Presentation to the 17th Information Quality Conference

Mark Bennett – mbennett@ideaeng.com
VP of Engineering
New Idea Engineering, Inc. – http://www.ideaeng.com

“Poor Data Quality Gives Search Engines a Bad Rap”

September 20, 2005

new idea ENGINEERING

http://www.ideaeng.com

Agenda

- Part 1: Search Engines vs. Databases
- Part 2: Index Creation Data Quality Issues
- Part 3: Search-time Quality Issues

- Discussion / Q&A

new idea ENGINEERING

http://www.ideaeng.com

“Poor Data Quality Gives Search Engines a Bad Rap”

Part 1: Databases vs. Full-Text Search Engines

More alike than different…

new idea ENGINEERING

http://www.ideaeng.com

Search Engines & Databases: Similar Technologies

- Similarities:
  - Search through large volumes of data
  - Return matching records
  - Have a specific query syntax

- Search Engines differ in:
  - Mostly read-only
  - Weighted matches
  - Very different “indexing” procedures
  - No “joins”

new idea ENGINEERING

http://www.ideaeng.com

Search Engines & Databases: Different Terms for the Same Thing

Databases	Search Engines
a “database”	Collection, Document Index or Catalog
Table	Segment or Partition*
Record	Document, Page, URL, Record, Hit
Field	Field, Doc Field, Meta Data/Field, Zone
Blob	Zone
Index (verb)	Indexing, **Spidering**, Crawling
Index (noun)	Collection, Doc Index

new idea ENGINEERING

http://www.ideaeng.com

Search Engines & Databases: “Indexing” and “Spidering”

- Web Content:
  - Spider / Crawler
  - Can also crawl a file system
- Non-Web Content
  - Uses a “Gateway” or “Connector”
  - Search Engine API
  - Other “Indexing Process”

new idea ENGINEERING

http://www.ideaeng.com

Understanding Search Engine "Spiders" used for Web Content

- Unlike traditional databases…
  - Search Engines can't just "load" or "import" their data
  - They need to go out and find it!
- Roots
  - The spider is given a few seed URLs to start at
  - It then follows the links on those pages to find other pages
- Include and Exclude patterns
  - Spiders have rules for which links they should follow and index
  - For large sites these rules can be rather complex

new idea ENGINEERING

http://www.ideaeng.com

“Poor Data Quality Gives Search Engines a Bad Rap”

Part 2: Indexing Issues

Getting the right data into the Search Engine

new idea ENGINEERING

http://www.ideaeng.com

Indexing Issues Overview

- Basic Indexing
  - All data being indexed?
  - Checking site growth and content
  - Monitoring the spider process
- Thorough Site Audit
  - When every document counts
- Document Meta Data
  - Basic fields
  - Vertical applications

