

➤ Web Information Retrieval

HCI
SOSE2023

Frank Hopfgartner, Stefania Zourlidou
Institute for Web Science and Technologies

Credit for these slides

These slides have been adapted from

- Web IR (Zeyd Boukhers-WeST, SOSE 2020)

- PageRank
 - Web graph
 - Origins
 - Motivation
 - Idea of PageRank
 - Recursive formalization
 - Random surfer
 - Formal Model

Objectives of this Lecture

- Query Specification
- Query Reformulation
- Retrieval Results Display
- Visualizing Research Results
- Design & Evaluation

➤ 1. Query Specification

- Query specification
- Retrieval result display
- Organizing search results
- Visualization in search interfaces
- Design and evaluation

- The primary methods for users to express their information need(s) are either
 - entering words into a **search entry** form
 - selecting links from a **directory** or other information organization display
- For Web search engines, the query is specified in textual form
- Typically, Web queries today are very short consisting of one to three words

- Short queries reflect the standard usage scenario
 - If the results do not look relevant, then the user reformulates their query
 - If the results are promising, then the user navigates to the most relevant-looking Web site
- This search behavior is a demonstration of the **orienteering strategy** of Web search

Query Specification

- Before the Web, search systems regularly supported **Boolean operators** and **command-based syntax**
 - However, these are often difficult for most users to understand
- Jansen *et al* conducted a study over a Web log with 1.5M queries, and found that
 - 2.1% of the queries contained Boolean operators
 - 7.6% contained other query syntax, primarily double-quotation marks for phrases
- White *et al* examined interaction logs of nearly 600,000 users, and found that
 - 1.1% of the queries contained one or more operators
 - 8.7% of the users used an operator at any time

- The standard interface for a textual query is a **search box entry form**
- Studies suggest a relationship between query length and the width of the entry form
 - Results found that either small forms discourage long queries or wide forms encourage longer queries

Query Specification Interfaces

- Some entry forms are followed by a form that filters the query in some way
- For instance, at yelp.com, the user can refine the search by location using a second form
- Notice that the yelp.com form also shows the user's home location, if it has been specified previously

The screenshot shows the Yelp website's search interface. At the top, there is a red header with the Yelp logo and the tagline "Real People. Real Reviews.™". Below the header, there is a navigation bar with links: "Welcome", "About Me", "Write a Review", "Find Reviews", "Invite Friends", and "Messaging". The main search area has two input fields: "Search for (e.g. taco, salon, Max's)" and "Near (Address, Neighborhood, City, State or Zip)". The first field contains the text "resturants" and the second field contains "washington, dc". A "Search" button is located to the right of the "Near" field. Below the "Near" field, a dropdown menu is open, showing "My Saved Locations" and "Recently Used Locations". Under "My Saved Locations", there is a list item "Home (Primary) Berkeley, CA 94705". Under "Recently Used Locations", there are two items: "Orinda, CA" and "Berkeley, CA". To the right of the dropdown menu, there is a "Memt" button and a "1 to 10 of 52" indicator. Below the search fields, the main content area displays "resturants Washington, DC" and "Did you mean: restaurants".

Query Specification Interfaces

- Some search forms show hints on what kind of information should be entered into each form
- For instance, in zvents.com search, the first box is labeled “what are you looking for”?

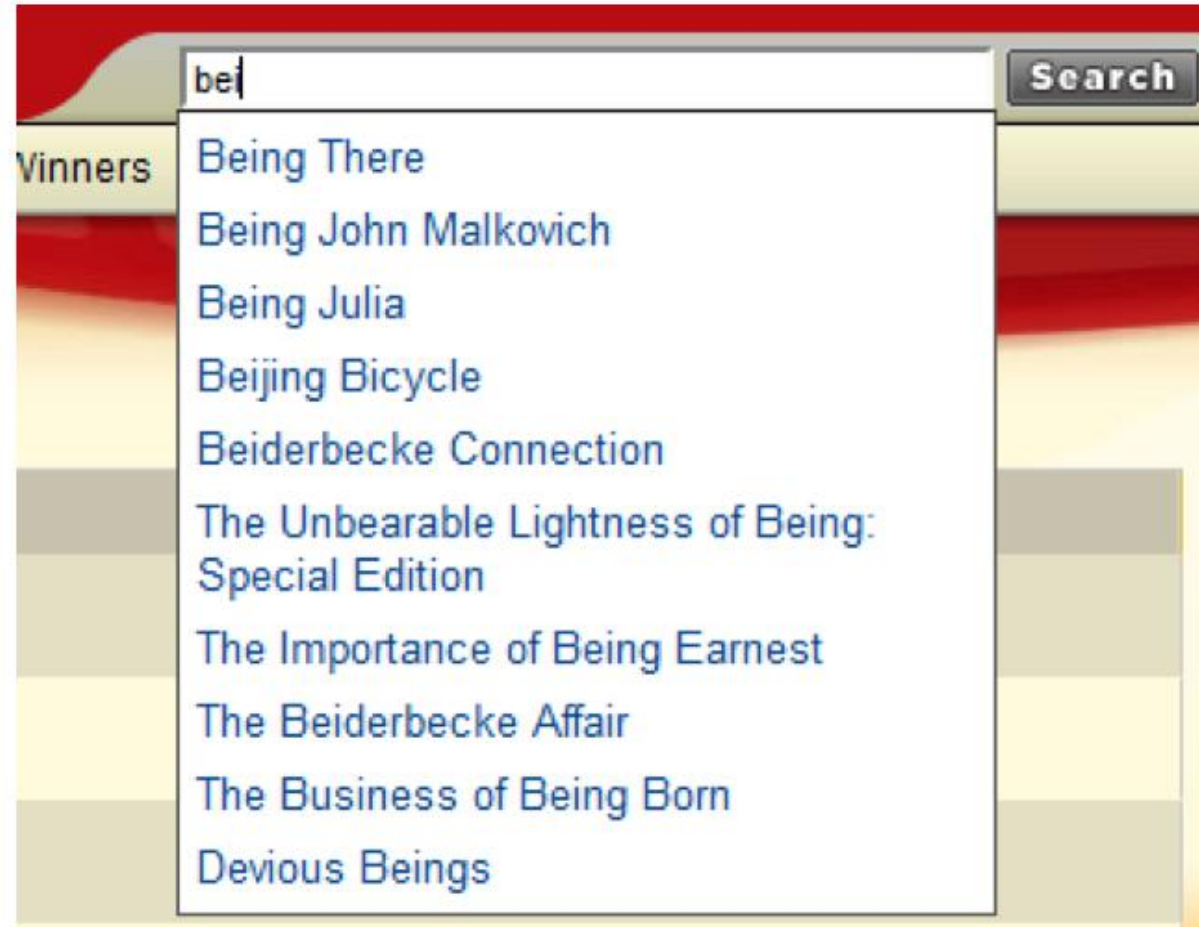
The screenshot displays the zvents.com search interface. On the left is the zvents logo with the tagline "Discover Things To Do". To the right is a search form with two input boxes: "what are you looking for?" and "when (tonight, this weekend, ...)". Above the first box is a navigation bar with links: "all" (highlighted in orange), "events", "movies", "restaurants", "venues", and "performers". Below the search form is a dark grey navigation bar with buttons for "events", "movies", "restaurants", "venues", "performers", and a "+ 6" button. A "search tip" link is visible on the right side of the search form.

