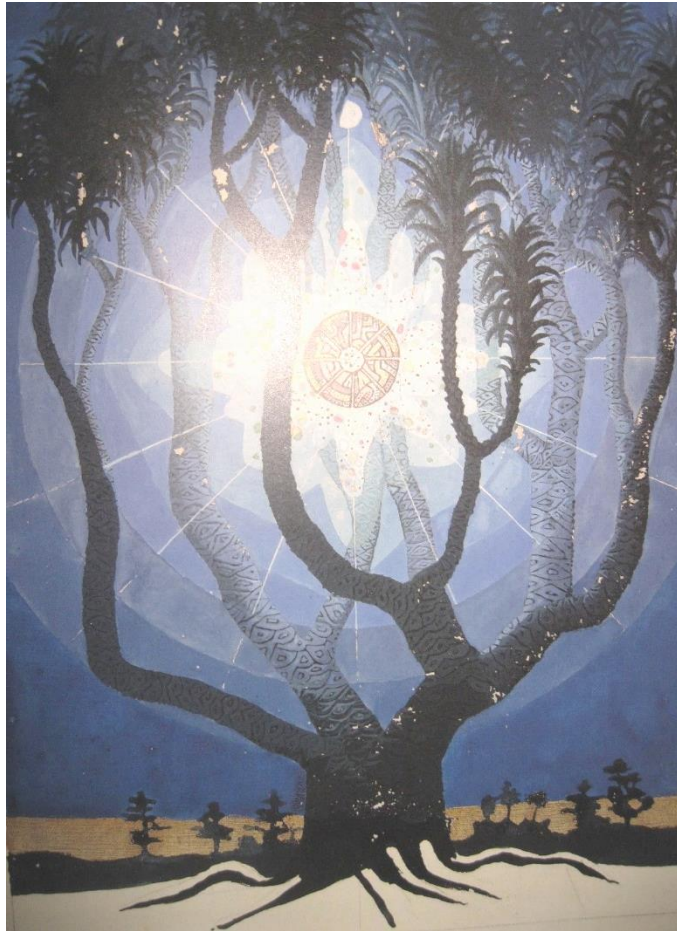


Introduction to Information Retrieval

Essentials on IR and search interfaces



Outline



IR definition and origins

Data: structure levels

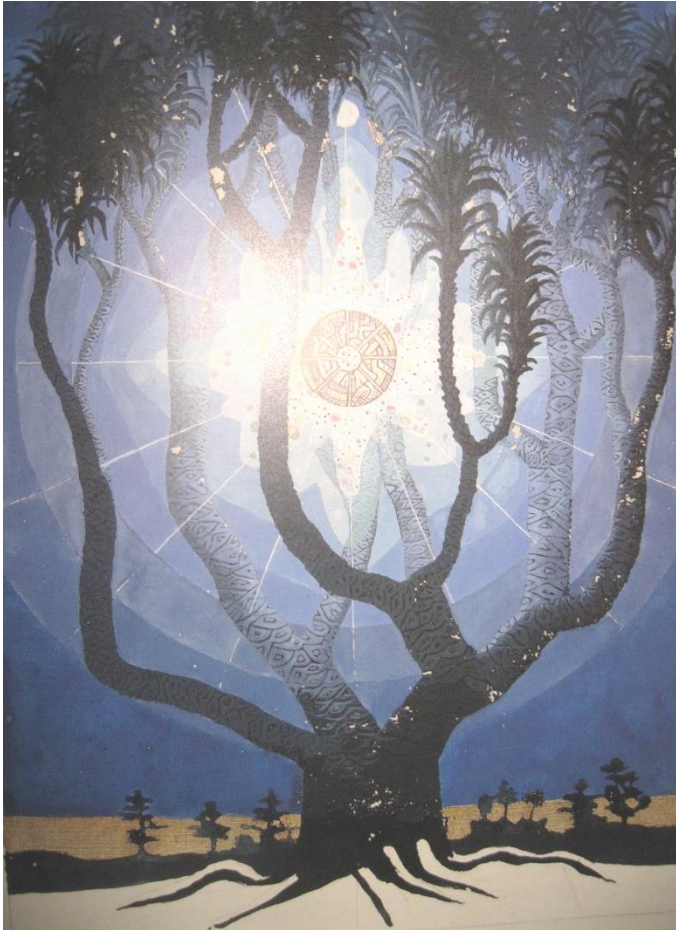
IR systems

Dynamic search model

Search interfaces

- **query support**
- **results display**
- **evaluation**

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Information retrieval

What is *Information Retrieval* (IR)?