new idea ENGINEERING

http://www.ideaeng.com

Index Data Quality: Basic Indexing

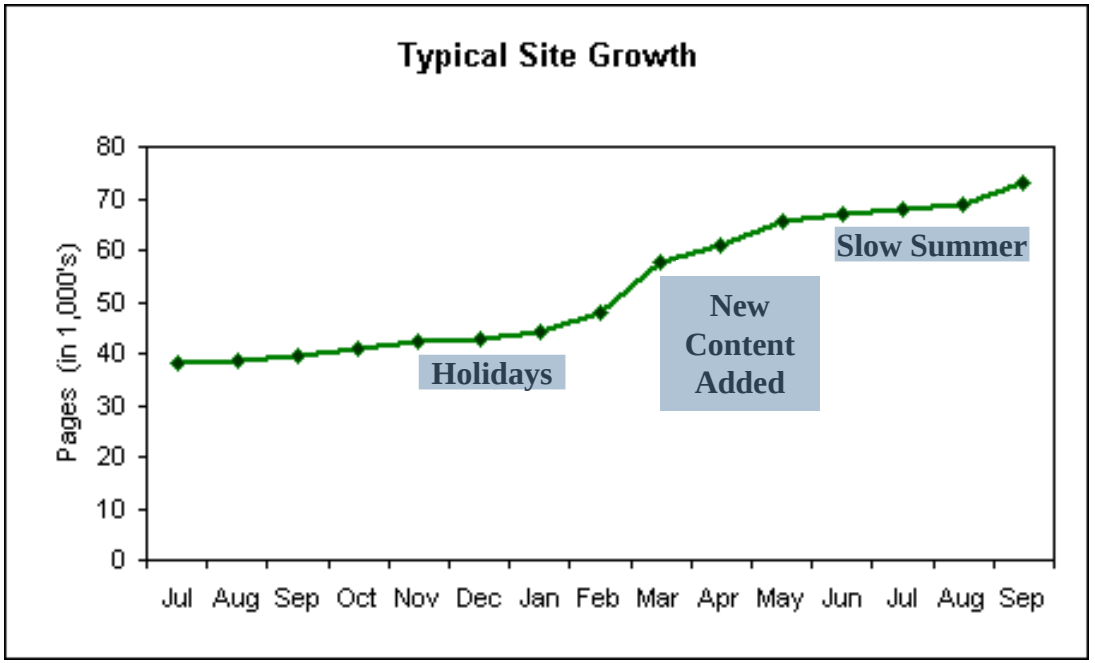
- Is your entire site being Spidered and Indexed?
  - How many documents does your search engine say you have?
  - Compare this count to other sources
- Getting a ballpark estimate from another source:
  - Public site: compare with Google or FreeFind
  - Intranet sites: try with another spider
  - Intranet files: compare with filesystem info
  - Database driven: compare with number of records

new idea ENGINEERING

http://www.ideaeng.com

Index Data Quality: Overall Site Growth

- Is your entire site being Spidered and Indexed?
  - Check the Total # of docs your search engine knows about
  - Capture and track this number over time to spot problems

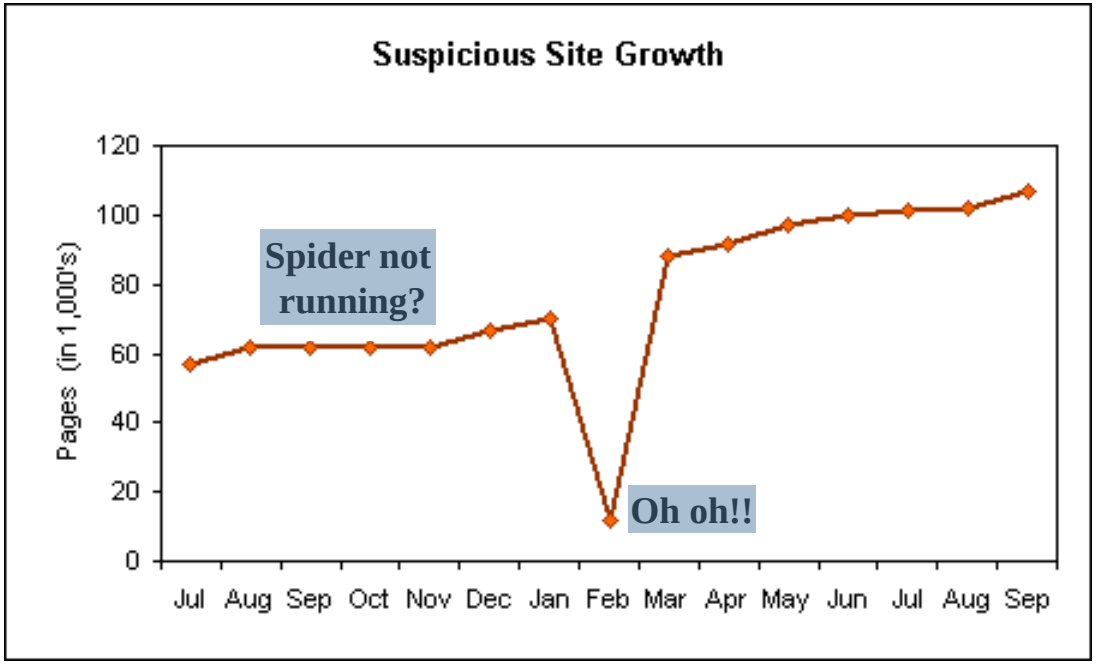


new idea ENGINEERING

http://www.ideaeng.com

Index Data Quality: Overall Site Growth

- Trend lines show problems, look for:
  - Radical changes in page count
  - Zero change in page count - is the spider even running?