- The previous example also illustrates specialized input types that some search engines are supporting today
 - The zvents.com site recognizes that words like “tomorrow” are time-sensitive
 - It also allows flexibility in the syntax of dates
- To illustrate, searching for “*comedy on wed*”, it automatically computes the date for the nearest future Wednesday
 - This is an example of how the interface can be designed to reflect how people think

- Some interfaces show a list of query suggestions as the user types the query
 - This is referred to as **auto-complete**, **auto-suggest**, or **dynamic query suggestions**
 - Anick *et al* found that users clicked on dynamic Yahoo suggestions one third of the time
- Often the suggestions shown are those whose prefix matches the characters typed so far
 - However, in some cases, suggestions are shown when only interior letters match
- Further, suggestions may be shown that are synonyms of the words typed so far

Query Specification Interfaces

- Dynamic query suggestions, from Netflix.com



- The dynamic query suggestions can be derived from several sources, including
 - The user's own query history
 - A set of metadata that a Web site's designer considers important
 - All of the text contained within a Web site

Query Specification Interfaces

- Dynamic query suggestions, grouped by type, from NextBio.com

The screenshot shows the NextBio search interface. At the top left is the 'NEXTBIO' logo. Below it, a search bar contains the text 'emb'. To the right of the search bar is a 'search' button. Below the search bar, a dropdown menu displays suggestions grouped by type. The suggestions are as follows:

- compound > EMB (Emb)
- gene > EMB (MGC71745, Gp70, AL022799, MGC21425)
- compound > EMB (Ethambutol)
- compound > EMB (Methylurethane)
- gene > Embl1
- gene > Embl2
- compound > EMBBA (Embba)
- tissue > Embryo
- compound > Embarin (Allopurinol)
- compound > Embutox (Butoxone)

On the left side of the interface, there is a 'search > embr' breadcrumb and a box labeled 'experiments(0)'. On the right side, there is a 'relevance by' label with a link 'by'.

➤ 2. Query Reformulation

- There are tools to help users **reformulate** their query
 - One technique consists of showing terms related to the query or to the documents retrieved in response to the query
- A special case of this is **spelling corrections** or **suggestions**
 - Usually only one suggested alternative is shown: clicking on that alternative re-executes the query
 - In years back, the search results were shown using the purportedly incorrect spelling

- **Term expansion:** search interfaces are increasingly employing related term suggestions
- Log studies suggest that term suggestions are a somewhat heavily-used feature in Web search
- Jansen *et al* made a log study and found that 8% of queries were generated from term suggestions
- Anick *et al* found that 6% of users who were exposed to term suggestions chose to click on them

- Some query term suggestions are based on the entire search session of the particular user
- Others are based on behavior of other users who have issued the same or similar queries in the past
 - One strategy is to show similar queries by other users
 - Another is to extract terms from documents that have been clicked on in the past by searchers who issued the same query

- **Relevance feedback** is another method whose goal is to aid in query reformulation
- The main idea is to have the user indicate which documents are relevant to their query
 - In some variations, users also indicate which terms extracted from those documents are relevant
- The system then computes a new query from this information and shows a new retrieval set

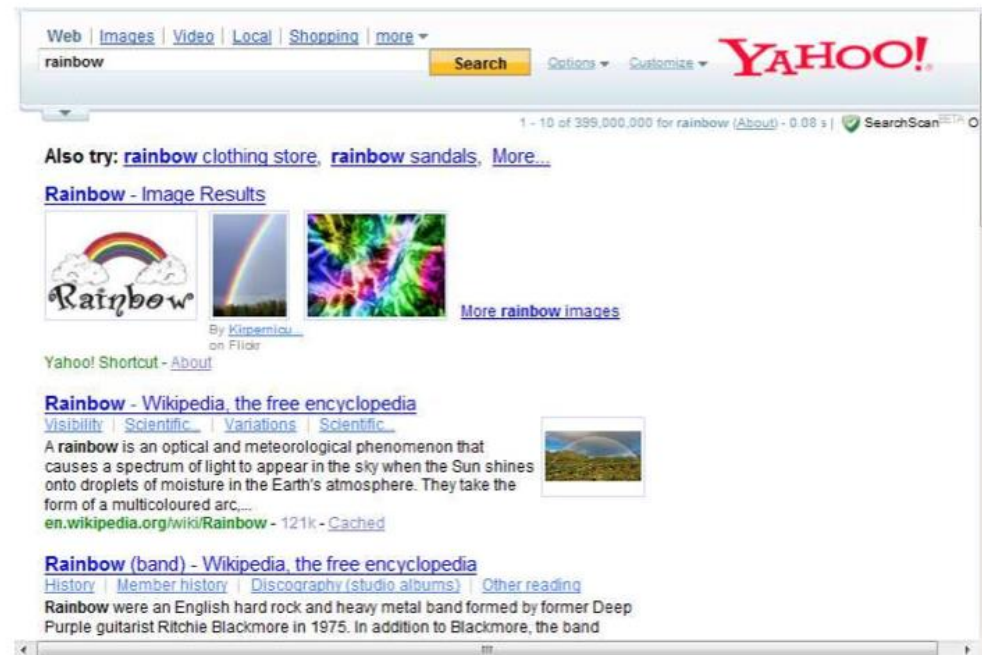
- Nonetheless, this method has not been found to be successful from a usability perspective
 - Because of that, it does not appear in standard interfaces today
- This stems from several factors
 - People are not particularly good at judging document relevance, especially for topics with which they are unfamiliar
 - The beneficial behavior of relevance feedback is inconsistent

