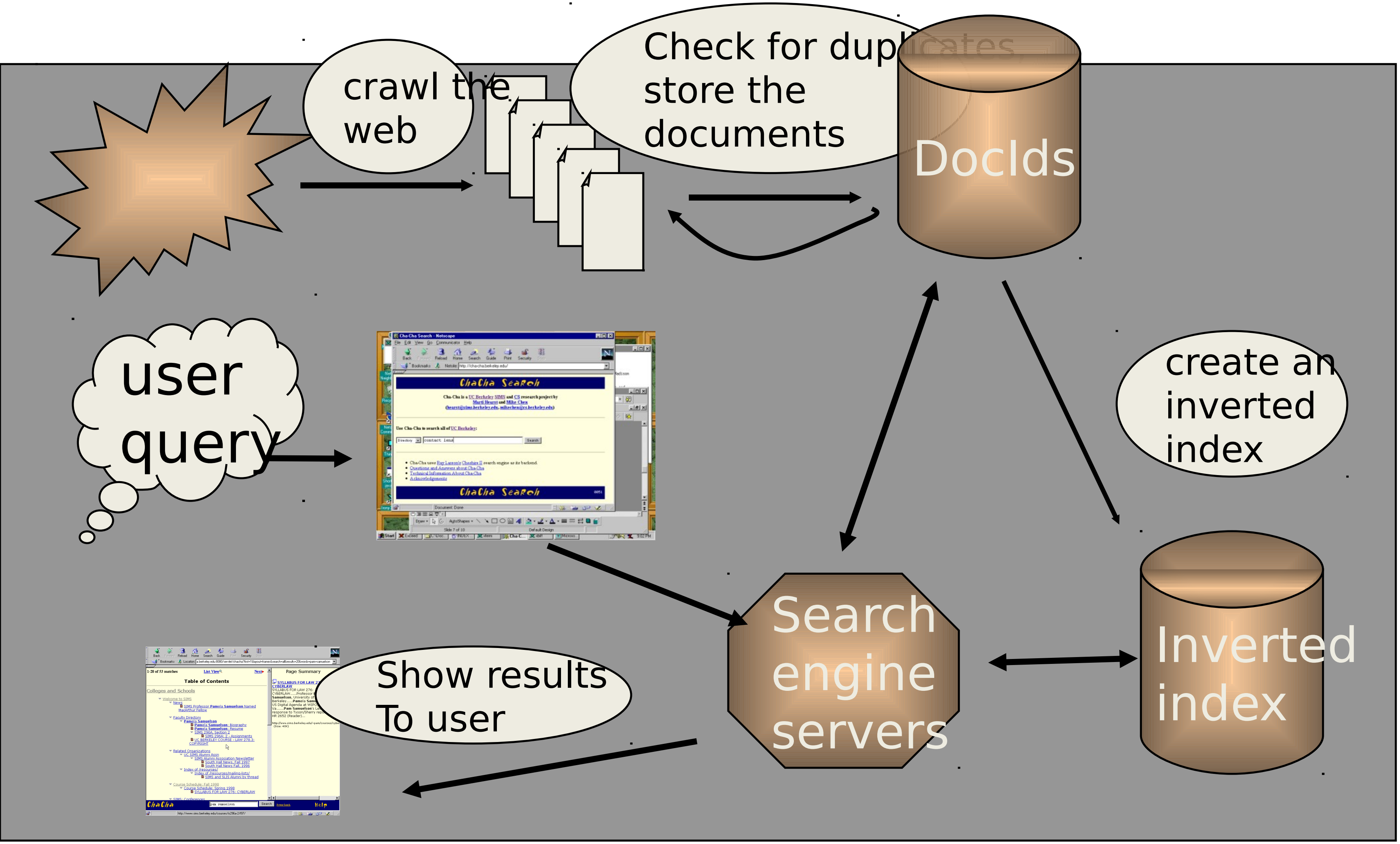
Searching the web

Searching the web

- Search engines determine how to rank pages using automated methods that look at the Web itself,
- Without any help from external source of knowledge
- There are enough information intrinsic to the Web and its structure to figure this out
- Example: go to Google and type “u2,” the first result it shows you is www.u2.com the home page of rock band U2.
- How did Google “know” that this was the best answer?

Standard Web Search Engine Architecture



crawl the web

Check for duplicates, store the documents

DocIds

create an inverted index

user query

Search engine servers

Inverted index

Show results To user

Cha-Cha Search - Netscape

http://cha-cha.berkeley.edu/

ChaCha Search

Cha-Cha is a UC Berkeley SIMS and CS research project by Marti Hearst and Mike Chen (hearst@sims.berkeley.edu, mikechen@cs.berkeley.edu)

Use Cha-Cha to search all of UC Berkeley:

Directory | contact lens | Search

- Cha-Cha uses Ray Larson's Cheshire II search engine as its backend.
- Questions and Answers about Cha-Cha
- Technical Information About Cha-Cha
- Acknowledgements

Document: Done

Location: a.berkeley.edu:8080/servlet/chacha?first=1&layout=frames&search=all&results=20&words=pam+samuelson

1-20 of 53 matches | List View | Next

Table of Contents

Colleges and Schools

- Welcome to SIMS
  - News
    - SIMS Professor Pamela Samuelson Named MacArthur Fellow
  - Faculty Directory
    - Pamela Samuelson
      - Pamela Samuelson: Biography
      - Pamela Samuelson: Resume
      - SIMS 296A, Section 2
        - SIMS 296A: 2 - Assignments
      - UC BERKELEY COURSE - LAW 278.3: COPYRIGHT
  - Related Organizations
    - UC SIMS Alumni Assn
      - SIMS Alumni Association Newsletter
        - South Hall News, Fall 1997
        - South Hall News Fall, 1996
    - Index of /resources/
      - Index of /resources/mailing-lists/
        - SIMS and SLIS Alumni by thread
  - Course Schedule: Fall 1998
    - Course Schedule: Spring 1998
      - SYLLABUS FOR LAW 276: CYBERLAW
  - SIMS: Conferences

Page Summary

SYLLABUS FOR LAW 27 CYBERLAW

SYLLABUS FOR LAW 276: CYBERLAW......Professor P Samuelson, University of Berkeley......Pamela Samu US Digital Agenda at WIPO Va.......Pam Samuelson's Le response to Tyson/Sherry repo HR 2652 (Reader)...

http://www.sims.berkeley.edu/~pam/courses/cyber (Size: 40K)

ChaCha | pam samuelson | Search | Power Search | Help

http://www.sims.berkeley.edu/courses/is296a-2/f97/

Historical Background – Searching the web

- Precursor: automated information retrieval systems of 1960s -
  - designed to search repositories of newspaper articles, scientific papers, patents, legal abstracts, and other document collections in response to keyword queries.
- Information retrieval systems have always had to deal with the problem that keywords are a very limited way to express a complex information need;
- it suffered from the problems of:
  - Synonymy(multiple ways to say the same thing)
  - Polysemy (multiple meanings for the same term)

Contd…

- With the arrival of the Web, where everyone is an author and everyone is a searcher, the problems surrounding information retrieval exploded in scale and complexity
- Diversity in authoring styles makes it much harder to rank documents according to a common criterion
- a single topic, one can easily find pages written by experts, novices, children, conspiracy theorists—and not necessarily be able to tell which is which

