking high enough
    - good results missing (could be crawler, indexing, or ranking)
    - Competitors evaluating better
    - How to meaningfully mine clickstreams (fewer clicks could mean you are doing better)
  - Test using ‘average people’ and have them compare
  - Retention rates, unique users (driven by marketing)

“Relevance” - Concerns

- Manager’s top question: “why did result X not rank for query Y?”
- Major “issues”
  - Too much spam making it through
  - Expected results ‘missing’ or not ranking highly
  - “Freshness”
  - Bad interpretation of query - i.e. “Jack Black” returning results about “black jack”
  - Drop in coverage - or worsening measures
    - Lower picks
  - Embarrassing examples

Relevance - What are the goals?

- Factors considered when performing evaluations:
  - What is the objective? maximize value to the ‘average user’, or a specific group of users
    - i.e. gear towards 18-34 Yr. old Males who like sports
- Cost - both to measure “relevance” as well as value of a customer
- Some goals:
  - Search Engines - maximize $$$ (this in general would mean market share, but could mean ad revenue)
  - “benevolent academics” - Show higher ‘performance’ on some measure
  - Other: maximize *perceived* value (play UI tricks, offer something unique)

Relevance - Challenges

- Independence? Search engines might apply some form of ‘mixing’ or ‘diversity’
- Not all users want the same thing
  - Use UI to reduce difficulty for ‘non-average’ user
- UI - both to allow users to see value, as well as to deal with ‘multiple meanings’
- Words on page vs “meaning” of page
  - How do you ‘score’ “click here to enter”
- Temporal relevance: q=iraq war --- 1991 vs 2007?

And the lecture moves on...

- Part III - Problems researchers in the ‘web space’ face, and what can we do about them?

Problems Facing Researchers (an example)

- My interview question in the area of web page classification

Problem statement:

An engineer builds a ‘home page classifier’. Inside this home page classifier is an oracle. The oracle knows the ‘correct answer’ for all web pages.

For each input, 99% of the time, the oracle correctly states if the input URL is a home page, 1% the oracle gives the wrong answer

Assumptions:

There exists a correct answer for all web pages

The Oracle always gives the same answer for the same URL

Problems Facing Researchers (example cont)

- The engineer takes sets of labeled examples and for all experiments correctly concludes that the system is exactly 99% accurate (balanced)
- 1000 home pages in, 990 labeled correctly
- 1000 non-home pages in, 990 labeled correctly

- This system is implemented on a real web search engine, as follows:
  - If a result URL is classified as home page, then the search engine draws a green box around it
- **Search engine users notice about every other green box is an error - explain.**

STATISTICS STATISTICS STATISTICS

- **statistics are everything**

- One of of the hardest things to grasp is that the success or failure of a particular algorithm can come down to the statistics. If the experiment and the application have different statistics, the "system" might not work.

- 99% "accuracy" might win a best paper award, but if the end user *sees* 50% accuracy you are fired!
- If only 1% of the web are home pages, then the number of false-positives will equal the number of false-negatives hence 50% precision

Five Important Classes of Problems Faced by Many Researchers

- Biases
- Bad assumptions, or statistics about users, queries, or web contents
- Insufficient or missing data
- Inconsistent evaluations or objectives
- Policies or external factors including resource limitations

Biases

- Always done this way before
  - remove stopwords, use bag-of-words, metrics, limited to a singular ML approach, ignore case and punctuation, stemming
- Previous problems that no longer apply
  - Typically resources constraints (i.e. low storage/CPU power)
  - Why not try all combinations? Why not score 10M result sets?
- “Natural Properties” used as artificial constraints/boundaries
  - Tending towards using data because of properties which are artificial
    - I.e. “anchortext” vs “anchortext windows”, “words” vs “character sequences”, English grammar vs ‘web-vernacular’
  - 32-bits to represent internal IDs (limit size)
- Source Data/Tools
  - Using wordnet because it is there, assume DMOZ is ‘perfect’

Biases

- Community/Academically ingrained
  - Num Features must be less than num examples
  - All errors are equal (a consequence of simplifying measures like precision and recall)
  - Entropy “always” picks the best features
  - If another group used this measure, that means it makes sense for me

Problem: Assumptions (Statistics)