➤ 3. Retrieval Results Display

- In Web search, the **page title** is usually shown prominently, along with the URL and other metadata
- In search over information collections, metadata such as **date published** and **author** are often displayed
- Text **summary** (or **snippet**) containing text extracted from the document is also critical
- Currently, the standard results display is a vertical list of textual summaries
- This list is sometimes referred to as the *SERP* (Search Engine Results Page)

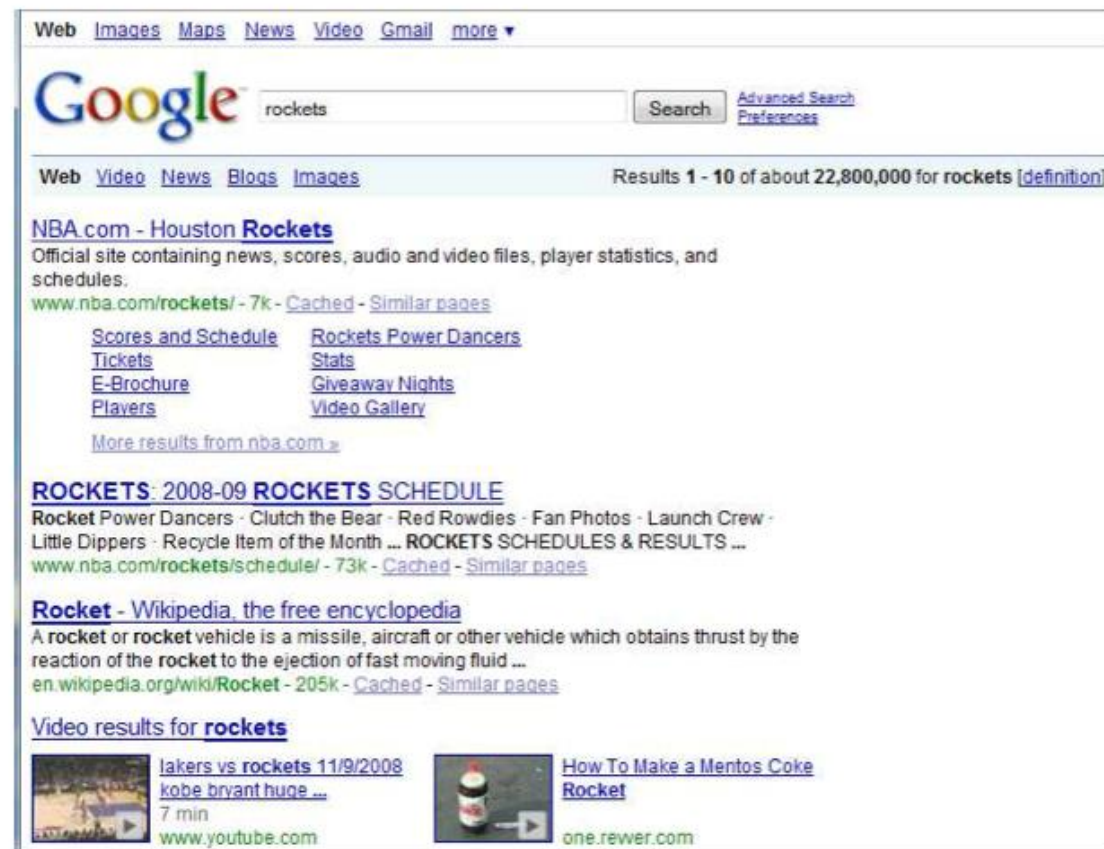
Retrieval Results Display

- In some cases the summaries are excerpts drawn from the full text that contain the query terms
- In other cases, specialized kinds of metadata are shown in addition to standard textual results
 - This technique is known as **blended results** or **universal search**



Retrieval Results Display

- A query on the name of a sports team might retrieve the latest game scores and a link to buy tickets



- Nielsen notes that in some cases the information need is satisfied directly in the search results listing
 - This makes the search engine an “answer engine”
- Displaying the query terms in the context in which they appear in the document
 - Improves the user’s ability to gauge the relevance of the results
 - It is sometimes referred to as **KWIC** - keywords in context
 - It is also known as query-biased summaries, query-oriented summaries, or user-directed summaries

- The visual effect of query **term highlighting** can also improve usability of search results listings
 - Highlighting can be shown both in document surrogates in the retrieval results and in the retrieved documents
- Determining which text to place in the summary, and how much text to show, is a challenging problem
- Often the summaries contain all the query terms in close proximity to one another
- However, there is a trade-off between
 - Showing contiguous sentences, to aid in coherence in the result
 - Showing sentences that contain the query terms

- Some results suggest that it is better to show full sentences rather than cut them off
 - On the other hand, very long sentences are usually not desirable in the results listing
- Further, the kind of information to display should vary according to the intent of the query
 - Longer results are deemed better than shorter ones for certain types of information need
 - On the other hand, abbreviated listing is preferable for navigational queries
 - Similarly, requests for factual information can be satisfied with a concise results display

Retrieval Results Display

- The page results below show figures extracted from journal articles alongside the search results

The screenshot displays the BioText Search Engine interface. At the top, the search query "CXCR4 HIV-1" is entered in the search bar. Below the search bar, there are options for "Search Over" (Full Text & Abstracts, Figure Captions (List), Figure Captions (Grid), Tables) and "Sort By" (Relevance). The "Results Page" is set to 20. The search results show 120 of 168 results for "full text".