new idea ENGINEERING

http://www.ideaeng.com

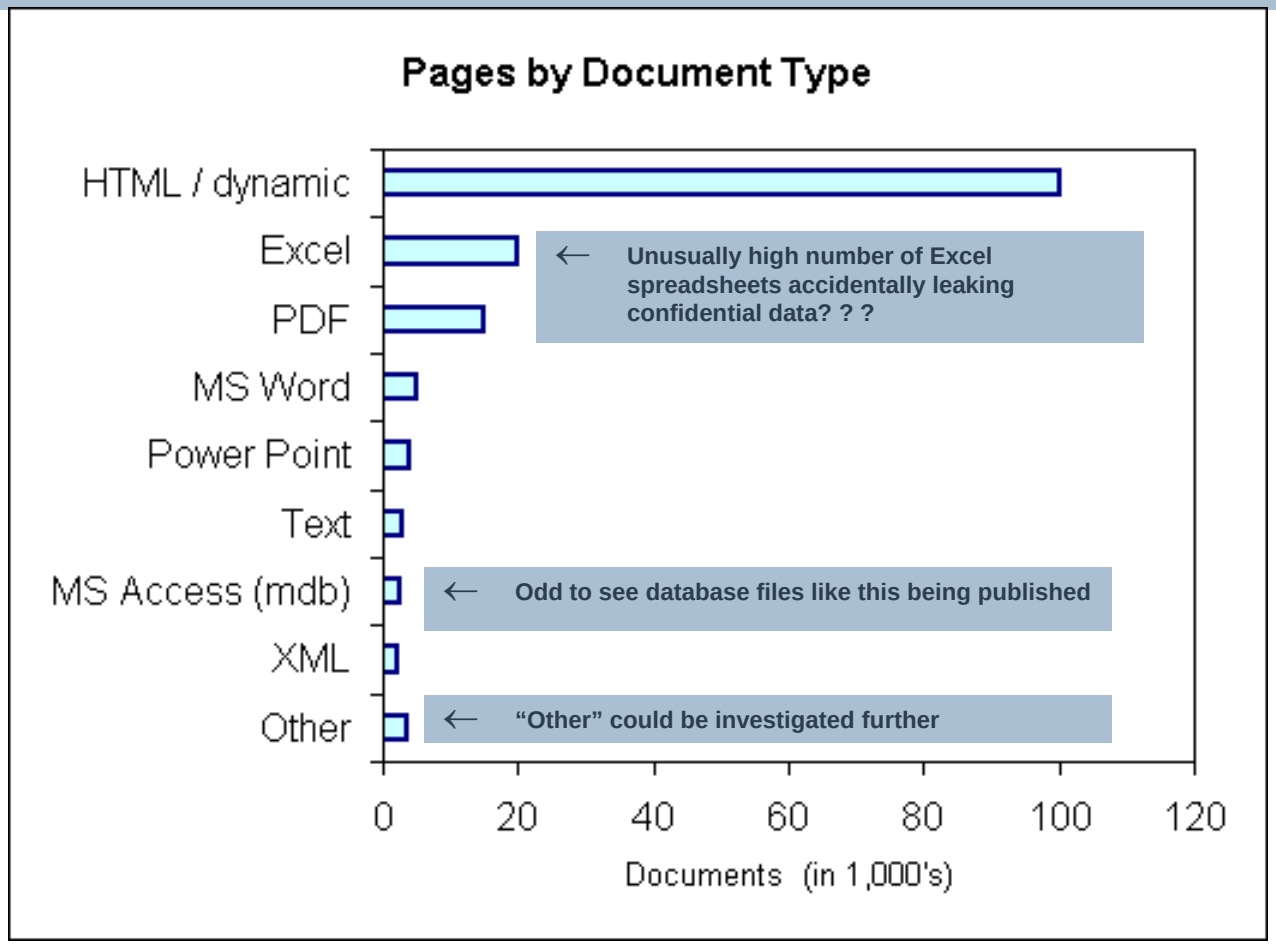
Index Data Quality: Other Basic Indexing Checks