Contd…

- There is a correspondingly rich diversity in the set of people issuing queries, and the problem of multiple meanings becomes particularly severe
- But the Web also introduces new kinds of problems:

1. Dynamic and constantly-changing nature of Web content&
2. The genuinely of the content

Web Crawling

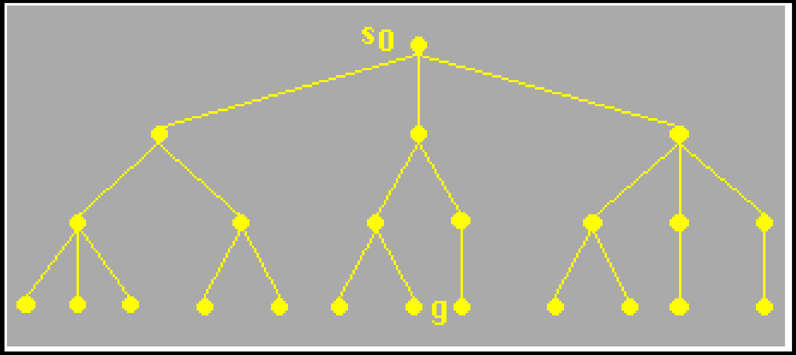
- How do the web search engines get all of the items they index?
- Main idea:
  - Start with known sites
  - Record information for these sites
  - Follow the links from each site
  - Record information found at new sites
  - Repeat

Web Crawlers

- How do the web search engines get all of the items they index?
- More precisely:
  - Put a set of known sites on a queue
  - Repeat the following until the queue is empty:
    - Take the first page off of the queue
    - If this page has not yet been processed:
      - Record the information found on this page
        - Positions of words, links going out, etc
      - Add each link on the current page to the queue
      - Record that this page has been processed
- In what order should the links be followed?

Page Visit Order

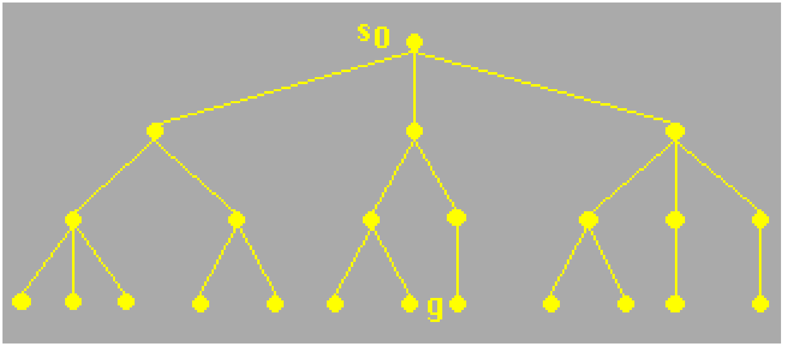
- Animated examples of breadth-first vs depth-first search on trees:
  - http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch



Structure to be traversed

Page Visit Order

- Animated examples of breadth-first vs depth-first search on trees:
  - http://www.rci.rutgers.edu/~cfs/472 html/AI SEARCH/ExhaustiveSearch

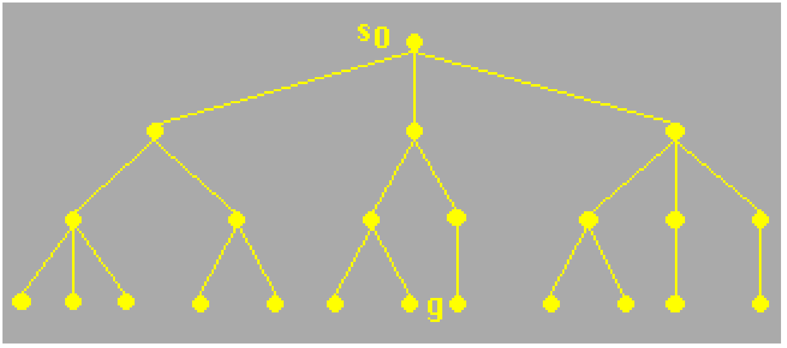


Breadth-first search

(must be in presentation mode to see this animation)

Page Visit Order

- Animated examples of breadth-first vs depth-first search on trees:
  - http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch.html



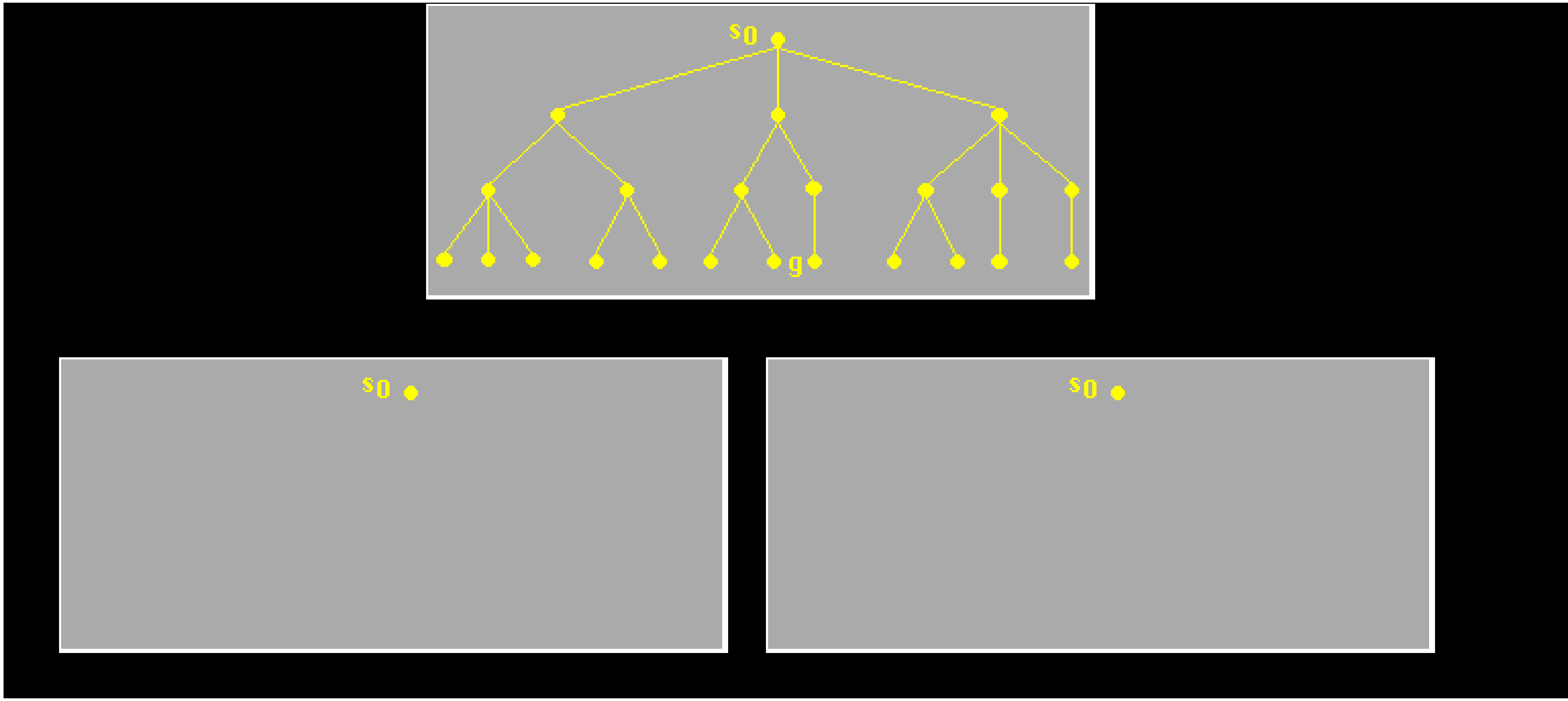
Depth-first search

(must be in presentation mode to see this animation)