The first result is titled "Down-regulation of cell surface CXCR4 by HIV-1" by Choi, B., Gatti, P., Fermin, C., Vigh, S., Haislip, A., Garry, R. (2008) *Virology Journal*. The abstract states: "CXCR4 chemokine receptor 4 (CXCR4), a member of the G-protein-coupled chemokine receptor family, can serve as a co-receptor along with CD4 for entry into the cell of T-cell tropic X4 human immunodeficiency virus type 1 (HIV-1) strains. Productive infection of T-lymphoblastoid cells by X4 HIV-1 markedly reduces cell-surface expression of CD4, but whether or not the co-receptor CXCR4 is down-regulated has not been conclusively determined. ... [Show Full Abstract](#)".

The "FULL-TEXT EXCERPTS" section shows: "...family function as coreceptors with the primary receptor CD4 to allow entry of various strains of human immunodeficiency virus type 1 (HIV-1) into the cells [5-8]. T-cell-tropic X4 HIV-1 use CD4 and chemokine receptor CXCR4 for entry into target cells, whereas macrophage-tropic R5 HIV-1 use CD4 and chemokine receptor CCR5. Dual-tropic strains can use either CCR5 and CXCR4 as co-receptors... [Show Full Excerpts](#)".

Below the abstract, there is a link to "VIEW FULL ARTICLE: [HTML](#) | [PDF](#)".

To the right of the abstract, there is a section titled "FIGURES FROM ARTICLE:" which displays several small thumbnail images of figures from the article, including flow cytometry plots and microscopy images.

The second result is titled "Differential control of CXCR4 and CD4 downregulation by HIV-1 Gag" by Vallathan, R., Resh, M. (2008) *Virology Journal*. The abstract states: "The ESCRT (endosomal sorting complex required for transport) machinery functions to sort cellular receptors into the lumen of the multivesicular body (MVB) prior to lysosomal ...".

Below the abstract, there is a link to "VIEW FULL ARTICLE: [HTML](#) | [PDF](#)".

To the right of the abstract, there is a section titled "FIGURES FROM ARTICLE:" which displays several small thumbnail images of figures from the article, including flow cytometry plots and microscopy images.

➤ 4. Organizing Search Results

- Organizing results into meaningful groups can help users understand the results and decide what to do next
- Popular methods for grouping search results: **category systems** and **clustering**
- **Category system:** meaningful labels organized in such a way as to reflect the concepts relevant to a domain
 - Good category systems have the characteristics of being coherent and relatively complete
 - Their structure is predictable and consistent across search results for an information collection

Organizing Search Results

Images from the Collections of the Fine Arts Museums of San Francisco, Legion of Honor and de Young Museums, <http://www.thinkart.org>

Save Search History and Settings [View Search](#) [New Search](#)

☒ all items ☐ in current results

Refine your search within these categories:

MEDIA: [all](#) > Print

aquatint (4)	lithograph (21)
drypoint (10)	mezzotint (14)
engraving (50)	woodcut (12)
etching (77)	

LOCATION: [all](#) > Europe ([group results](#))

Austria (1)	Italy (14)
Belgium / Flanders (5)	Scotland (5)
Bohemia (8)	Spain (1)
France (27)	Switzerland (2)
Germany (19)	more...
Holland (24)	

OBJECTS ([group results](#))

Clothing (69)	Musical Instruments (4)
Containers (21)	Vehicles (56)
Food and Meals (45)	Weapons (27)
Fuel (2)	Writing Tools (13)
Lighting (2)	

BUILT PLACES ([group results](#))

Bridge (18)	Dwelling (107)
Building (56)	Part of Building (44)
Built Open Space (14)	Road (21)

ANIMALS AND PLANTS ([group results](#))

Birds (19)	Mammals, Hoofed (43)
Cattle and Domestic	Mammals, Other

These terms define your current search. Click the to remove a term.





keyword "castle"

LOCATION: Europe





MEDIA: Print

197 items, grouped by MEDIA ([view ungrouped items](#))

aquatint (4)