- Are your indexes up to date?
- Check index process logs for errors?
  - How long would it take you to notice if indexer was failing!?
- Add automated checks to your index scripts
  - Track how long spidering / indexing usually takes
- Have your spider log URLs that were and were not indexed

new idea ENGINEERING

http://www.ideaeng.com

Index Data Quality: Check by Document Type



new idea ENGINEERING

http://www.ideaeng.com

Index Data Quality: Doing a Thorough Audit

Going beyond simple document counts

- Consider a Thorough Audit:
  - Compare actual URLs and/or document keys
  - Dump URLs from search engine index
  - Dump expected URLs from database or “find” script
  - URLs may need to be normalized

new idea ENGINEERING

http://www.ideaeng.com

Index Data Quality: Document Meta Data / Fields

- Common Examples:
  - Title, Date, Author, Source and Summary
- *Vocabulary:* ***Vertical Search Application***
  - Specialized search application with large amounts of text
  - Often used by expensive “Knowledge Workers”
- Vertical Application Meta Fields
  - Examples: part #, customer #, model #, version, SSN, etc.
  - Often similar to traditional database apps
- Did the Search Engine see your Meta Data?

new idea ENGINEERING

http://www.ideaeng.com

Document Meta Data: Document Dates

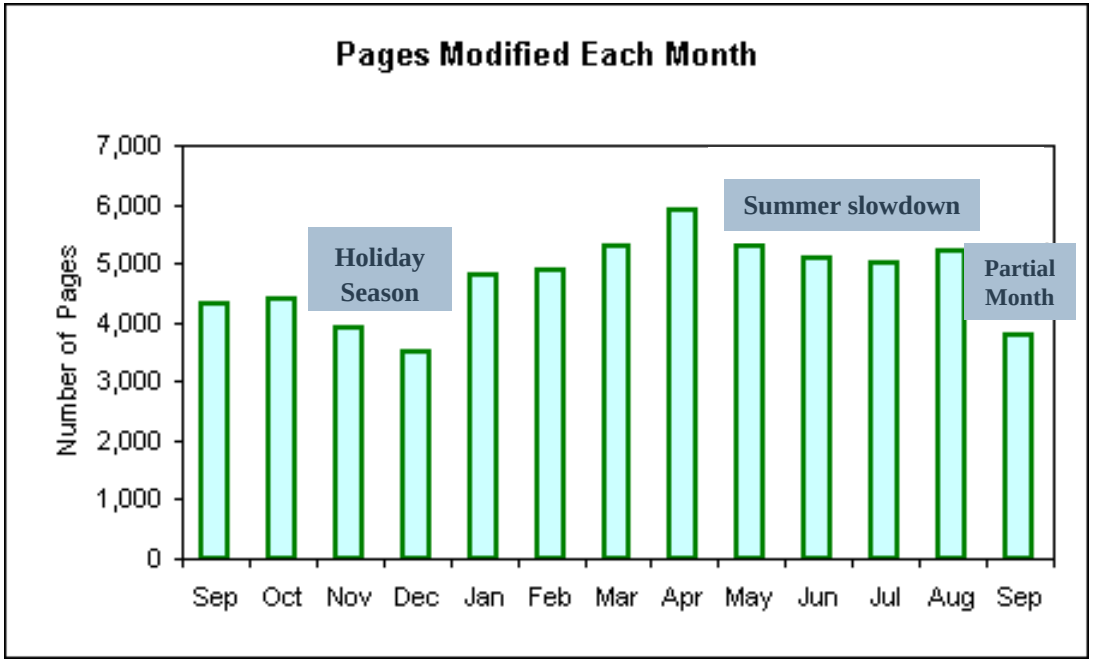
- **Should reflect the creation or last modification of that information**
  - Do your dates look correct?
- **Quick check for problems in the results list:**
  - Blank dates
  - All docs have today's date, or other recent date
- **Thorough Audit**
  - Dump all dates to a file or database and graph them

new idea ENGINEERING

http://www.ideaeng.com

Document Meta Data: Distribution of Document Dates

- Look at the distribution of dates the search engine has stored for each time period

