Context
What is Enterprise Search?
How is it different from Internet search?
Content
User interface
Language
Search context
Results
Search Logs
Some context

- Largest not-for-profit HMO in the US
- 8.2 million members
- About 145,000 employees, plus 15,000 or so doctors
- 8 Regions
- About 35 Unions
- 30 hospitals
- 431 medical office buildings
- Operating revenues of $22.5 billion
Enterprise search: definition

- Search implemented within a business context for internal users
  - Index internal content
  - Serve internal clients and end-users
- As opposed to companies “optimizing” search on the Internet, which is more of an external, marketing and PR function
  - Buying ads
  - Optimizing content
- I am going to talk specifically about internally-facing search for employees – not the public site
Technology is not really the issue

- It’s tempting to assume that “the best” search engine is the answer to all search problems.
- There are many enterprise search products on the market, each has its strengths and weaknesses.
- It’s important to know the business climate of the organization when choosing one.
  - Content
  - Users – end users and business clients
  - What capabilities users expect
  - Capabilities users need
  - Organizational commitment to search
Enterprise search: how it’s different

- Search is not the business of the enterprise
  - Most organizations do not invest any time in search
  - As opposed to companies whose life-blood is improving search
- User task is different
  - Internet: find an answer or a starting point
  - Intranet: find the right answer, know it’s the right answer
- Statistical data is not there
  - Fewer searches per day
  - Less content
- In-links are not there
  - Important in improving relevancy on the Internet
My job is about

- Serving many individual business clients, each concerned about their own content and priorities
- Stepping back and advocating for
  - The end user
  - Organizational information strategy
  - Constant improvement vs. project focus
- Devising strategies to get the right information to the user
- Leverage search to make the organizational complexity transparent to the user
Our biggest challenge: Content

- Clutter
  - Many of the same problems as the Internet – where the solution is to improve ranking
    - Lack of in-link analysis really hurts this for an Intranet
    - We have to come up with different strategies

- Disparate repositories
  - Different technologies (databases, DMS)
  - Managed by different groups
  - Different navigation and metadata schemes

- “Over the wall” attitude
  - Content is our job, search is your job.
    - Not always a lot of interaction between search team and content owners

- “Wild west” – no centralized strategy
Some statistics

- About 600 Intranet sites
- Number of documents almost doubles each year
Document Metadata

Bar chart showing the distribution of document metadata in various file formats such as MS Word, Adobe PDF, MS Excel, MS PowerPoint, MS Project, HTML, Text, and XML. The chart measures the percentage of documents containing titles, uniquely titled description, keywords, and subjects.
Security

Five cases

- The content shouldn’t be in the search indexes at all
- The content should be in the index, but only be returned for certain individuals
- The content should be returned in results, but different information should be shown in the results for different individuals, and access to the full document should be restricted
- The content should be returned in results for everyone, but gaining access to the full document should be restricted
- The content should be returned and accessible to everyone
The core task: finding answers

- Can users find the search box?
- Are we indexing the right information?
- Are we protecting information that should not be searchable?
- Most searchers enter only one or two-word queries, what can we do on the backend to improve that query before sending it to the search engine?
- What search terms are people using that are not returning results?
- What search terms are people using that are returning too many or inappropriate results?
- Once the results are returned, what can be done to categorize them, describe them, rank them.
- What next steps can we offer the user if their initial results do not answer their question?
Can the user find the search box?

- Inconsistent placement.
- Size – text box usually too small to accommodate a reasonable query.
- Wording around search box/button is inconsistent.
- Text entry boxes are “ugly” – designers want them to be as small and inconspicuous as possible.
“White box” confusion

- There are so many white boxes out there…
- Our top three searches are for Google.com, Yahoo.com and msn.com
  - Anecdotally, this is very common on Intranets
  - Might indicate that users don’t necessarily understand the distinction between the Intranet and Internet
  - Or perhaps even the difference between the browser and the page
- Makes sense really
  - Many users navigate the Web with search (rather than using the address bar)
  - Many browsers throw unknown addresses to a search
Language

- All the usual problems
  - The terms a searcher uses are not the same as the words the author used
  - Term ambiguity (bank)
  - Euphemisms

- Especially bad in an enterprise setting
  - Misspellings: especially in a medical domain
  - Use of acronyms: frequently a many-to-one phenomenon
  - Jargon: doesn’t translate between departments or frequently within departments!
Query Interpretation and Expansion

- Query expansion through thesaurus
  - We have found that there can be as many as 7 meanings for the same acronym! Do we include all of them? Or only in specific contexts?

- Dropping stop words (words with no meaning, the, these, etc)

- Interpreting Boolean: OR
  - Is it Boolean?
  - Or an abbreviation for Operating Room or Oregon?
  - (it’s most often the latter)
Context – what did I search?

- Partly a white box problem, who knows what’ll happen when you hit enter.
- Since some sites search the whole Intranet (or Internet in some cases), and some just search “themselves”, it’s hard to know what you’re going to get back.
- Even if we report “your search of Biomed National returned…”
  - Do searchers know what that means? (If they see it?)
- Sometimes the content just isn’t there.
  - Because there are “off-site” content sources that we don’t know about yet (Docushare, LN databases, etc.)
Results presentation

- How should documents be ranked?
  - By relevance alone?
  - What about by recency?
  - What about by popular sites?

- Results from different content sources
  - Display together or apart?

- Next steps
  - Narrowing (faceted browsing)
  - Expanding
  - Sorting
Informing decisions: search logs

- We’ve learned from our logs
  - Content users were looking for that we didn’t offer (phonebook, acronyms, best bets)
  - The character of queries – length, number of words
  - How often particular queries occur

- How search logs benefit site information architecture
  - Content & features that should be added
  - What terms people are using – you can add them to metadata
  - Where users searched from
  - What searchers are not able to find through navigation
Search touches many areas