- **processing** of **libraries** to answer an **information need**

The processing of libraries commonly entails:

- **finding** and **organizing** material (usually **documents**) of an **unstructured** nature (usually text)
 - from a **large collection** (usually in digital format)
-
- **document** = book, webpage, media file, article, form
 - **library = document collection**
 - also known as **corpus** if the collection refers to a well-defined topic
 - *static*: well-defined set of documents
 - *dynamic*: documents can be added and removed along time

Information retrieval: application domains

When we think of IR we think first of ***web search***, ***fact retrieval*** and ***question answering***... yet:

- e-mail and laptop search
- corporate knowledge bases
- academic bases (peer-reviewed articles)
- legal information retrieval

... retrieval at different levels:

- **global** (web)
- **organizational** (institutions, companies, academy)
- **individual** (personal computers, e-mail, social media)

Information retrieval: tasks

Early core task of IR:

- **searching** (also known as querying)
Given a collection, retrieve **relevant** documents to the user's **information need** (helping the user complete a **task**)

Nowadays, research in IR:

- **crawling** and **indexing** massive collections
- **fact retrieval**
- **question answering** using **large language models**
- **browsing** (also known as navigation)
- document **annotation** (e.g., **fake text detection** and **sentiment analysis**)
- corpus **organization**
- **cross-language** retrieval
- document **summarization** and restyling...

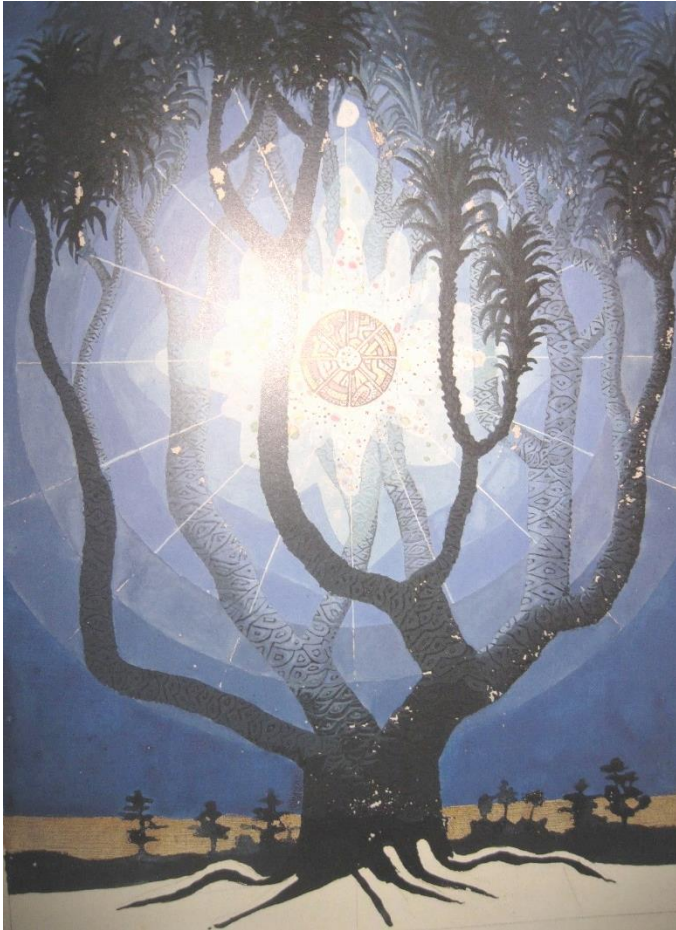
The beginning...