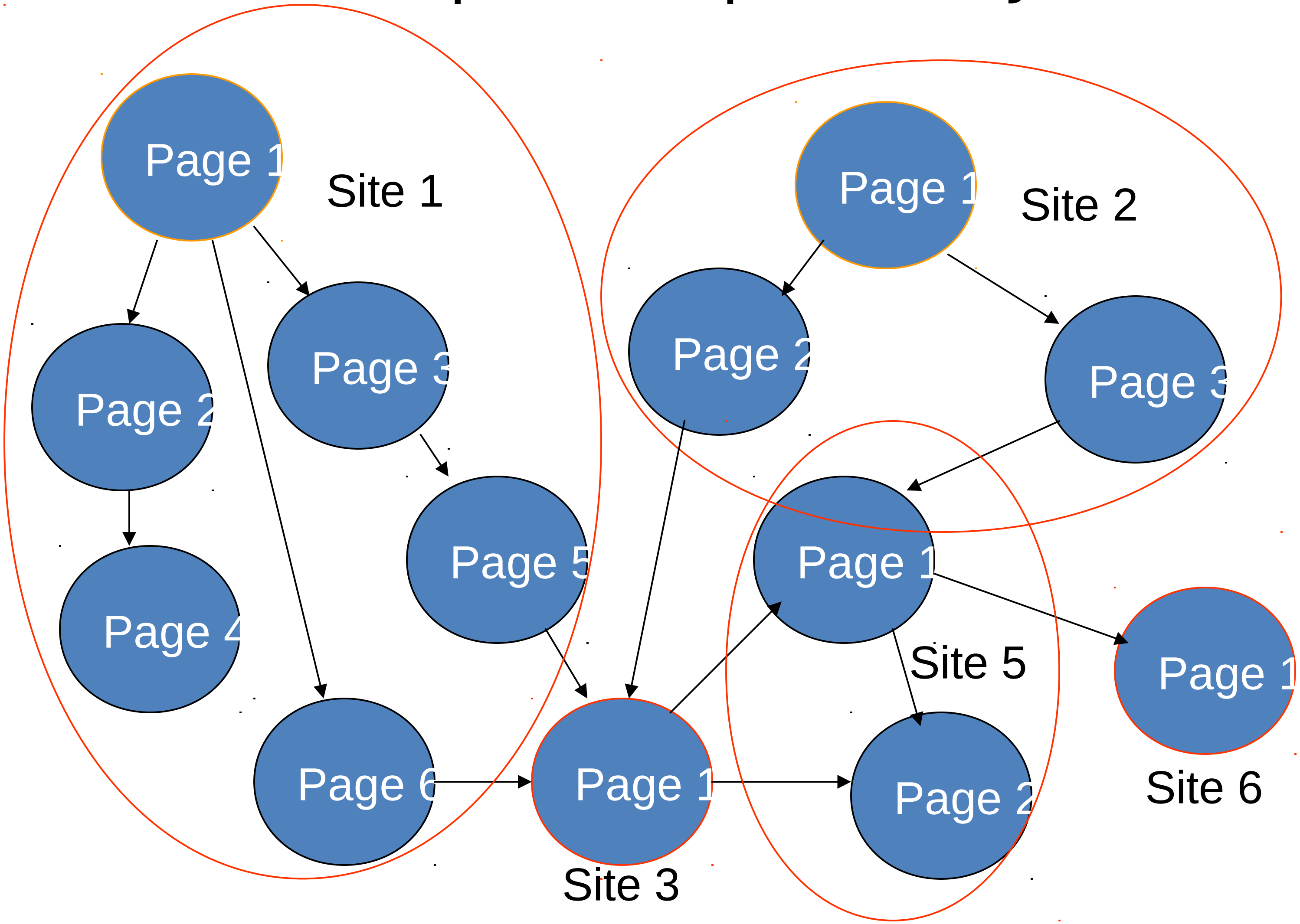
Page Visit Order

- Animated examples of breadth-first vs depth-first search on trees:

  http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch.html



Sites Are Complex Graphs, Not Just Trees



Web Crawling Issues

- Keep out signs
  - A file called robots.txt tells the crawler which directories are off limits
- Freshness
  - Figure out which pages change often
  - Recrawl these often
- Duplicates, virtual hosts, etc
  - Convert page contents with a hash function
  - Compare new pages to the hash table
- Lots of problems
  - Server unavailable
  - Incorrect html
  - Missing links
  - Infinite loops
- Web crawling is difficult to do robustly!

Today

- Web Search Engines and Algorithms

Directories vs. Search Engines

- Web Directories
  - Hand-selected sites
  - Search over the contents of the descriptions of the pages
  - Organized in advance into categories

- Search Engines
  - All pages in all sites
  - Search over the contents of the pages themselves
  - Organized after the query by relevance rankings or other scores

