

Using Search Engines for Data Discovery on Intranets

Avi Rappoport, Search Tools Consulting

www.searchtools.com

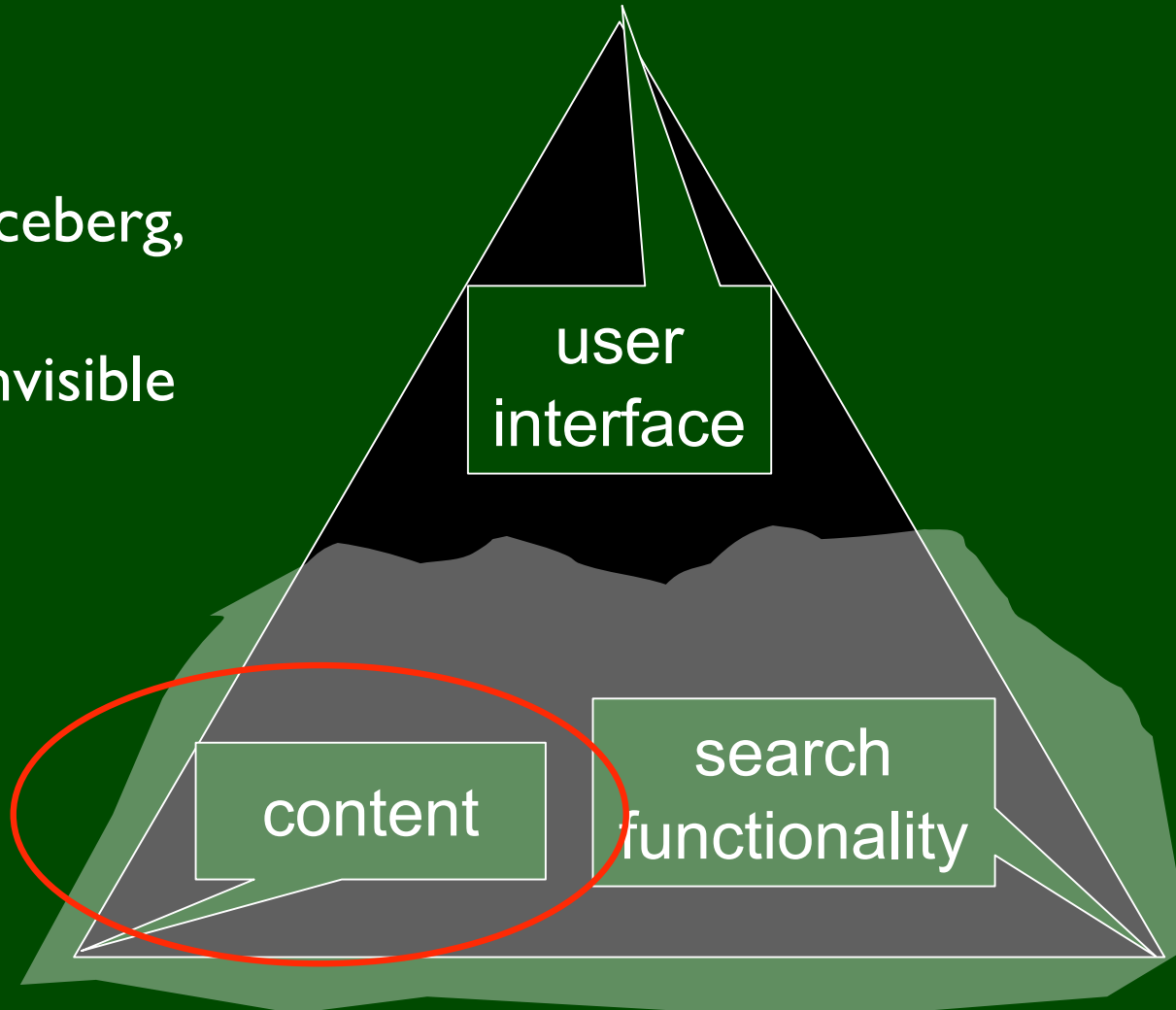
analyst5@searchtools.com

Why Use Search for Data Discovery?

- Intranets tend towards chaos
- Portals & CMSs can't keep up
- Search is often the only access to content
- Better than content inventory tools?
 - Search spiders are the best
 - Deep access to content
 - Integration with search and taxonomies

Three Parts of Usable Search

Like an iceberg,
search is
mostly invisible



Iterative Process

- Turn on spider and see what happens
 - Keep per-server load low
 - Enable verbose reports and logs
- Cast a wide net
- Keep copious notes
- Repeat several times

Spider-eye View of the Intranet

- List of servers
 - Some clearly departmental
- List of pages and documents
 - How many with titles?
- Metrics
 - File types: HTML, text, PDF, Word, etc.
 - Server protocols
 - Character sets and languages

Technical Issues

- Some items should be there but aren't
 - Problem links: bad syntax, JavaScript
 - Disability access issues
 - Redirects and aliases
 - Graphical text, funky PDFs
 - Missing servers
 - Changed name
 - Server crashed or removed
 - Wrongly configured robots.txt

Unauthorized Content

- Obsolete
 - Pre-merger rules
 - Old product price lists
- Hijacked servers
 - Spam services
 - Software piracy
 - Music or video sharing
 - Porn

Security Failures

- Insecure server software
 - Vulnerable to viruses and spyware
- Unauthorized content
 - HR information
 - Confidential plans
 - Financial records
 - Partner data
 - Client data

Link Analysis

- Incoming Links
 - Authoritative pages
 - Add to portal
 - Increase relevance weight
 - Watch for errors
- Outgoing Links
 - Hub pages, sometimes flag problems
 - Indicate useful external content

File Attributes and Facets

- Format (extension and MIME types)
- Character sets and languages
 - Explicit
 - Derived from analysis
- Metadata
 - Explicit and implicit
 - Anchor text content
- Content types
 - Forms, releases, calendars, contracts

Duplicate Detection

- Entire duplicate servers
- Identify duplicate documents
 - Checksum is easy
 - Title
 - Similarity metric
- Mark current or definitive version
- Reduce server & search load

Classification and Entity Extract

- Leverage effort of opening documents
- Classification
 - Define "aboutness", cluster by similarity
 - Update training sets and rules
- Entity Extraction
 - *People, places, and things*
 - Job codes, abbreviations, chemical names
- Create and store metadata
 - No need to re-build on the fly

Regulatory Compliance

- Sarbanes-Oxley
 - track business decisions and transactions
 - retain documents on the process of gathering and calculating financial data
- HIPAA and others
- Use both classification and search tools
 - Identify documents
 - Get *actual* dates

Iterations and Maintenance

- Adapt to changes in Intranet scope
- Identify problems, missing data, servers
 - Help intranet publishers
- Resource for content management efforts
- Identify new sources of valuable content
 - Work with portal creators
- Reduce future overhead

Intranet Inventory

- Definitive list of servers and documents
 - Reduce duplication
 - Rules for date & language deduction
 - Track core content
- Identify server problems, access control
- Reports
 - Metrics
 - Publishing and training priorities
 - Topics

Integrate with Search & Taxonomy

- Identify most important content
 - Link analysis
 - Use for search relevance weighting
- Offer reliable document dates
- Remove duplicates
- Store metadata
 - Entities, attributes and topic identification
 - Support faceted metadata

Search for Data Discovery

- Dynamic view of the Intranet
 - Inventory and metrics
 - Identify unauthorized and exposed content
 - Find duplicate servers and documents
 - Regulatory issues
- Create and maintain taxonomies
- Improve search index and relevance
 - Core vs. archive content
 - Facets!