- For over 5000 years, man has organized information for later retrieval and searching
 - compiling, storing, organizing, and indexing **hieroglyphics**, **papyrus** and **books**
- To hold the various items:
 - *physical libraries* were built
 - oldest known library in Elba in the Fertile Crescent (between 3000 and 2500 BC)
 - Great Library at Alexandria (300 BC)
 - since then: libraries everywhere
 - with the advent of informatics: documents are digitalized and maintained within **digital libraries**
 - from librarians to information and companies everywhere
 - nowadays: a global digital library – the **Web**

From manual to automatic IR

- Volume of information in physical, digital and global libraries growing...
 - necessary to build specialized data structures for efficient search —*indexes*
 - for centuries: *manual indexing* (e.g., card catalogs' searching)
 - since 50s: *automatic indexing* and Boolean querying
 - today: enriched with advanced query support, graphical search interfaces, hypertext features
- Brief **timeline**
 - IR concepts introduced in 50's
 - in late 60's: the *TF-IDF term weighting scheme*, largely used up to date
 - in 70's: first ACM SIGIR International Conference on Information Retrieval held in Rochester

Outline



IR definition and origins

Data: structure levels

IR system

Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Structured vs unstructured data

- **Structured data**

- ... from measuring systems along time
 - geophysical, digital, mechanical, physiological, biological, societal, organizational, hybrid systems
- typically organized in series (sensorized systems), relational and multi-dimensional formats

- **Unstructured data**

- ... from ad-hoc and knowledge systems
 - communication acts (whether spontaneous, opinions, published material)
from individuals and organizations
- typically in reference to free text (written or transcribed)

- What about **media data**?

- audio, image and video

Structured data

- searching structured data: ***data engineering*** (databases)
 - well defined structure and semantics
 - a single erroneous object among a thousand of retrieved objects means failure

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

... seeks exact match queries and allows diverse numerical operators

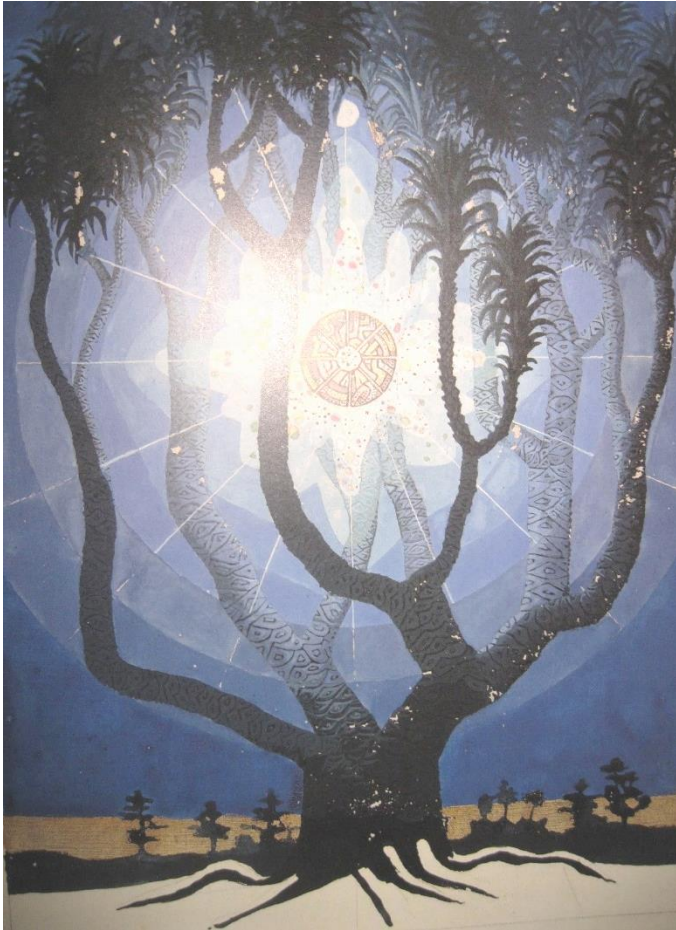
Salary < 60000 AND Manager = Smith

- learning models from structured data: ***machine learning***

Semi-structured data

- In fact almost no data is *unstructured*
 - *this slide* has distinctly identified zones such as *Title* and *Bullets*
 - as well as **linguistic structure!**
- Semi-structured data
 - research articles: zones?
 - web pages: zones? enriched text?
 - what about XML data?
- Examples on searching semi-structured data:
 - *title* contains data AND *bullets* contain search
 - *title* is about object-oriented programming AND *author* is stro*rup
where * is a wildcard operator