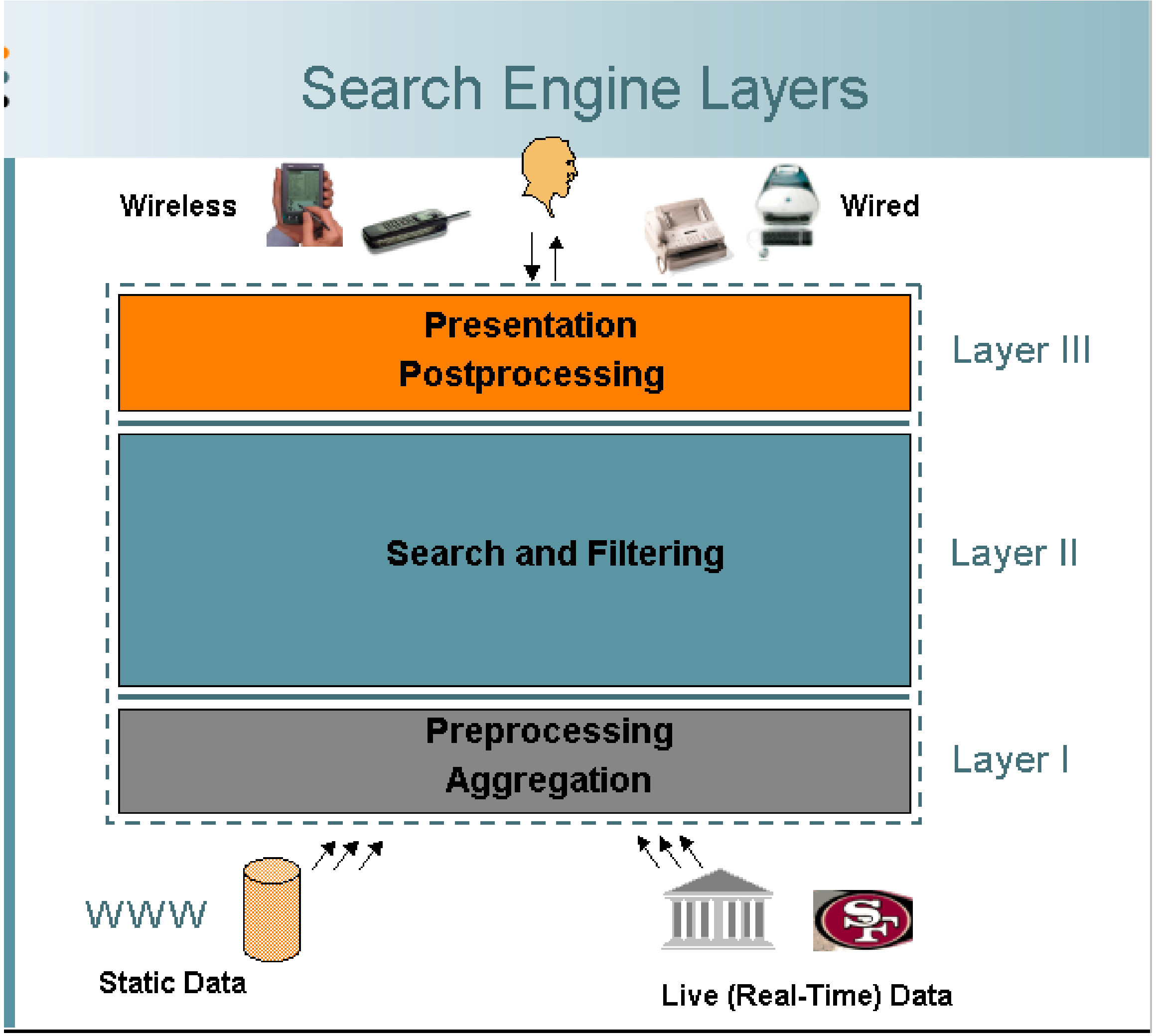
Web Search Queries

- Web search queries are SHORT
  - ~2.4 words on average (Aug 2000)
  - Has increased, was 1.7 (~1997)
- User Expectations
  - Many say “the first item shown should be what I want to see”!
  - This works if the user has the most popular/common notion in mind

Search Engines

- Crawling
- Indexing
- Querying

Web Search Engine Layers



Search Engine Layers

Wireless

Wired

Presentation
Postprocessing

Layer III

Search and Filtering

Layer II

Preprocessing
Aggregation

Layer I

WWW

Static Data

Live (Real-Time) Data

From description of the FAST search engine, by Knut Risvik
http://www.infonortics.com/searchengines/sh00/risvik_files/frame.htm

Google

- Google maintains (probably) the worlds largest Linux cluster (over 15,000 servers)
- These are partitioned between index servers and page servers
  - Index servers resolve the queries (massively parallel processing)
  - Page servers deliver the results of the queries
- Over 8 Billion web pages are indexed and served by Google

Search Engine Indexes

- Starting Points for Users include
- Manually compiled lists
    - Directories
- Page “popularity”
    - Frequently visited pages (in general)
    - Frequently visited pages as a result of a query
- Link “co-citation”
    - Which sites are linked to by other sites?

Starting Points: What is Really Being Used?

- Todays search engines combine these methods in various ways
  - Integration of Directories
    - Today most web search engines integrate categories into the results listings
    - Lycos, MSN, Google
  - Link analysis
    - Google uses it; others are also using it
    - Words on the links seems to be especially useful
  - Page popularity
    - Many use DirectHit’s popularity rankings

Web Page Ranking

- Varies by search engine
  - Pretty messy in many cases
  - Details usually proprietary and fluctuating
- Combining subsets of:
  - Term frequencies
  - Term proximities
  - Term position (title, top of page, etc)
  - Term characteristics (boldface, capitalized, etc)
  - Link analysis information
  - Category information
  - Popularity information

Ranking: Hearst ‘96

- Proximity search can help get high-precision results if >1 term
  - Combine Boolean and passage-level proximity
  - Proves significant improvements when retrieving top 5, 10, 20, 30 documents
  - Results reproduced by Mitra et al. 98
  - Google uses something similar

Ranking: Link Analysis

- Why does this work?
  - The official Toyota site will be linked to by lots of other official (or high-quality) sites
  - The best Toyota fan-club site probably also has many links pointing to it
  - Less high-quality sites do not have as many high-quality sites linking to them

Navigating the web

- Definition:
  - Web navigation refers to the process of navigating a network of information resources in the World Wide Web, which is organized as a hypertext or hypermedia
- The user interface that is used to do so is called a web browser
- A central theme in web design is the development of a web navigation interface that maximizes usability

Contd…

- The first and most important way of navigation is using a search engine.
- As described earlier, a search engine is a web site that allows you to find web pages based on the words you enter or ‘search query' –this can be a question, or just a collection of words that you think will be on the web page you're after.

Navigation

- Principles for good navigation design
  - A site must:
    1. Let me know where I am at all times
    2. Clearly differentiate hyperlinks from content
    3. Let me know clearly where I can go from here
    4. Let me see where I've already been
    5. Make it obvious what to do to get somewhere
    6. Indicate what clicking a link will do

Good navigation

- Easily learned
- Consistent
  - In terms of their placement, offerings, and appearance
- Provides feedback
- Requires an “economy of action and time”
- Users understandable labels
- Is appropriate to site’s purpose
- Supports users’ goals
- Provide contextual clues and flexibility

Where should you put navigation?

- Depends on the type of navigation
- The golden rules are:
  - Put the most useful navigation where it’s closest to hand
  - Put navigation where the user is likely to look for it.