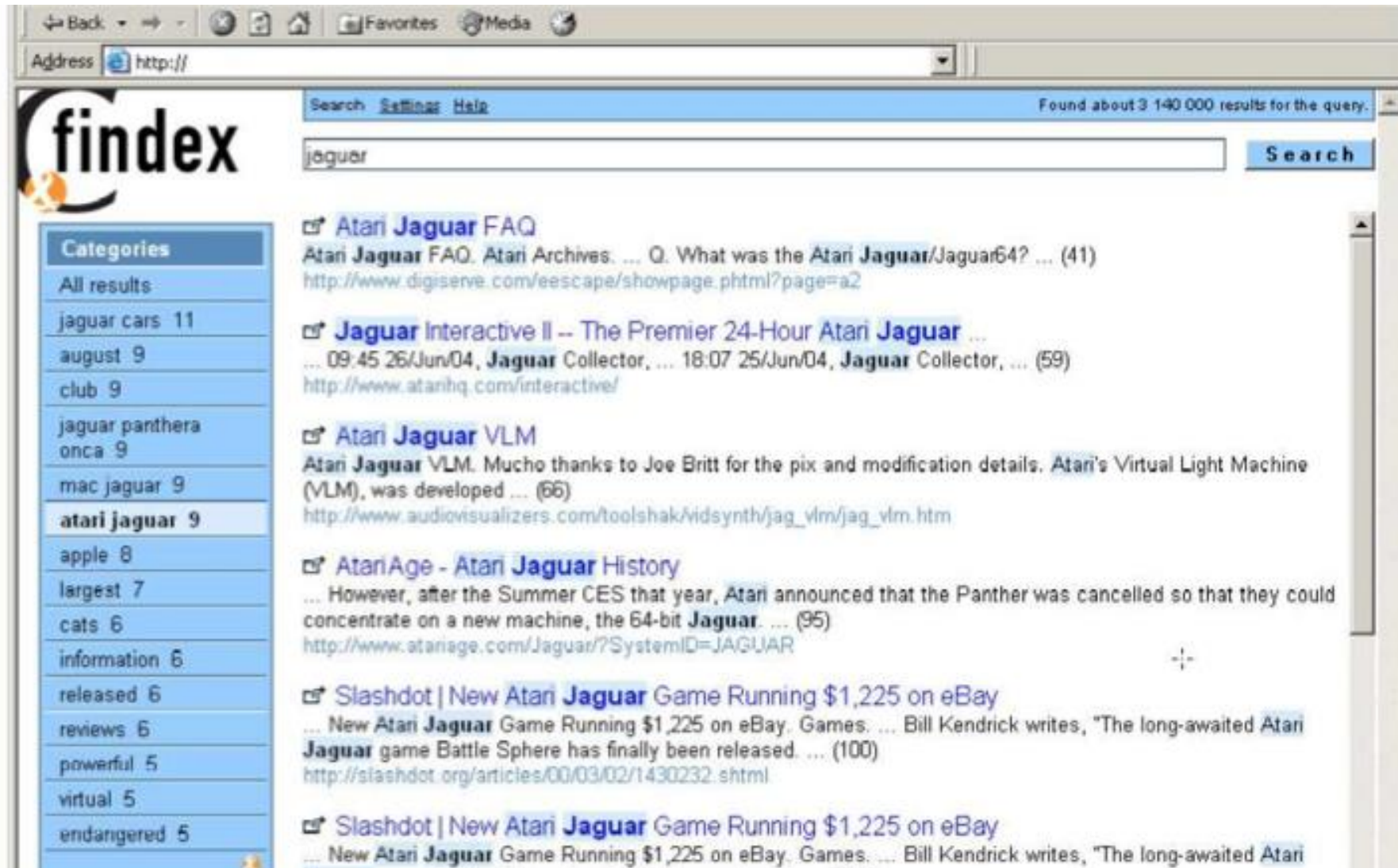
			
Caernarfon Castle... 18th - 19th century	Duntanborough Castle 1808	Edinburgh Castle N... 1801	Untitled (landscap... circa 1780

drypoint (10)

			
Lindesfarne Castle 19th - 20th century	Stirling Castle, N... 19th - 20th century	Castle Moly 19th - 20th century	landscape with a ... 19th - 20th century

- **Clustering** refers to the grouping of items according to some measure of similarity
- It groups together documents that are similar to one another but different from the rest of the collection
 - Such as all the document written in Japanese that appear in a collection of primarily English articles
- The greatest advantage of clustering is that it is fully automatable
- The disadvantages of clustering include
 - an unpredictability in the form and quality of results
 - the difficulty of labeling the groups
 - the counter-intuitiveness of cluster sub-hierarchies

Organizing Search Results



Organizing Search Results



Web [News](#) [Images](#) [Video](#)

senate

Search

Results **1-20** of about **133,005,879** | [Details](#)

[Sources](#) [Sites](#) [Time](#) [Topics](#)

Top 879 Clusters remix

- + U.S. Senate (160)
- + State Senate (118)
- + Police, Reform (40)
- + Small Businesses (41)
- + Relief (29)
- + University, Student (35)
- + Biden (24)
- + China (22)
- + Mitch McConnell (26)
- + History (24)
- + Senate Hearing (24)
- + Photos, Biography (29)
- + Floyd (12)
- + Senate confirms (17)
- + Fauci (11)
- + Burr, Chair (11)
- + Immigration (13)
- + Ohio (15)
- + Kentucky, McGrath (12)

1. [U.S. Senate](#) [new window](#) [preview](#)

Date: 2020-06-24T08:27:00

Senate Stories. Presented to enlighten, amuse, and inform, our new **Senate** history blog explores the forces, events, and personalities that have shaped the modern **Senate**. The **Senate** and Women's Fight for the Vote. Learn about the **Senate**'s long road to approving the Nineteenth Amendment.

<https://www.senate.gov> - Yippy Index V

2. [U.S. Senate: Senators](#) [new window](#) [preview](#)

Date: 2020-06-24T08:30:00

• Senators who have Cast more than 10,000 Votes • Senators who have Delivered Washington's Farewell Address. Political Parties • Senators Representing Third or Minor Parties • Senators who Changed Parties during **Senate** Service (since 1890) Diversity in the **Senate** • Foreign-born Senators • Ethnic Diversity in the **Senate** • Women ...

<https://www.senate.gov/senators/contact> - Yippy Index V

3. [United States Senate - Wikipedia](#) [new window](#) [preview](#)

Date: 2020-06-24T20:44:00

The United States **Senate** is the upper chamber of the United States Congress, which, along with the United States House of Representatives—the lower chamber—constitutes the legislature of the United States. The **Senate** chamber is located in the north wing of the Capitol Building in Washington, D.C.. The composition and powers of the **Senate** are established by Article One of the United States ...