- Common statistical mistakes
  - Forget about A-priori probability of a positive event
  - Overlooking “minor” but significant factors
    - Data all from one source/site, or data all of one ‘type’ (i.e. horror movie reviews, but no dramas)
    - User properties - All ‘researchers’, ‘educated’, using high-speed Internet?
  - Statistics of experiment not same as statistics of ‘target’ - i.e. test on TREC, design for ‘real web’
    - Real web pages don’t always have ‘text’
    - Real web probably more spam than non-spam

Problem: Assumptions

- Common assumptions
  - Small scale == large scale wrt algorithms and experiments
    - Pagerank on 10M graph from XYZ.edu is different than Pagerank on the second iteration of a breadth-first crawl of DMOZ
  - User judgments are stationary
    - query=Iraq, a web page about the 1991 US War judged as ‘highly relevant’ in 1991
    - query = “Clinton” - in 2001 it meant Bill, in 2007 Hillary
  - “Relevance” is directly connected to content of page
    - “click here to enter” as relevant to some query?
  - Ignore ‘undesirable users’ and ‘undesirable content’
    - Spammers adapt too... Old training data...
  - Filtering of bad content is easy, or a static list is sufficient
  - Recall doesn’t matter

Data (insufficient or missing)

- Researchers often complain of insufficient data
  - How to evaluate “relevance” without millions of labeled queries?
  - Pagerank computation on small graphs
  - Classification without large training/labeled sets
- “Real” query logs often not available to researchers
  - Or old lists mess up calculations
- Crawls of large sites as opposed to a few random pages
- Important data is restricted or expensive
  - DNS data, user demographics
  - Free “manual lists” could be of low quality or unavailable
  - How to get ‘current news feed’?

(Inconsistent) Evaluations and Objectives

- Is the goal 99% on some test set, or users “think” the system is accurate?
  - Biggest problem is improper mapping from measure to real purpose
- It all depends on the application
  - Highlighting all predicted homepages is very different than accuracy on a balanced set
- Maybe 75% is actually good enough? Does Recall matter? Who are the users?
- Independence assumption on your evaluation, but not in reality (or vice versa)
- Not all errors are equal
- Inferring conclusions on a statistically different data

Evaluation Example

- Which engine is better?
- Engine A
  - 85% of the time the ‘best result’ is ranked at #1, and the remaining top 9 are all totally random
  - 15% of the time all ten results are random
- Engine B:
  - 80% of the time the ‘best result’ is ranked at #1, with the other top 9 as mediocre (weakly relevant)
  - 20% of the time the “best result” is ranked #9, other 9 mediocre
- Engine C: Same as B, but 19.9% “best result” is #10
- Engine D:
  - 90% of the time the ‘best result’ is ranked at #1 with the top 9 very good, but 10% of the time all 10 results are offensive

Policies and External Factors

- Google has tens of thousands of machines, what can be done with just one?
- Are you “allowed” to crawl fifty pages per second?
- ERB not allow collection of personal user data?
- Can’t access the local-server logs?
- Artificial requirement to utilize a specific dataset?
- Other? Can only crawl on weekends?
- Indexing tool is required and does not support position information

Dealing With Problems/Challenges

- Most important is understand the “real problem” with minimal biases or assumptions
- General Approach
  - Solve a different but ‘close enough’ problem
  - Find a different problem (or different way to state the given problem) that maps to easier or more reasonable solutions
  - Consider “divide and conquer”
    - Split the “problem” into many separate problems whose combined solution addresses the ‘real problem’
- Always try to:
  - Use statistics to your advantage if 1% is bad, then operate in the 99% space
  - Know your biases and try to use data and user biases positively
    - Assuming users have a bias favoring A train more A

Dealing With Problems an Example

- Consider the homepage classifier example from before - with an oracle that is 99% balanced accuracy.
- The problem (as defined by a PM, not a researcher):
  - Clearly identify to users homepages present in search results
- We can start with: A 99% balanced accuracy classifier - which fails due to the low A-priori probability of being a homepage
- Approach: Define a different sub-problem:
  - **Alter the statistics** so 99% balanced works (i.e. raise the A-priori probability of a homepage
  - Possible method: A very simple classifier/filter which removes the ‘obviously not homepages’