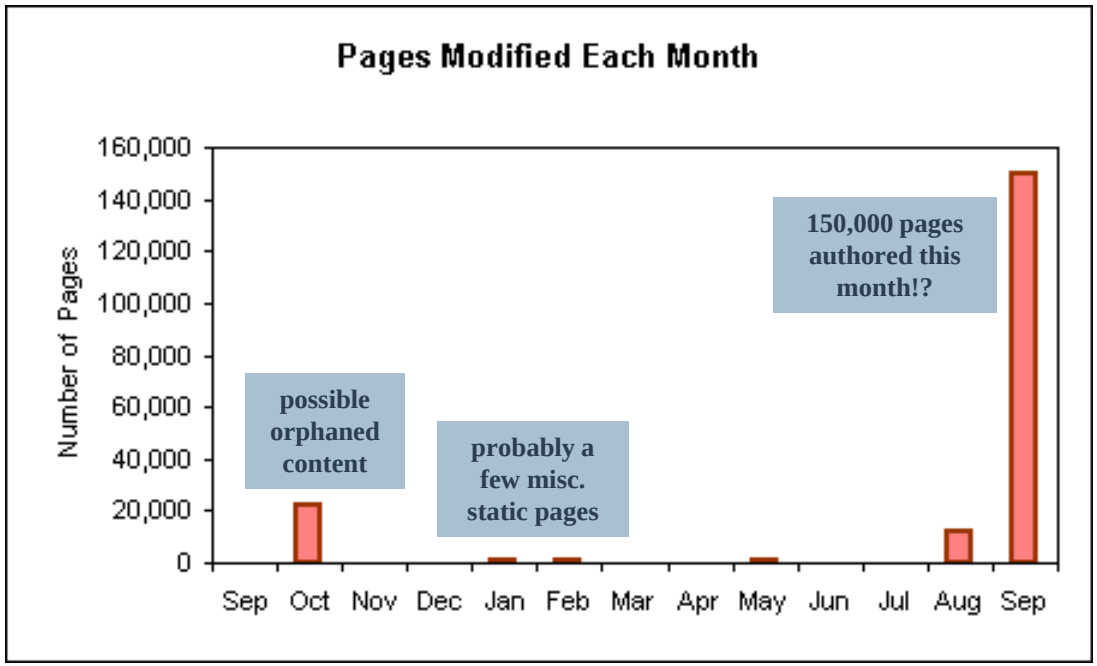


new idea ENGINEERING

http://www.ideaeng.com

Document Meta Data: Incorrect Document Dates

- Search engines often have this wrong
  - May need to tweak the web server's settings



new idea ENGINEERING

http://www.ideaeng.com

Document Meta Data: Document Titles

- Normalize titles before doing analysis
- Signs of trouble:
  - Red flag: Null, empty or gibberish titles
  - Too short or too long
  - Duplicate titles (some may be legit)
- Special case: Titles with long common prefixes
  - Sample problem:
    - Acme Online Customer Support: Frequently Asked Question: *rest of title…*
  - Improved:
    - Acme FAQ: *rest of title…*

new idea ENGINEERING

http://www.ideaeng.com

Document Meta Data: Site Specific Fields

- Vertical Applications often have important, custom document meta-data
  - Are you using it? Should you be?
  - Do you have a master list of fields?
- Signs of trouble:
  - Missing / empty meta data
  - Review list of unique values for each field
  - Not normalized (format, white-space, case)

new idea ENGINEERING

http://www.ideaeng.com

“Poor Data Quality Gives Search Engines a Bad Rap”