https://en.wikipedia.org/wiki/United_States_Senate - Yippy Index V

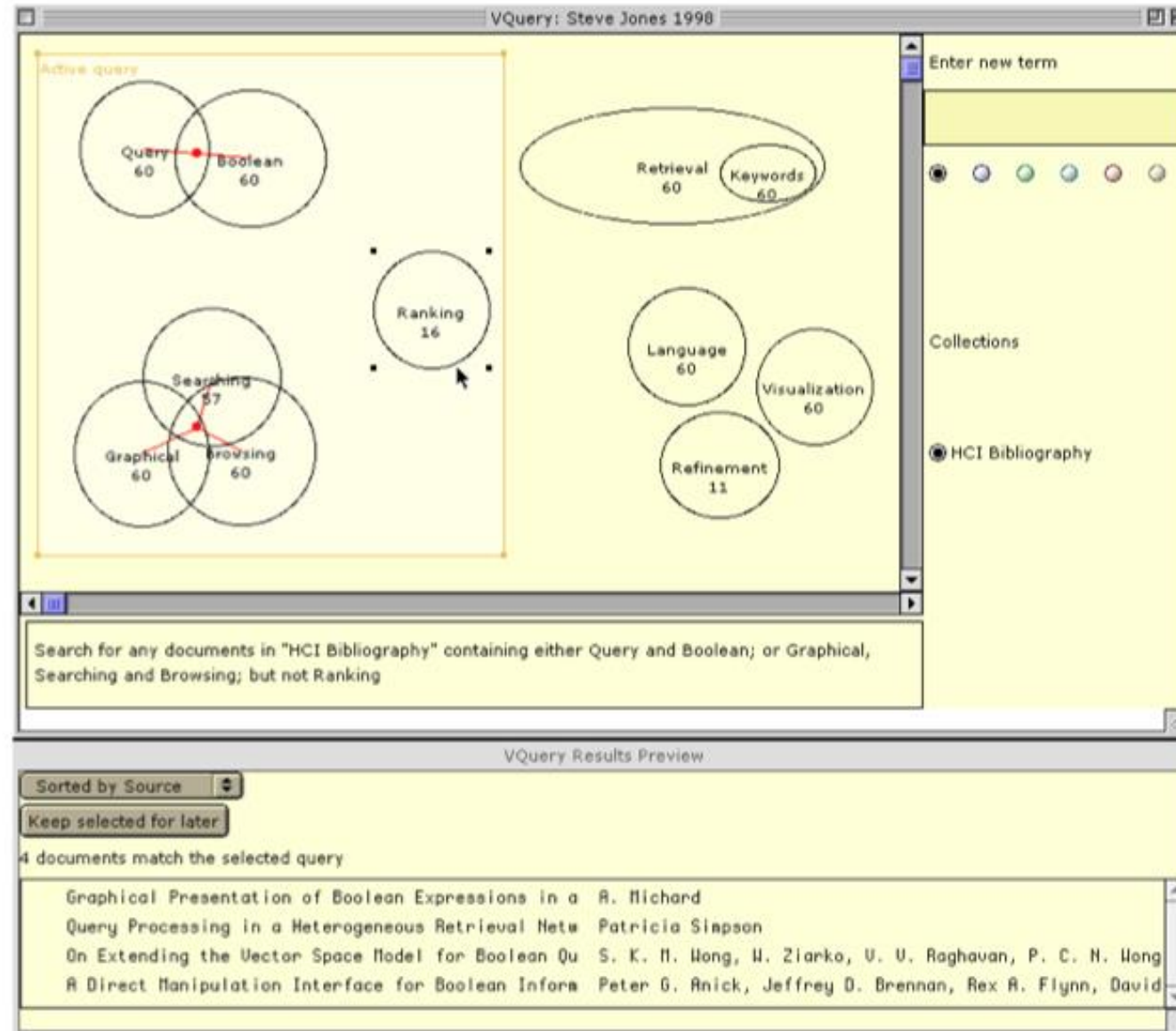
4. [U.S. Senate | USAGov](#) [new window](#) [preview](#)

➤ 5. Visualization in Search Interfaces

- Experimentation with visualization for search has been primarily applied in the following ways
 - Visualizing Boolean syntax
 - Visualizing query terms within retrieval results
 - Visualizing relationships among words and documents
 - Visualization for text mining

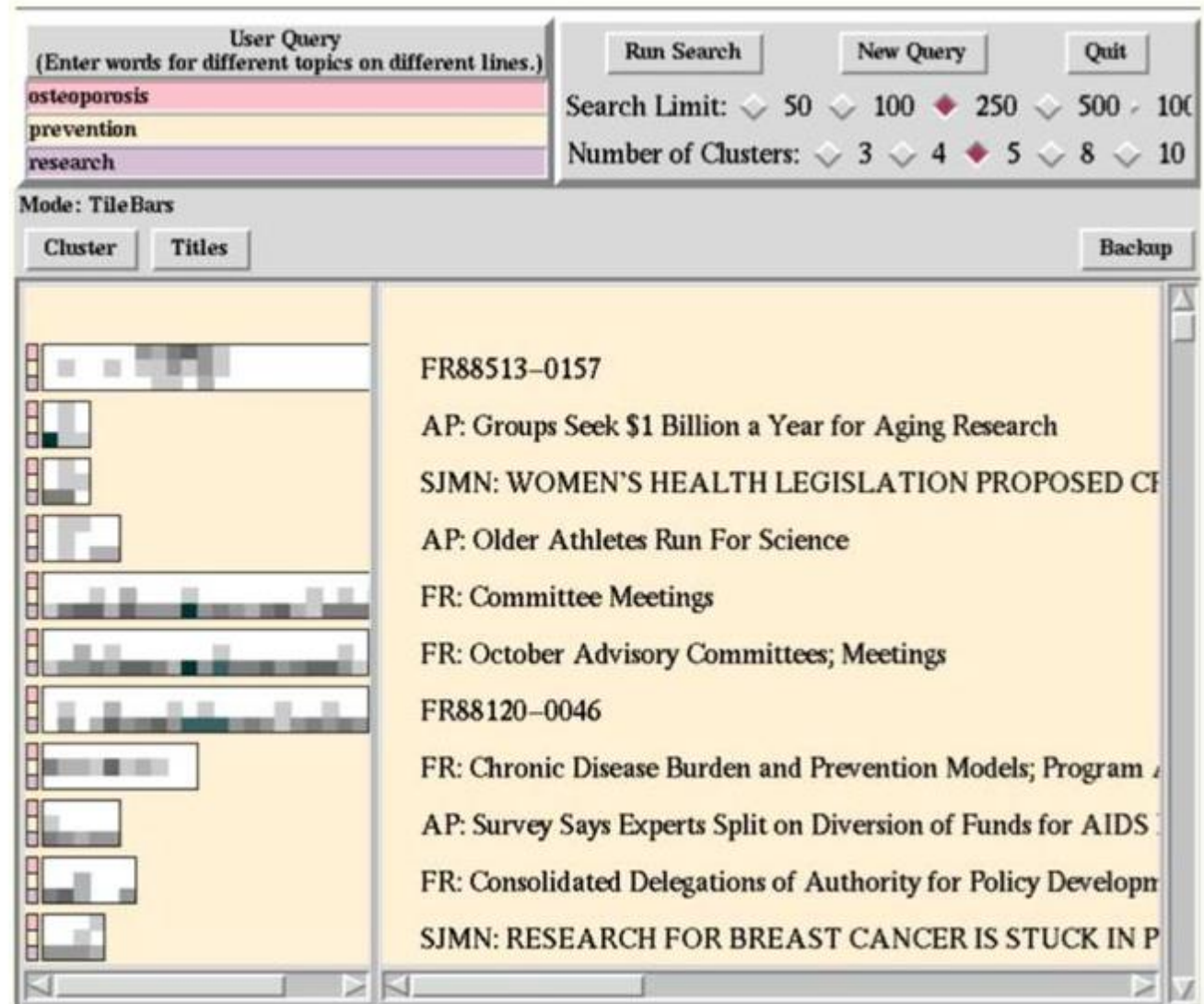
- Boolean query syntax is difficult for most users and is rarely used in Web search
- For many years, researchers have experimented with how to visualize Boolean query specification
- A common approach is to show **Venn diagrams**
- A more flexible version of this idea was seen in the **VQuery system**, proposed by Steve Jones