Navigation Models

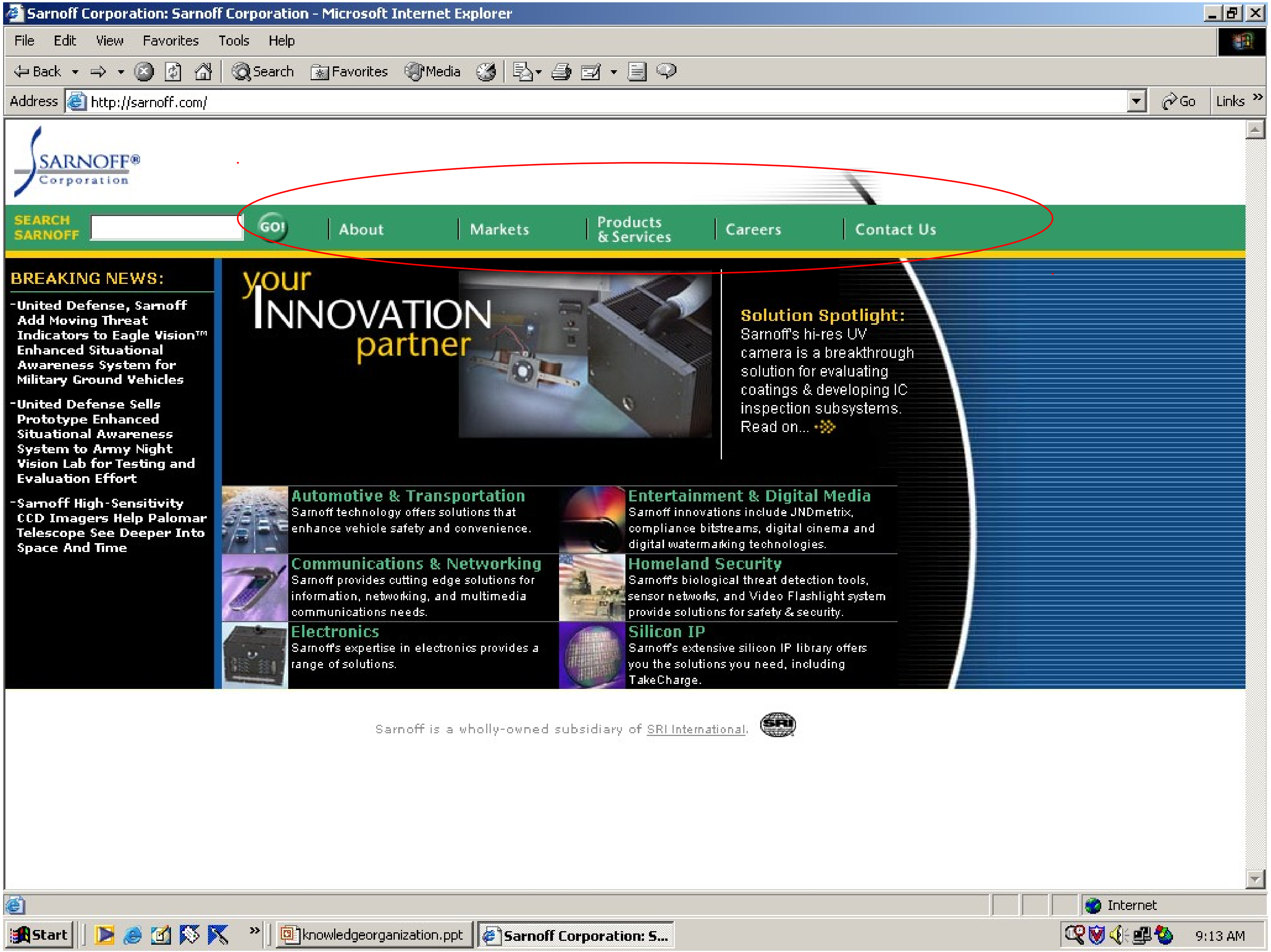
- common IA and navigation conventions.
  - List of contents
  - Breadcrumb trail
  - Horizontal top bar
  - Tabs
  - 2-level top (bar or tabs)
  - Top and side bars
  - Buttons bar with revealed drop-down
  - Multiple-level tree nav
  - Paging

Where should you put navigation?

- Depends on the type of navigation
- The golden rules are:
  - Put the most useful navigation where it’s closest to hand
  - Put navigation where the user is likely to look for it.

Types of navigation systems

- Hierarchical navigation systems
  - Similar to the information hierarchy
  - Require additional navigation systems
- Global navigation systems (site-wide navigation systems)
  - Able to jump back to the main page
  - Important to extend the global navigation system throughout the sub-site
- Local navigation systems
- Ad hoc navigation
  - Embedded links



BREAKING NEWS:

- United Defense, Sarnoff Add Moving Threat Indicators to Eagle Vision™ Enhanced Situational Awareness System for Military Ground Vehicles
- United Defense Sells Prototype Enhanced Situational Awareness System to Army Night Vision Lab for Testing and Evaluation Effort
- Sarnoff High-Sensitivity CCD Imagers Help Palomar Telescope See Deeper Into Space And Time

Solution Spotlight:
Sarnoff's hi-res UV camera is a breakthrough solution for evaluating coatings & developing IC inspection subsystems. Read on...

Automotive & Transportation
Sarnoff technology offers solutions that enhance vehicle safety and convenience.

Entertainment & Digital Media
Sarnoff innovations include JNDmetrix, compliance bitstreams, digital cinema and digital watermarking technologies.

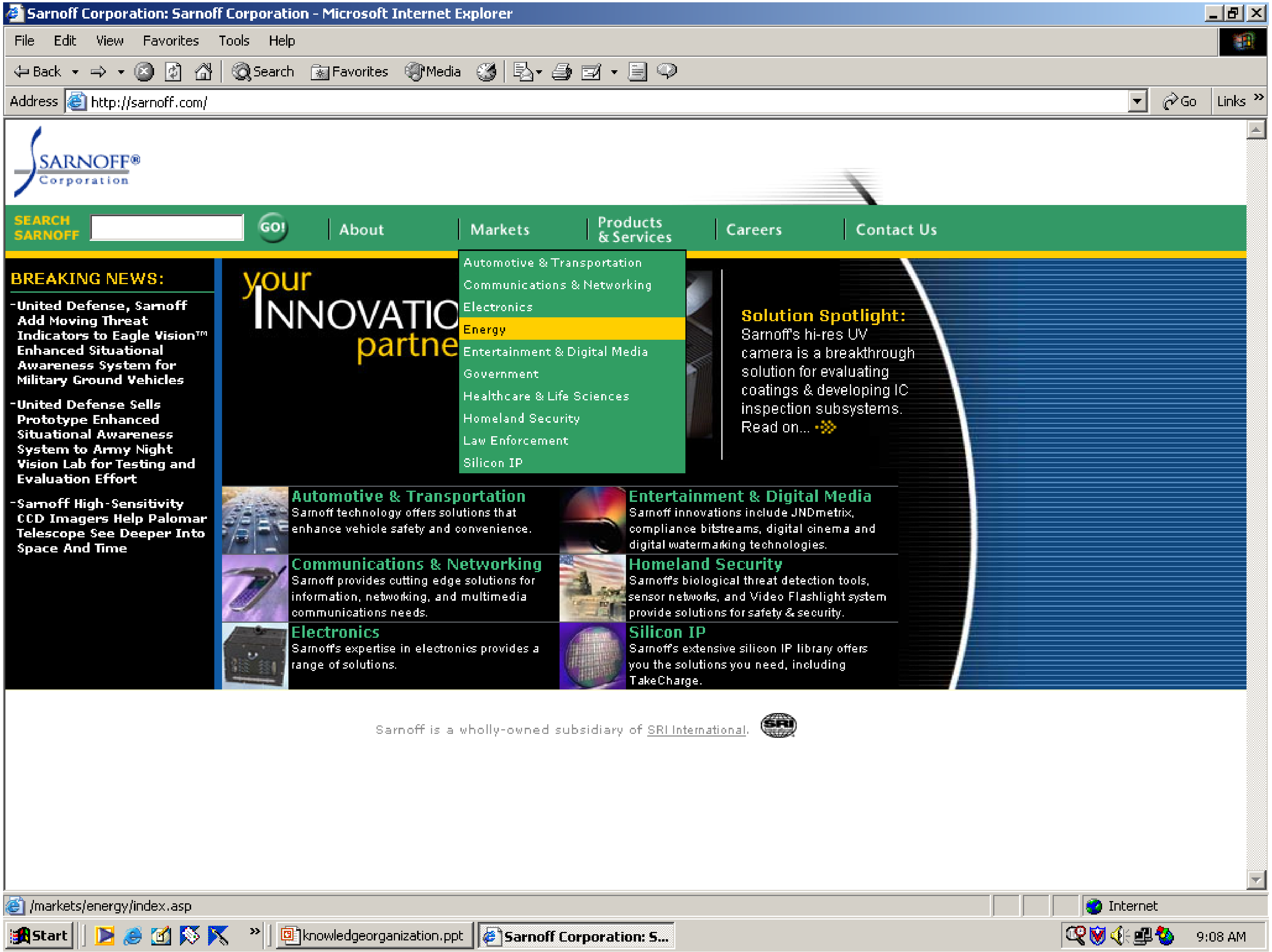
Communications & Networking
Sarnoff provides cutting edge solutions for information, networking, and multimedia communications needs.