Part 3: Search Results

Understanding and Adjusting Search Engine Results

new idea ENGINEERING

http://www.ideaeng.com

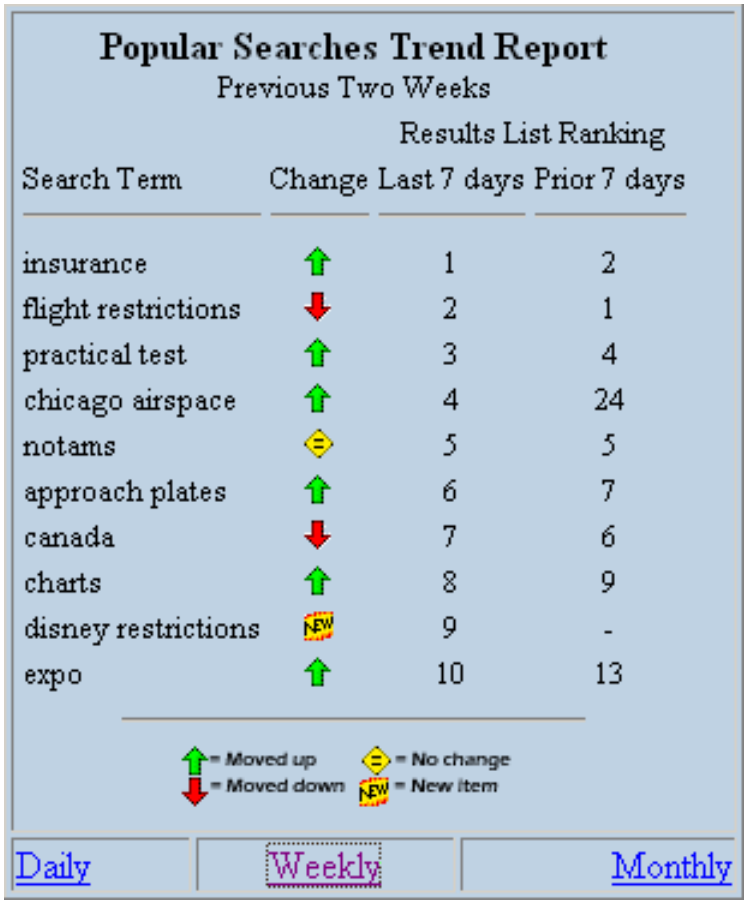
Search Results DQ Overview

- Your Top Searches
  - Returning good results?
  - Checking the Search Engine's scoring
- Effectiveness
  - Click-through
  - Users' Vocabulary
- Other Issues
  - Reliability
  - Performance

new idea ENGINEERING

http://www.ideaeng.com

Search Results Data Quality: Initial Spot Check



Popular Searches Trend Report

Previous Two Weeks

Search Term	Change	Results List Ranking Last 7 days	Prior 7 days
insurance	↑	1	2
flight restrictions	↓	2	1
practical test	↑	3	4
chicago airspace	↑	4	24
notams	=	5	5
approach plates	↑	6	7
canada	↓	7	6
charts	↑	8	9
disney restrictions	NEW	9	-
expo	↑	10	13

↑ = Moved up ↓ = Moved down = = No change NEW = New item

Daily | Weekly | Monthly

- Run your top 10 searches:
  - Do the documents returned seem relevant?
  - Can you think of better documents to return?