Visualizing Boolean Syntax



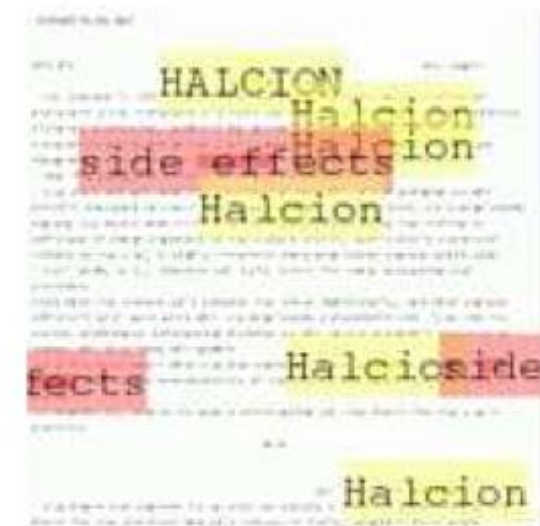
Visualizing Query Terms

The TileBars interface



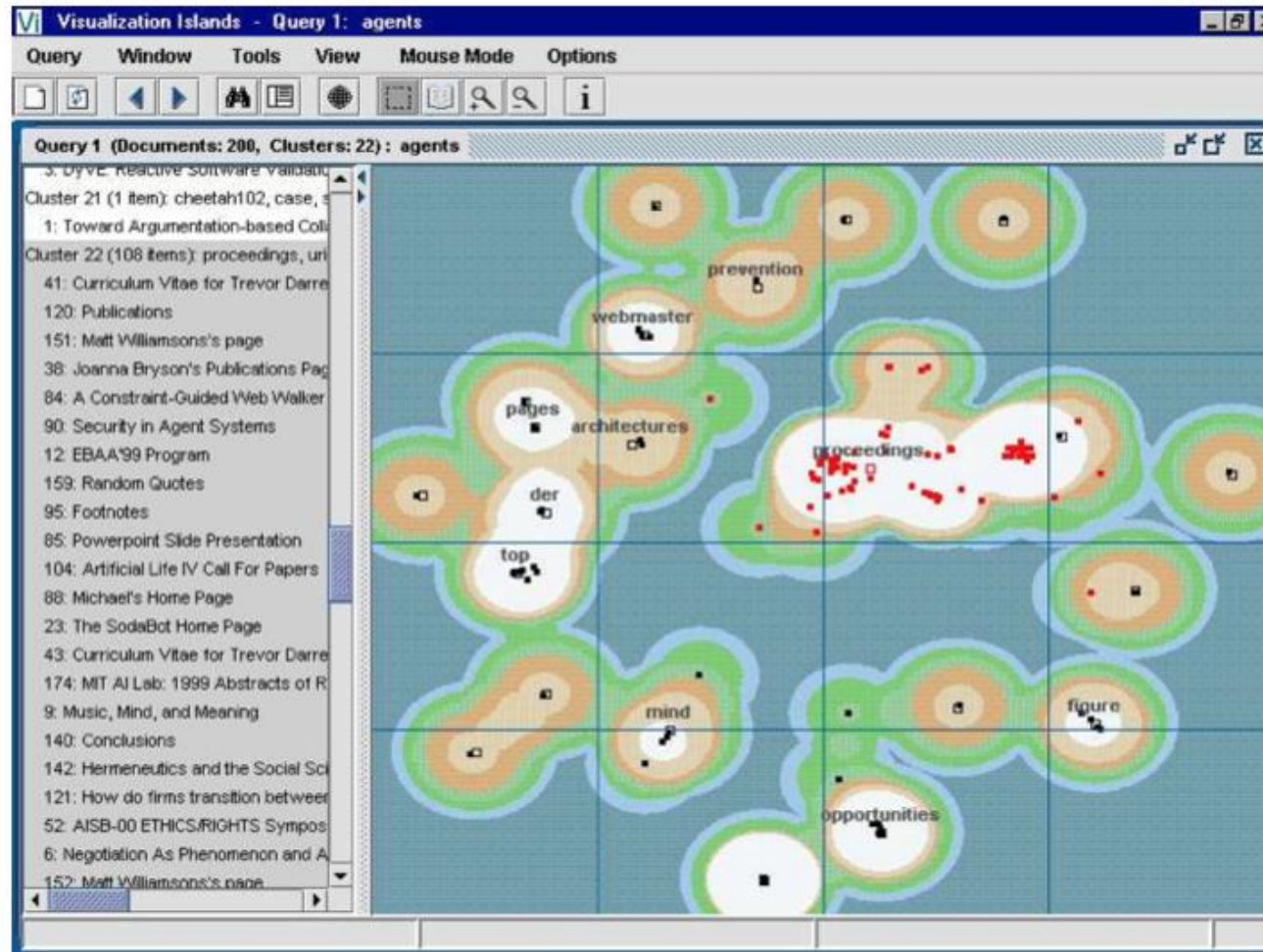
Visualizing Query Terms

Textually enhanced
thumbnails



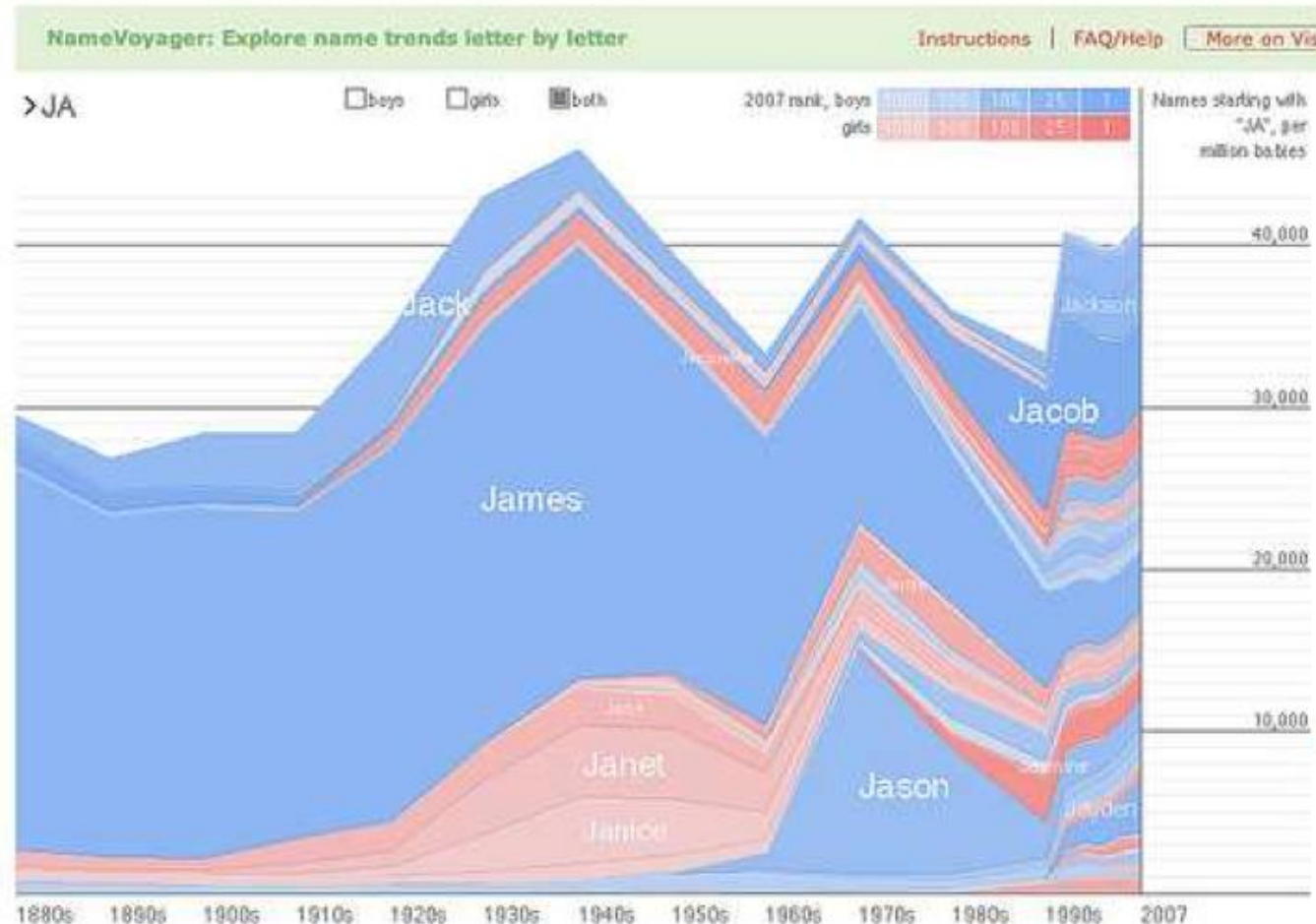
Words and Docs Relationships

xFIND's VisIslands, from Andrews *et al*



Words and Docs Relationships

- The popularity of baby names over time (names beginning with JA), from babynamewizard.com



➤ 6. Design & Evaluation

- User interface design: a field of Human-Computer Interaction (HCI)
- This field studies how people think about, respond to, and use technology
- User-centered design: a set of practices developed to facilitate the design of interfaces
- The design process begins by determining what the intended users' **goals** are
- Then, the interface is devised to help people achieve those goals by completing a series of **tasks**

- Evaluating a user interface is often different from evaluating a ranking algorithm or a crawling technique
 - A crawler can be assessed by crisp quantitative metrics such as coverage and freshness
 - A ranking algorithm can be evaluated by precision, recall, and speed
- The quality of a user interface is determined by how people respond to it
- Subjective responses are as, if not more, important than quantitative measures
- If a person has a choice between two systems, they will use the one they prefer

- How best to evaluate a user interface depends on the current stage in the development cycle
- When starting with a new design or idea, **discount** usability methods are typically used
 - Example: showing a few users different designs asking them to indicate which parts are promising and which are not
- Another commonly used discount evaluation method is **heuristic evaluation**
 - Usability experts “walk through” a design and evaluate the functionality in accordance with a set of design guidelines