Homeland Security
Sarnoff's biological threat detection tools, sensor networks, and Video Flashlight system provide solutions for safety & security.

Electronics
Sarnoff's expertise in electronics provides a range of solutions.

Silicon IP
Sarnoff's extensive silicon IP library offers you the solutions you need, including TakeCharge.

Sarnoff is a wholly-owned subsidiary of SRI International.



BREAKING NEWS:

- United Defense, Sarnoff Add Moving Threat Indicators to Eagle Vision™ Enhanced Situational Awareness System for Military Ground Vehicles
- United Defense Sells Prototype Enhanced Situational Awareness System to Army Night Vision Lab for Testing and Evaluation Effort
- Sarnoff High-Sensitivity CCD Imagers Help Palomar Telescope See Deeper Into Space And Time

Solution Spotlight:
Sarnoff's hi-res UV camera is a breakthrough solution for evaluating coatings & developing IC inspection subsystems. Read on...

Automotive & Transportation
Sarnoff technology offers solutions that enhance vehicle safety and convenience.

Entertainment & Digital Media
Sarnoff innovations include JNDmetrix, compliance bitstreams, digital cinema and digital watermarking technologies.

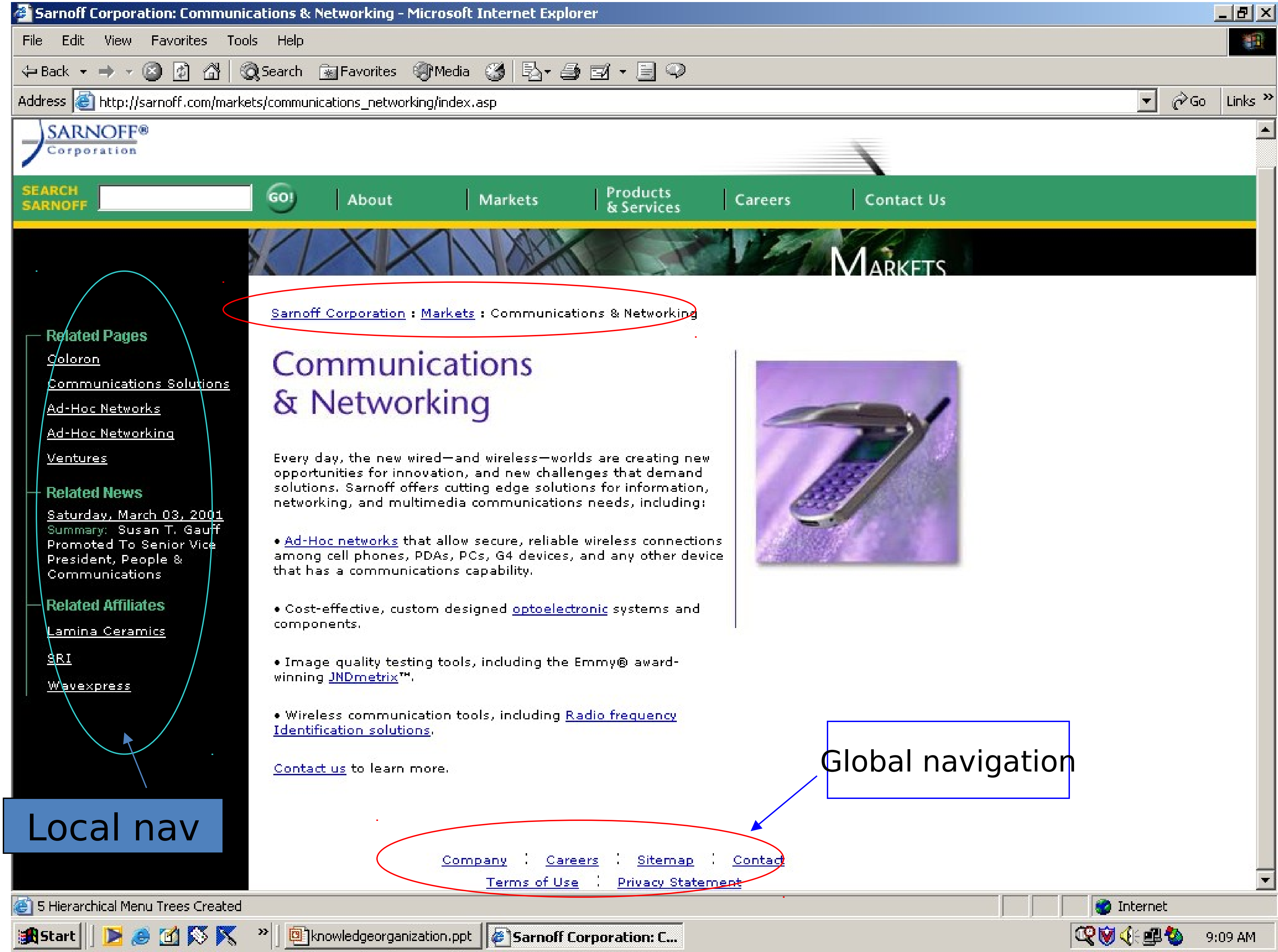
Communications & Networking
Sarnoff provides cutting edge solutions for information, networking, and multimedia communications needs.

Homeland Security
Sarnoff's biological threat detection tools, sensor networks, and Video Flashlight system provide solutions for safety & security.

Electronics
Sarnoff's expertise in electronics provides a range of solutions.

Silicon IP
Sarnoff's extensive silicon IP library offers you the solutions you need, including TakeCharge.

Sarnoff is a wholly-owned subsidiary of SRI International.


Sarnoff Corporation: Communications & Networking - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://sarnoff.com/markets/communications_networking/index.asp
Go
Links
SARNOFF® Corporation
SEARCH SARNOFF
GO!
About
Markets
Products & Services
Careers
Contact Us
MARKETS
5 Hierarchical Menu Trees Created
Internet
Start
knowledgeorganization.ppt
Sarnoff Corporation: C...
9:09 AM

Sarnoff Corporation : Markets : Communications & Networking

Related Pages

Coloron

Communications Solutions

Ad-Hoc Networks

Ad-Hoc Networking

Ventures

Related News

Saturday, March 03, 2001
Summary: Susan T. Gauff Promoted To Senior Vice President, People & Communications

Related Affiliates

Lamina Ceramics

SRI

Wavexpress

Communications & Networking

Every day, the new wired—and wireless—worlds are creating new opportunities for innovation, and new challenges that demand solutions. Sarnoff offers cutting edge solutions for information, networking, and multimedia communications needs, including:

- Ad-Hoc networks that allow secure, reliable wireless connections among cell phones, PDAs, PCs, G4 devices, and any other device that has a communications capability.
- Cost-effective, custom designed optoelectronic systems and components.
- Image quality testing tools, including the Emmy® award-winning JNDmetrix™.
- Wireless communication tools, including Radio frequency Identification solutions.

Contact us to learn more.

