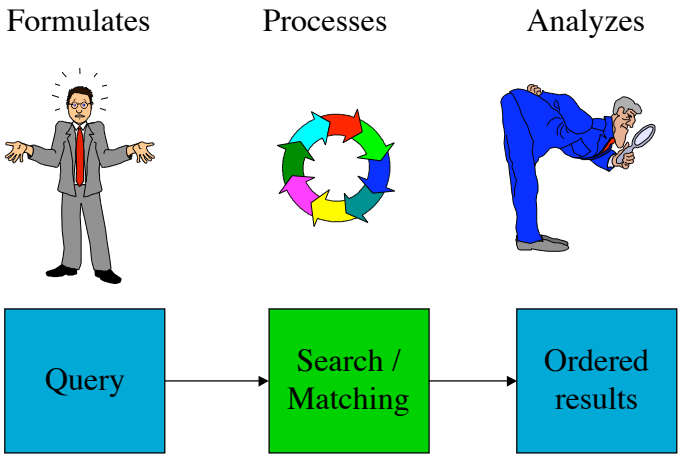
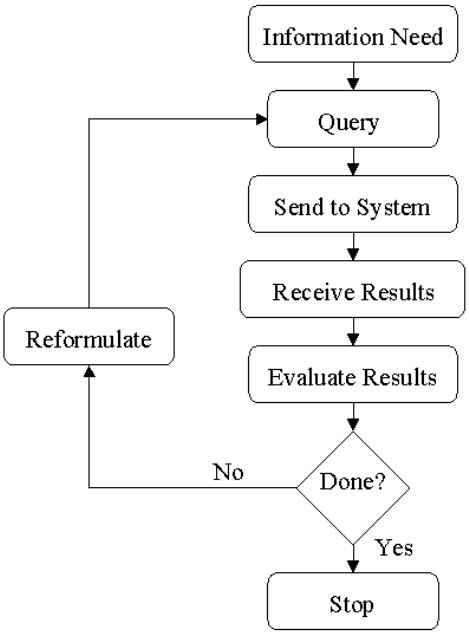
INFSCI 2140
Information Storage and Retrieval
Lecture 10: Search Interface and Information Visualization

Peter Brusilovsky
http://www2.sis.pitt.edu/~peterb/2140-051/

The ad-hoc search process



Step diagram for traditional information access process

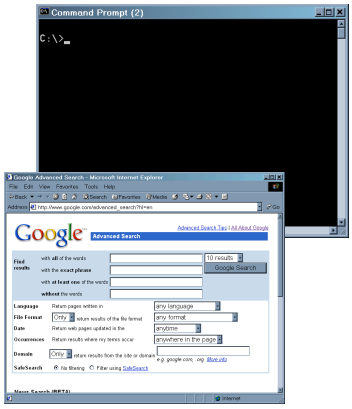


Search interfaces

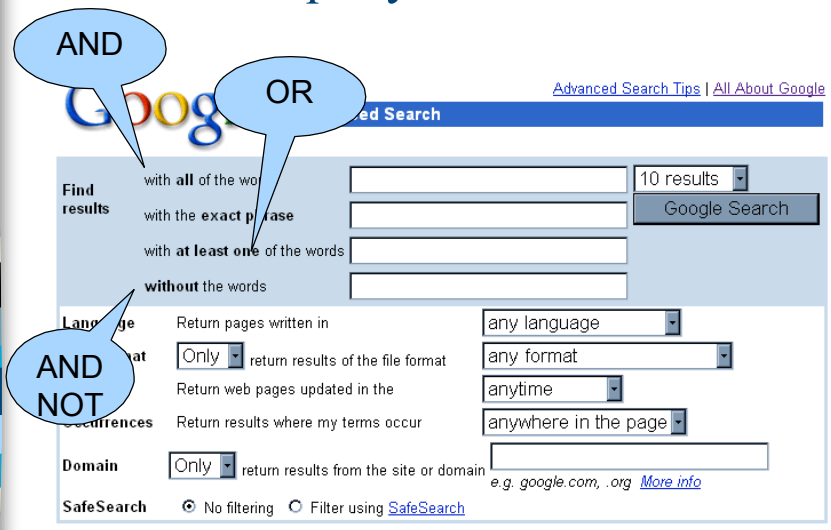
- Classic ad-hoc search is oriented to old teletype/command line interface style
  - Query is typed in
  - Results are returned as a flow of text
- Interfaces has got better over years
  - Rich text presentation with formatting
  - Graphical user interfaces
- Can we improve search interfaces too?