How To Improve Things

- Go through each ‘problem’ and make them work to your advantage - yes really!
- Know your biases - apply them correctly, or break out of bad habits
  - “Not All Errors Equal” can work to your advantage
    - Adjust ‘weight’ of training/evaluation errors
    - Adjust problem definition
    - Focus more effort on the ‘serious errors’ (fewer of them means easier methods might open up)
  - Minor changes can reduce artificial constraints
    - Why remove stopwords, why not remove ‘features which add low value due to high frequency’
    - Improved features can dynamically redefine ‘stopwords’
    - Example “the” in title is a negative feature for homepage classification

How To Improve Things (cont)

- When statistics don't work:
  - change the stats or redefine the problem
  - Know the world, and consider "it" throughout the entire process from problem definition, to methods, to implementation
  - Specialize or constrain your problem, be explicit
  - Multiple methods have different statistical implications
  - Split the 'bad statistics'
    - Maybe in one sub-space has 'good statistics'
    - Try to find good sub-spaces and separate the 'bad'
    - If you can't solve for all, then don't
      - It might be enough to identify the bad region

How To Improve Things - cont...

- Missing data? Insufficient data?
  - **First - be aware of the statistics, don't approximate without considering the effects**
  - Don't blindly try to get the 'same thing' - solve the problem
    - I.e. looking for 'real user queries' don't collect 'bad user queries'' simply because they are queries that came from users
  - Use proxy-data - can't collect user queries, find a data source and 'hack it' - Wikipedia titles represent 'important concepts'
  - Be warned - 'logs' may be edited, old, not-representative, include errors, missing frequency information
  - Know what you want the data for, maybe you don't need it

How To Improve Things - cont...

- Don’t make unrealistic assumptions, for your true goals
  - If you must make unrealistic assumptions, don’t go found a company based the solution
- Remember what matters to a search engine or to the ‘consumer’
- Don’t feel obligated to use a specfic method because everyone else does (if you do, use other more meaningful measures too)
- Solve the ‘right problems’ - or redefine problems so the assumptions are reasonable, and the evaluations will match
- Consider all evaluations and objectives ‘in context’
  - What does 86% accuracy mean? Does the evaluation represent what you want?
- If you evaluate as an academic your paper will work for academia
  - Don’t ignore existing methods, understand them and enhance them

General Advice

- I recommend you understand the ‘gist’ of each method, each piece of data each goal
- THEN decide how to use them to your advantage!
- Be creative for your problem definition: it is easier to find ‘not homepages’ than homepages’
- Be creative in your evaluation - but be sure it is meaningful!

Data Sets (Public)

- Where can researchers go to get “data”
  - NOTE: Before using these sets, please consider what they are and what they are not. I do not personally comment on the value or quality of these sets.
- Popular sets: TREC “Text REtrieval Conference (has large web content)
  - http://trec.nist.gov/data.html - lists all the TREC data sets
  - TREC Web data: http://ir.dcs.gla.ac.uk/test_collections/
- “relevance sets”
  - LETOR (LEarning To Rank) http://research.microsoft.com/users/tyliu/LETOR/default.aspx
- Sites with many different data sets (web/IE/classification/QA)
  - http://www.cs.umass.edu/~mccallum/code-data.html
    - above includes links to CORA - research paper set
  - http://kdd.ics.uci.edu/summary.data.date.html
    - links to multiple data sets

Data Sets (cont)

- Specialized data sets:
  - Wikipedia can be downloaded - http://en.wikipedia.org/wiki/Wikipedia:Database_download
    - There are also “external links” off of Wikipedia pages
  - Reuters Data set - useful for classification, IE, categorization: http://trec.nist.gov/data/reuters/reuters.html
  - Citeseer (academic papers BiBTex): http://citeseer.ist.psu.edu/oai.html
  - Open Directory (DMOZ) RDF dump:
    - http://rdf.dmoz.org/
  - WebKB dataset
    - http://www.cs.cmu.edu/~webkb/
  - AOL query logs (you need to search to find these)
  - ENRON (email) dataset: http://www.cs.cmu.edu/~enron/
  - WebSpam: http://www.yr-bcn.es/webspam/datasets/

Commercial Plug

We’re hiring!

If interested find me, and or contact:
jobs@searchme.com