Company : Careers : Sitemap : Contact

Terms of Use : Privacy Statement

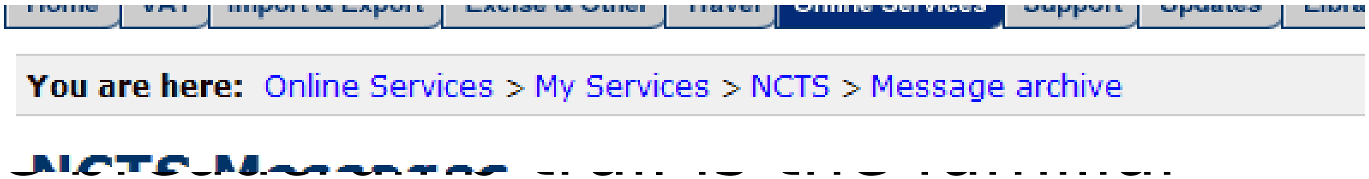
Local nav

Global navigation

Navigation Models

- common IA and navigation conventions.
  - List of contents
  - Breadcrumb trail
  - Horizontal top bar
  - Tabs
  - 2-level top (bar or tabs)
  - Top and side bars
  - Buttons bar with revealed drop-down
  - Multiple-level tree nav
  - Paging

Breadcrumb trail

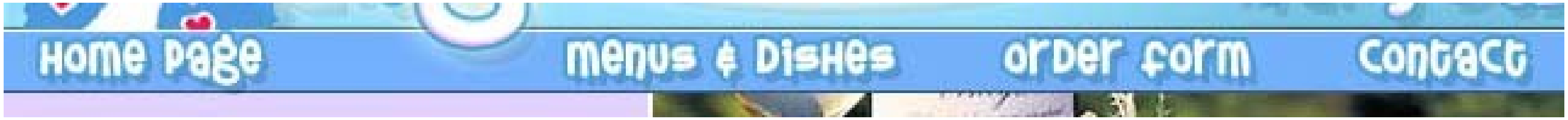


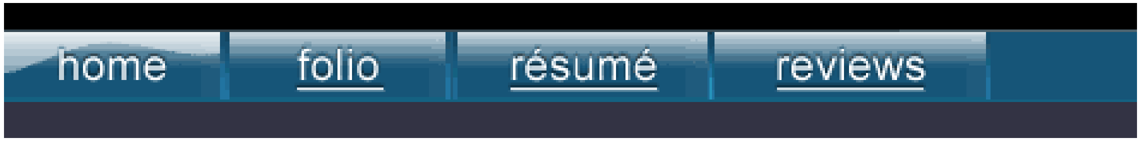
You are here: Online Services > My Services > NCTS > Message archive

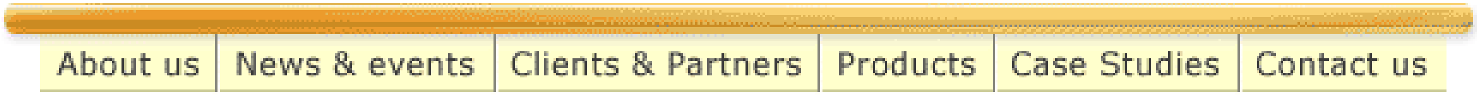
NCTS Messages

- Th[illegible] trail is the familiar navigation device that:
  - Shows you where you in are in a hierarchy, and
  - Lets you click back to any point above where you are now
- Breadcrumb trails are great in situations where:
  - You've got a particularly deep hierarchy, say four levels or more
  - The possible flow is such that a typical user might want or need to get back to a specific previous place

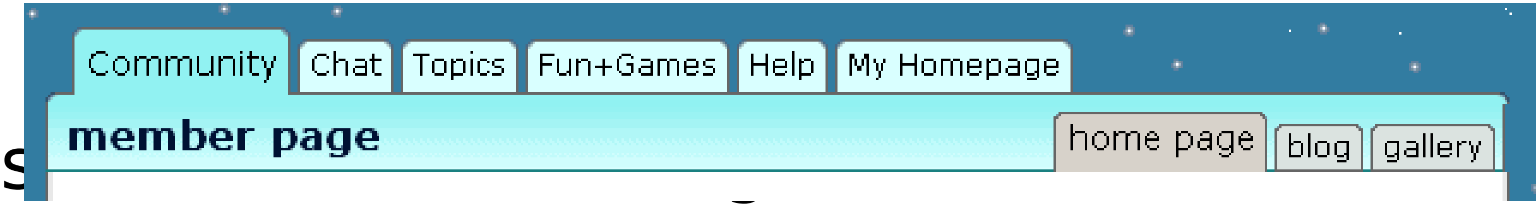
Horizontal top bar







Tabs

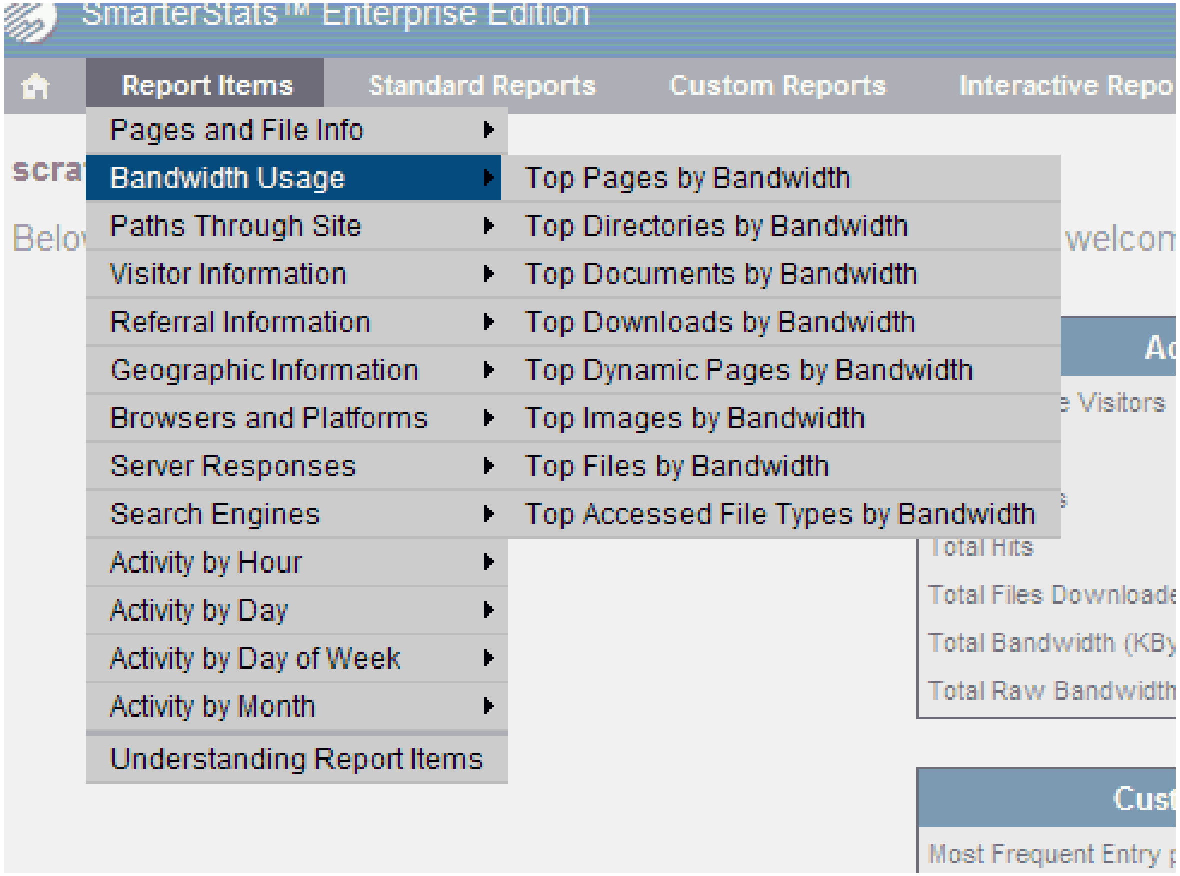


- S :
  - They serve to show the active section/selection very clearly
  - They naturally have a working visual hierarchy, with a real-world connection that makes them extremely clear. A tab is normally attached to (part of) a folder or sheet in a binder, and physically labels everything in the folder, or on the sheet.
  - They are unambiguously mutually-exclusive. It's physically impossible to have two tabs selected (because they would both have to be at the front).