new idea ENGINEERING

http://www.ideaeng.com

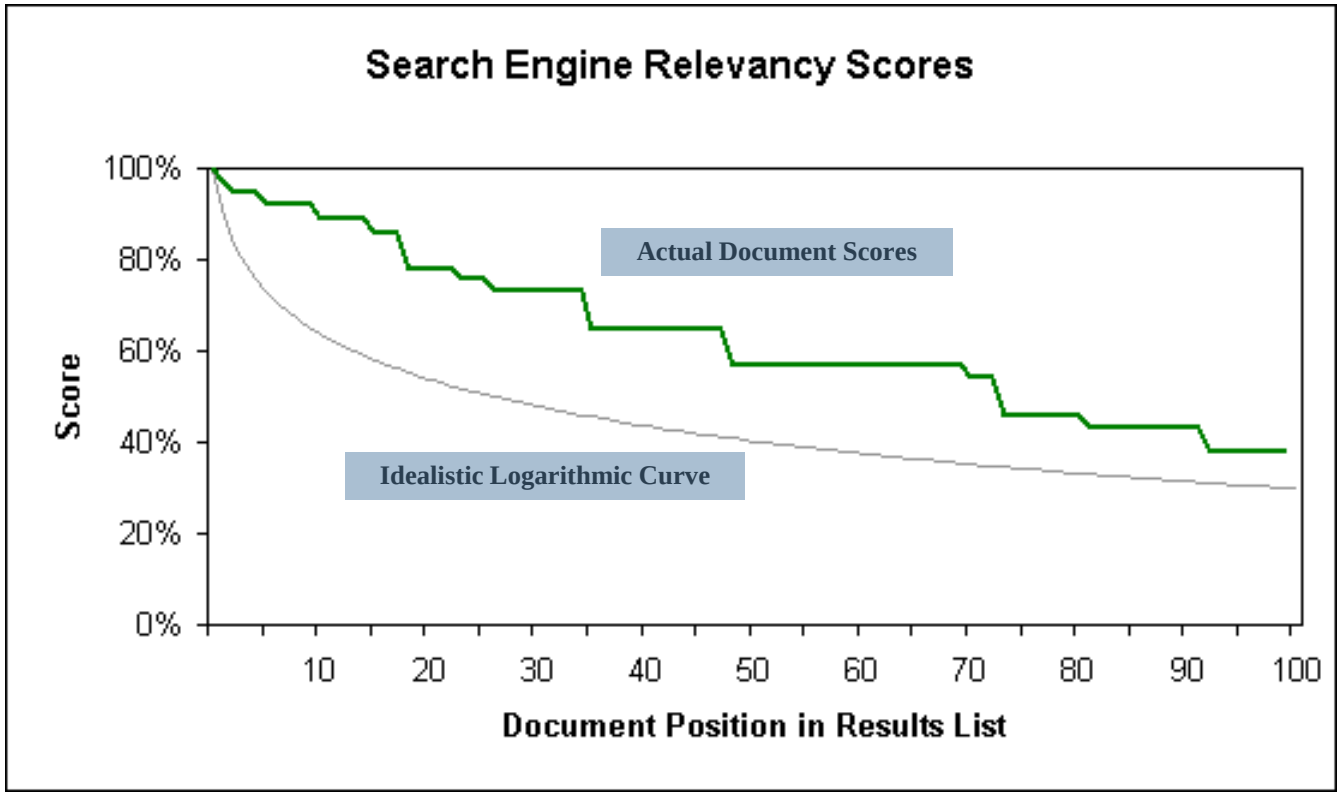
Search Results Data Quality: Relevancy Histogram of Matching Docs

- Most search engines can provide some type of “score”
  - Usually a percentage
  - May be useful in checking search engine ranking
- Remember:
  - Calculated relevance doesn’t always correspond to perceived relevance; the latter is much more important
  - Checking your scores doesn’t mean you have to display them to users; our advice is to not display them to users.
  - Just a guideline, one of many items to consider