Better query formulation interface

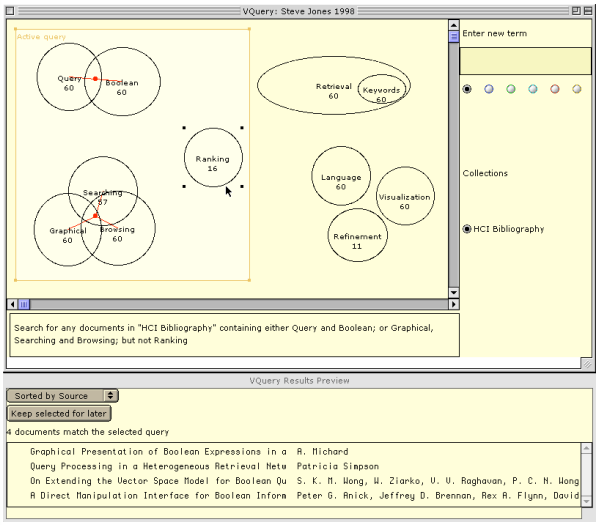
- Shneiderman identifies five primary HCI styles:
  - Command language
  - Natural language
  - Form filling
  - Menu selection
  - Direct manipulation
- We distinguish
  - GUI
  - Direct manipulation



Form-based query formulation

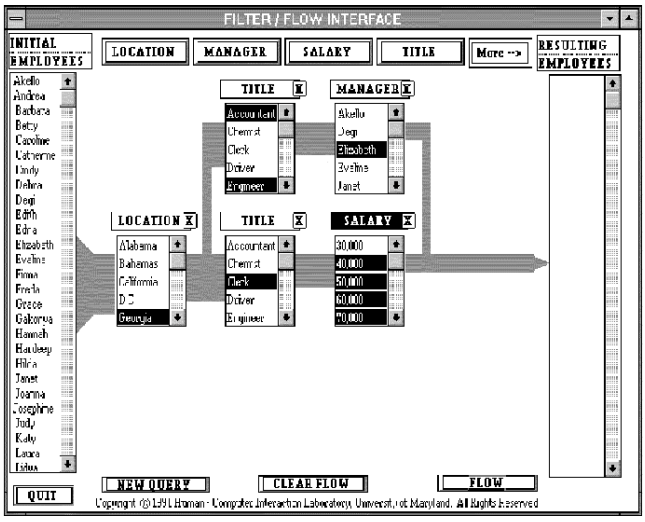


Graphical User Interface I



- Venn diagrams were proposed to improve Boolean query formulation

Graphical User Interface II



Filter flow model

- Two or more filter in sequence create the semantics of a conjunct (AND)
- Two or more in parallel create a disjunct (OR)

Direct manipulation for search?

- How we can use direct manipulation in the classic ad-hoc search process?
- The case of Stanford Digital Library (CHI'97)

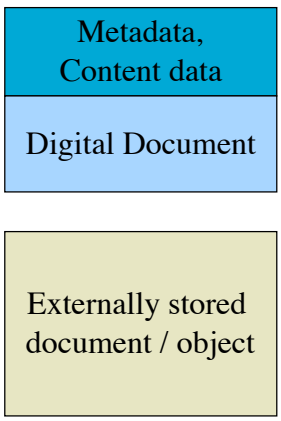
Why to bother about presentation?

- Looking through the search results is a part of the process of finding relevant documents
- The overall process could be improved if this part is improved
- The standard presentation is the ordered list of matched documents
- What can we improve?

What can we do?

- Decide what to present for a document
- Show context
- Explain relevance to the query
- Group the results
- Present results not in a linear list
- Present results graphically
- Let the user explore the results interactively

Documents and surrogates



- Digitally stored, used for search, presentation, and selection
- Digitally stored, used for presentation and selection, not used for search
- Externally stored, not used for search

What to present?

Why it is a bad idea?