Top and side bars

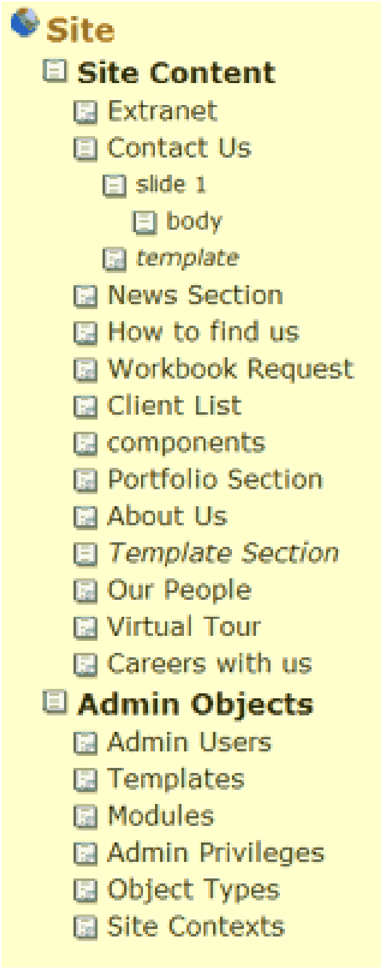
- Very common. The top bar is used for the site-level navigation/tools and often first-level navigation, because these are more fixed.

Nav bar with revealed drop-downs



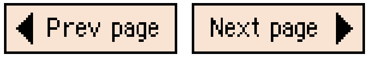
Multiple-level tree nav

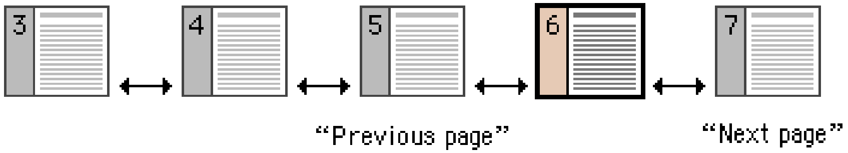
- Benefits
  - It is relatively familiar and intuitive (provided it is presented in a conventional format).
  - It can provide relatively simple access to a complex structure.

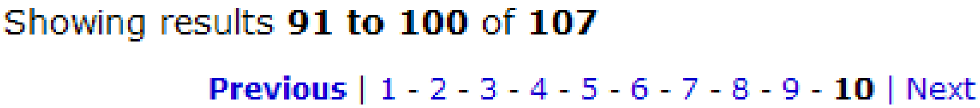


- Site
  - **Site Content**
    - Extranet
    - Contact Us
      - slide 1
        - body
      - *template*
    - News Section
    - How to find us
    - Workbook Request
    - Client List
    - components
    - Portfolio Section
    - About Us
    - *Template Section*
    - Our People
    - Virtual Tour
    - Careers with us
  - **Admin Objects**
    - Admin Users
    - Templates
    - Modules
    - Admin Privileges
    - Object Types
    - Site Contexts

Paging







Showing results **91 to 100** of **107**

Previous | 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - **10** | Next

- Where you get a piece of content that spans several pages (typically long articles, long indexes, forums, or search results).
- By providing paging buttons and links to local home pages and contents pages you give users the tools to understand how you have organized your Web site information, even if they have not entered your Web of pages through a home page or contents page.
- The buttons don't prevent people from reading the information in whatever order they choose, but they do allow readers to follow the sequence of pages you have laid out.

Remote Navigation Elements

- Provide an alternative bird’s-eye view of the site’s content
  - Tables of contents
  - Site Map
  - Index

Book marking

- When you find a web page that you find interesting, instead of having to remember the address of the webpage, you can simply save the address as a 'bookmark' in your browser (some browsers call them favorites).
- Means you can come back and find that exact web page again

Contd…

- There are websites that allow you to save your bookmarks to an account on the website, instead of to your browser.
- Means that you can access your bookmarks from anywhere–you don't have to be on the computer that you first found the site on
- Also, you can make your bookmarks public, meaning other people can see the interesting stuff you've found –called 'social bookmarking'

Contd…

- When you find something interesting on the internet, you can also share it with your friends.
  - One way to do this is to email your friends with a link to the particular web page, however this is increasingly being replaced by social media services.
  - When lots of people share something with their friends, and then their friends share it on, and so on, this is called a viral or in some cases an internet meme

Contd…

- Hyperlinks : hot links
- Whatever you call them, these links provide a connection between Web pages that allows for amazingly easy access to other Web pages.
- A link or hyperlink can be text, an icon, a picture, or an icon that moves a user from one Web page or Web site to another.
- A hyperlink(often underlined, colored etc.)has an unseen Web address imbedded in it.

Navigating Within a Web Page

- there are convenient ways to move around that particular page itself.
- Often a Web page holds more information than can fit on one screen.
- A Web page appears aligned to the upper left hand corner of your screen.
- There is often information that you cannot see farther down after the last line on the screen.
- Sometimes there is also more information to the right of the screen.

Contd…

Slider & Arrows

- Scrolling is an easy way to navigate on a Web page.
- You can scroll up and down and side to side by using either the horizontal or vertical onscreen scroll bars on the bottom and right side of the screen.
- To scroll using the onscreen scroll bars, simply position your cursor on the slider on the scroll bar.
- Hold the mouse button down and drag the slider up and/or down on the vertical scroll bar (or side to side on the horizontal scroll bar).
- You can also position your cursor over the arrows at the top and the bottom of the vertical scroll bar (left and right sides of the horizontal scroll bar) to move one line at a time.

Contd…

- Using Arrow Keys
  - The keyboard holds some other choices for helping you move around a Web page.
  - The first are the Page Up and Page Down keys on your keyboard.
- Using Mouse
- Other important components:
  - URL Bar/Address Bar/Location Bar
  - Home
  - Back and Forward
  - Refresh
  - sitemap
  - text links
  - Tabs for more navigation etc.

Contd…

The purpose of navigation is to:

1. Present readers with the most user-friendly path through the classification so that they can find the content they want quickly.
2. Ensure readers always know where they are on the site.
3. Allow readers to move quickly and logically through the web site.
4. Give readers the proper context of the document they are reading.
5. Highlight for the reader parts of the classification that the organization wants to promote.

- a golden rule of navigation design:
  - Start your design from the reader's point of view.
  - Get the people who will actually use the web site involved in the design from the earliest point possible.
- Website navigation is important to the success of website visitor’s experience to the website.
- The website’s navigation system is like a road map to all the different areas and information contained within the website.