new idea ENGINEERING

http://www.ideaeng.com

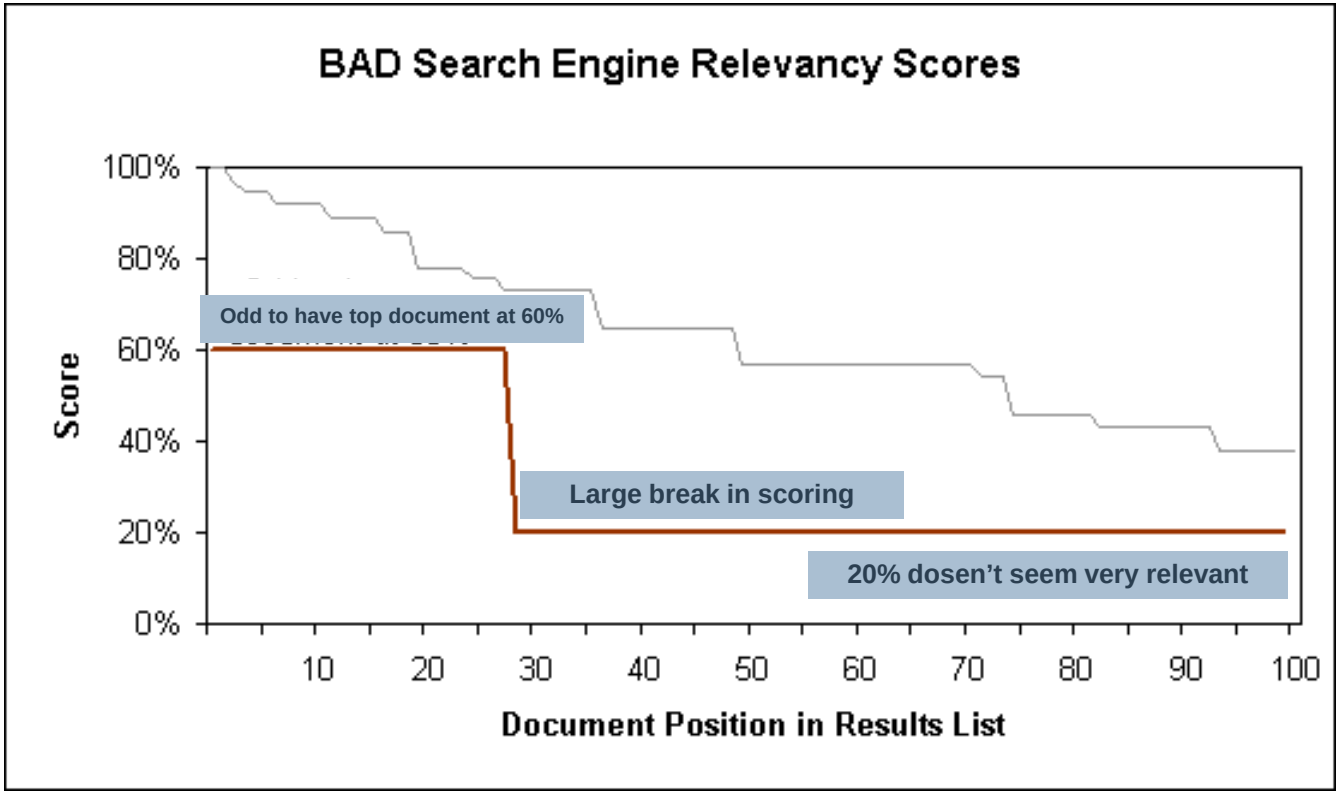
Search Results Data Quality: Relevancy Histogram of Matching Docs



new idea ENGINEERING

http://www.ideaeng.com

Search Results Data Quality: Bad Relevancy Histogram



new idea ENGINEERING

http://www.ideaeng.com

Search Results Data Quality: Click-Through and User Vocabulary

- Are users clicking on the top 1 or 2 documents returned by each search?
  - Or are they clicking the 4th or 5th document down?
- Vocabulary Issues
  - Do visitors use the same wording as your content?
  - Use your vendor's Thesaurus feature
  - Use Directed Results to suggest better answers

new idea ENGINEERING

http://www.ideaeng.com

Search Results Data Quality: Reliability and Performance

- Check search activity for serious problems
  - Red flag: searches with no results
  - Yellow flag: searches returning > 10% of site
  - Red flag: a sudden **drop** in searches per day
    - Consider a “ping” script to periodically run a known search
  - Yellow flag: a sudden **increase** in searches per day
  - Yellow flag: many searches from same IP address or domain
  - Yellow flag: searches taking > 2 seconds
  - Red flag: searches taking > 5 seconds

“Poor Data Quality Gives Search Engines a Bad Rap”

Wrapping Up

new idea ENGINEERING

http://www.ideaeng.com

Summary and Action Items

- Basic Indexing
  - Is your spider running? Getting errors?
  - Is your site index complete? Is it growing?
- Advanced Indexing
  - Consider document-by-document Audit
  - Check your Meta Data: dates, titles, etc
- Search Results Quality
  - Check your Top Searches and Click-through rates
  - Look at users' vocabulary vs. your content
  - Reliability and Performance

new idea ENGINEERING

http://www.ideaeng.com

Resources

- Links
  - Lucene is an open source search engine; very educational regardless of the engine you use in production.
  - **http://lucene.apache.org**
- Books
  - Mining the Web: Discovering Knowledge from Hypertext Data
  - Lucene In Action: Guide to the Java Search Engine
- NIE Enterprise Search Newsletter
  - **http://www.ideaeng.com/subscribe**

new idea ENGINEERING

http://www.ideaeng.com

Discussion

Q & A

Follow up questions:

Mark Bennett

mbennett@ideaeng.com

(408) 733 – 0387

new idea ENGINEERING

http://www.ideaeng.com