- Document ID
- Metadata, content data
  - Metadata: author, title, year, source
  - Keywords
  - Abstract
- An extract of the full document
  - First paragraph
  - Thumbnail
- Full document

Why it is a bad idea?

Two-step / three-step presentation

- Two steps:
  - Level 1 - list with minimal information
  - Level 2 - full information by request
- Three steps:
  - Level 1 - list with minimal information
  - Level 2 - more detailed information by request
  - Level 3 - full information by request

Example: Photo archive

- Photos are stored, but are not searchable
- Searchable are *descriptions*
- Description: what, when, where
  - Content (abstract vs. classifier)
  - Time (granularity!)
  - Location (coding scheme vs words)
- What to present?

Case study: Movie rental store

The case of search engines

- The choice:
  - Header
  - URL
  - Content
- Core elements: Header and URL
  - Why they are important?
  - Why they are not enough?

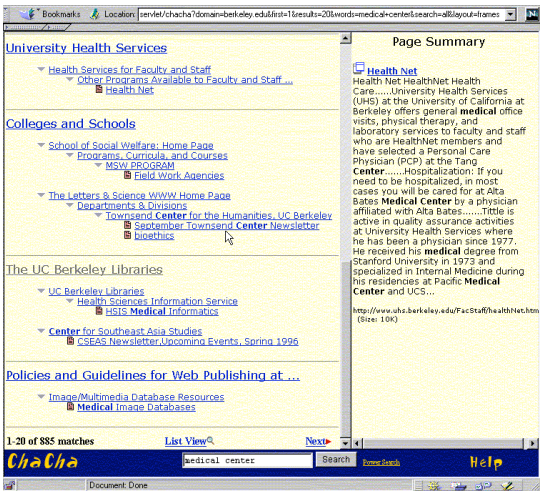
What else except the core?

- Classic design: Excite, 2 steps
  - Start of the document
- Modern design: Lycos, AltaVista
  - KWAC (keywords and context)
- Advanced design:
  - NorthernLight: relevance, category
  - Google: Link to cashed document

Showing the Context (tree)

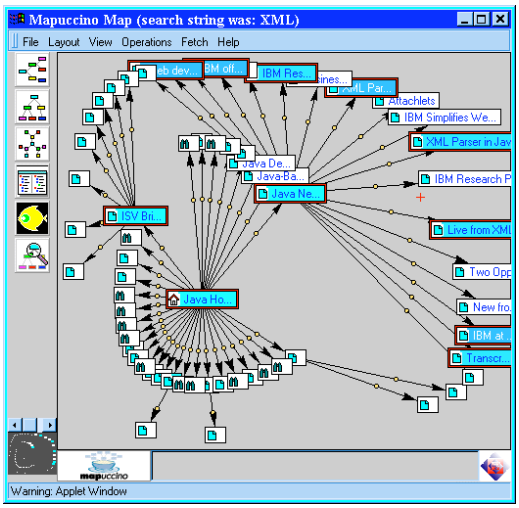
- Some systems try to show the results in a proper context
  - Cha-Cha system collect the Web pages that satisfy a query, then come up to their home pages and collect also them. This is made in order to show to the user a hierarchy (or a path) that goes to the query results and helps to give them a context

Show the context (path)



- Cha-Cha shows the user a path to each query result helping to see the context

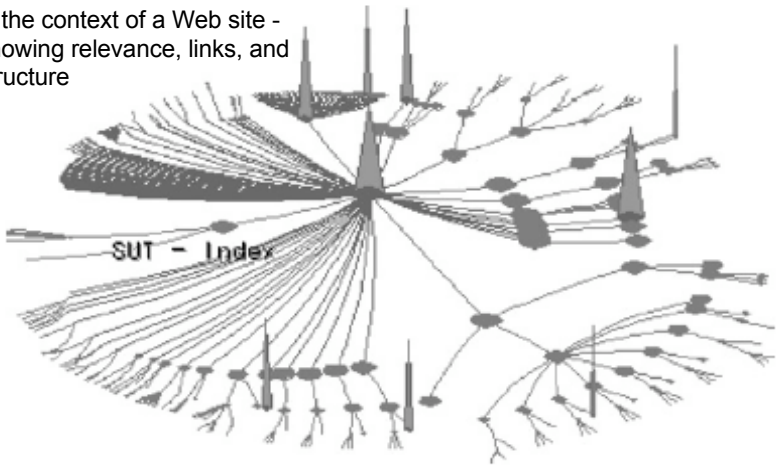
Relevance and Context (links!)