- A formal experiment must be carefully designed to take into account potentially confounding factors
 - For instance, it is important for participants to be motivated to do well on the task
- This kind of study can uncover important subjective results
 - Such as whether a new design is strongly preferred over a baseline
- However, it is difficult to find accurate quantitative differences with a small number of participants

- Another problem: the timing variable is not the right measure for evaluating an interactive search session
 - A tool that allows the users to learn about their subject matter as they search may be more beneficial, but take more time
- Two approaches to evaluating search interfaces have gained in popularity in recent years
 - One is to conduct a **longitudinal study**
 - Participants use a new interface for an extended period of time, and their usage is monitored and logged
 - Evaluation is based both on log analysis and questionnaires and interviews with the participants

- Another evaluation technique is to perform experiments on already heavily-used Web sites
- Consider a search engine that receives millions of queries a day
 - a randomly selected subset of the users is shown a new design
 - their actions are logged and compared to another randomly selected control group that continues to use the existing interface
 - this approach is often referred to as **bucket testing, A/B testing**

➤ 4. Summary

Summary

- Query Specification
- Query Reformulation
- Retrieval Results Display
- Visualizing Research Results
- Design & Evaluation

[1] <https://olat.vcrp.de/auth/RepositoryEntry/4071063853>

[2] <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze,
Introduction to Information Retrieval, Cambridge University Press. 2008

[3] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Professional, 2011.

► Chapter 10: User Interfaces and Visualization, available [here](#)