- Mappuccino allows the user to search on a specific web site. The pages that satisfy the query are shown together with the other linked pages. The idea is that the user will find what he needs in the results pages or in the linked pages

Relevance and Context (sctructure!)

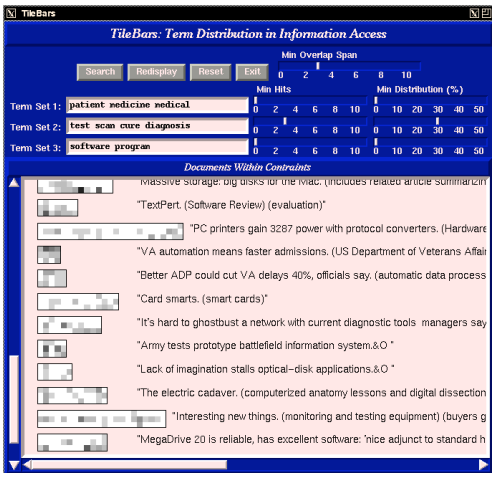
- WIDAS shows relevant documents in the context of a Web site - showing relevance, links, and structure



Relationship between results and the query

- The motivation: to show how the document relates to the query
  - If there was a year restriction -> show year
  - If there were keywords - show which are found (KWOC - KWIC - KWAC)
- Some efforts to better show keyword relevance between document and query
  - Semantic Highlighting / Google
  - TileBars

TileBars

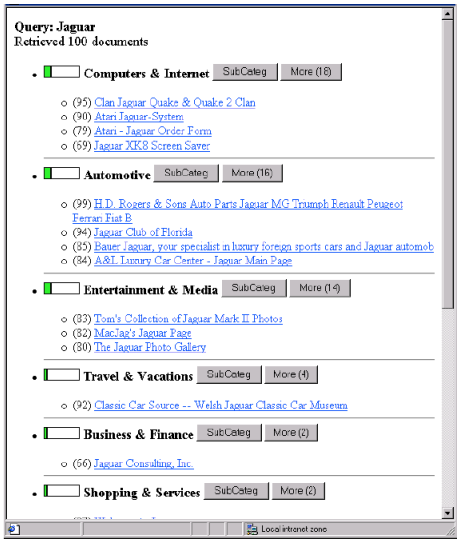


The system shows the degree of match for each query word in the documents (darker squares represent more frequent matches)

Semantic Grouping

- Semantic grouping: the idea is to group documents together by a semantic feature (taken from metadata or mined)
  - Source / Author
  - Media
  - Date / Time
- If no metadata is available the category can be deduced using automatic classification

Hierarchical Classification



- Dumais and Chen approach to present search results
- Uses automatic classification with CVM

Clustering

- If no category for classification is available, documents can be simply grouped by their similarity
- The idea of *clustering* is to group together documents with similar content
  - Based on keywords-level similarity between documents
  - There are many clustering algorithms that differ in speed, precision, presentation power
  - Hierarchical and 2D clustering
  - The problem of cluster naming

Managing quantity

- More is better?
- Quantity and quality
- Let the user choose
- Setting standard cut-off point
- Adaptation to the user's task and background
  - Adaptive filtering
  - Adaptive cut-off

Information Visualization for search result presentation

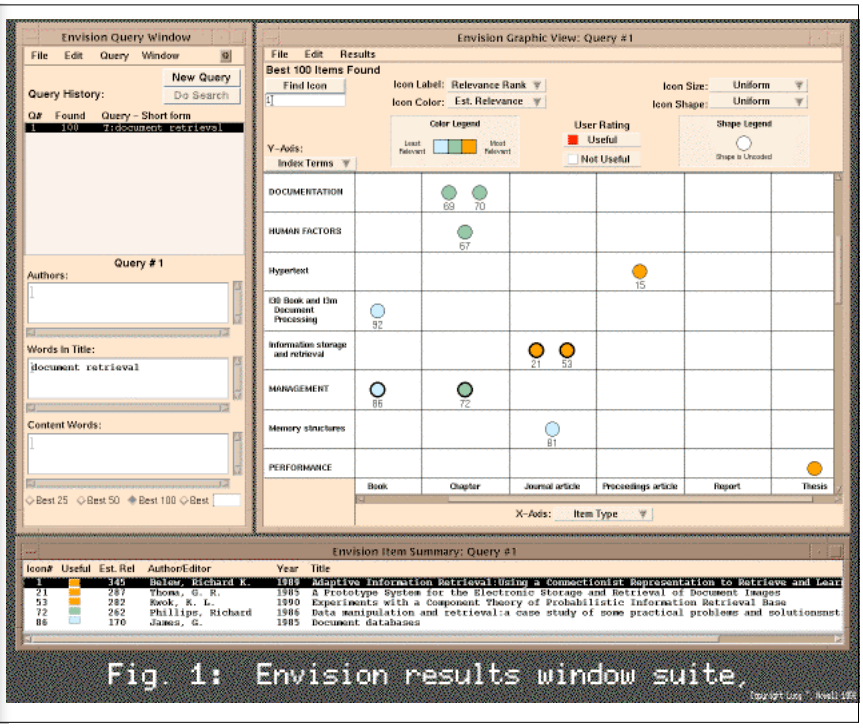
- Present results not in a linear list (2-3D)
  - Table: Envision, SenseMaker
  - 2D or 3D space: VIBE, InfoCrystal, LyberWorld, Lighthouse
- Let the user explore the results by manipulation with visualization
  - VIBE, BIRD, GUIDO, LyberWorld, Envision

Graphical results presentation

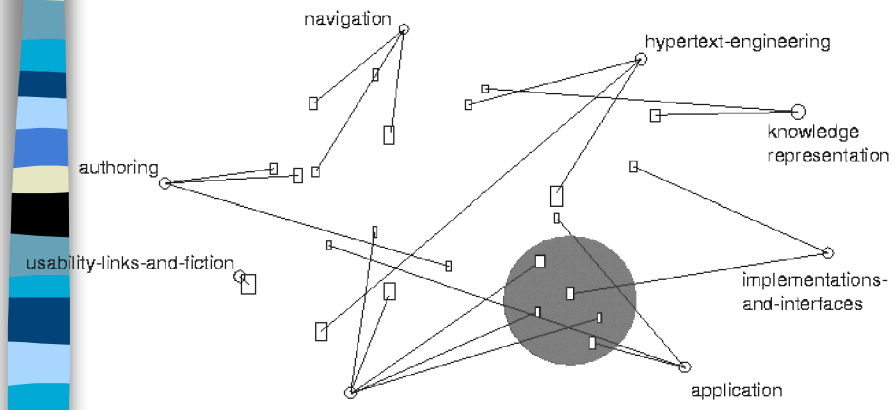
- Most graphical presentation approaches are based on the same ideas
  - Group similar documents
  - Show relevance to the query
- In a table similar documents can be shown in the same cell
  - Metadata-based: Envision
  - Similarity-based: SOM

Present results not in a linear list

- Envision
  - Virginia Tech Digital Libraries project
    http://www.dlib.vt.edu/projects/Envision/
- 2-D Table interface for data exploration
  - This user-controlled system facilitates examining very large data sets, displaying multiple aspects of the data simultaneously and efficiently, and interactive discovery of patterns in the data



Query terms hits between documents



- Vibe system places the query terms at the boundaries of a space and the documents are scattered inside this space

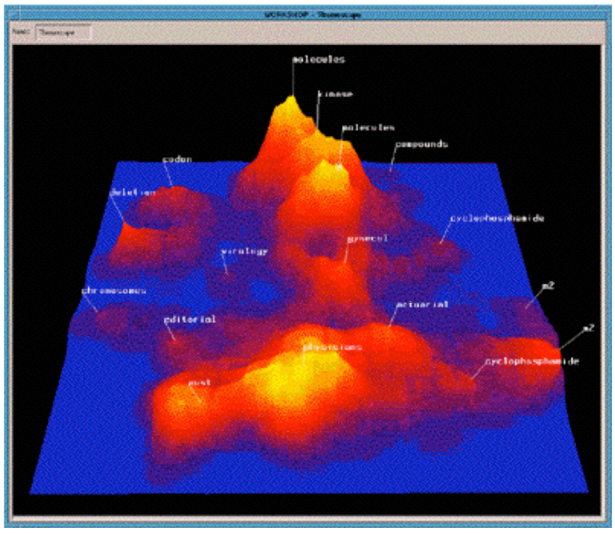
Information visualization beyond presentation of search results

- Information visualization can provide an alternative to search and used as a different information access paradigm
- Information visualization
  - Similar to browsing: finding documents by navigation and manipulation
  - Uses more expressive 2D and 3D representation
  - Allows to see “the whole picture”

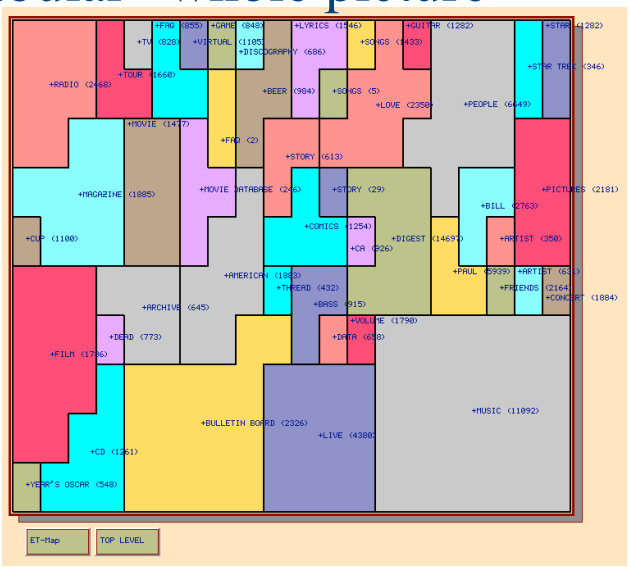
Some examples of information visualization

- Presenting “the whole picture”
  - Tabular
  - 2D or 3D
- Interfaces for exploration of specifically organized data (tables, hierarchies…)
  - TableLens, LifeLines
- Visualization of hypertext and the Web
  - Hyperbolic Browser
- Adaptive Information Visualization
  - Lighthouse, Knowledge Sea

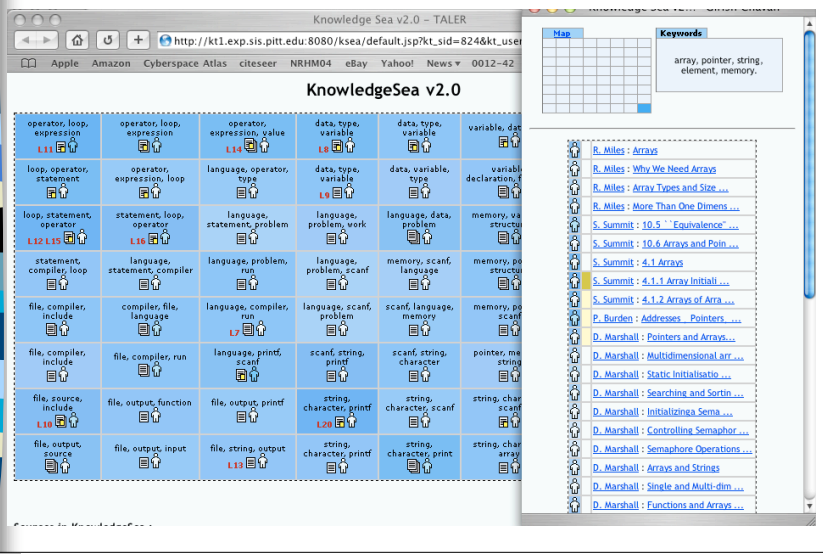
Graphical “whole picture”



Tabular "whole picture"



Knowledge Sea: Adaptive IV



Dynamic Queries

- Query is issued using GUI controls
- Query response is visible and visualized immediately
- Query can be dynamically modified
- Attributes can be explored
- There is tight coupling between displays and controls
- Examples: MovieFinder, LifeLines...

Dynamic queries: where else?

- Name 3 possible application areas
-
-
-