Outline



IR definition and origins

Data: structure levels

IR systems

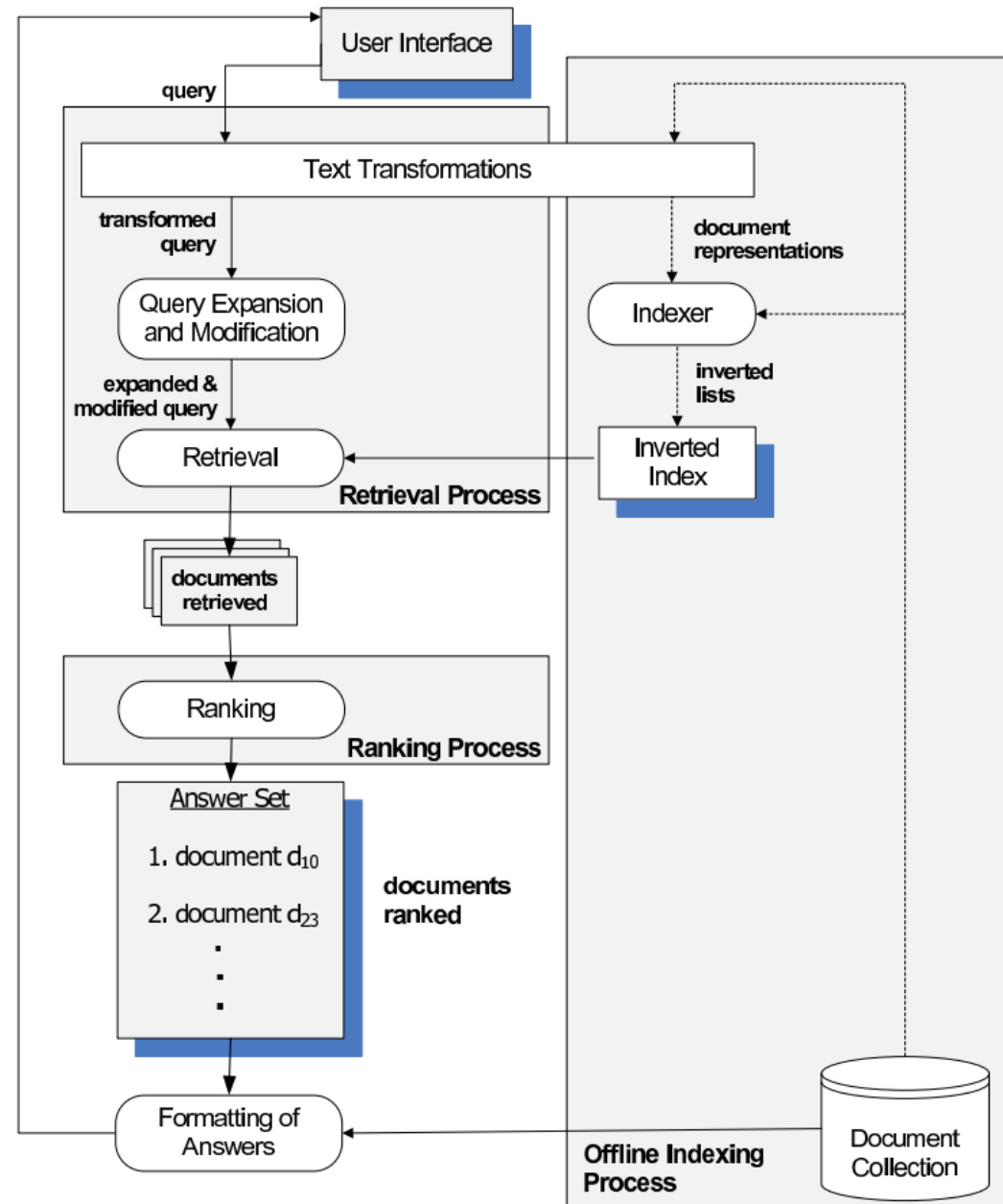
Dynamic search model

Search interfaces

- query support
- results display
- evaluation

IR system

- a **computational system** to support information retrieval
- document **processing, indexing, retrieval, ranking, and annotation** as core IR tasks
- let us focus on a specific task: **document retrieval**
 - *architecture on the left*



How good is the retrieved information?

IR systems may correctly answer a **query**, yet not necessarily satisfy the **information need**

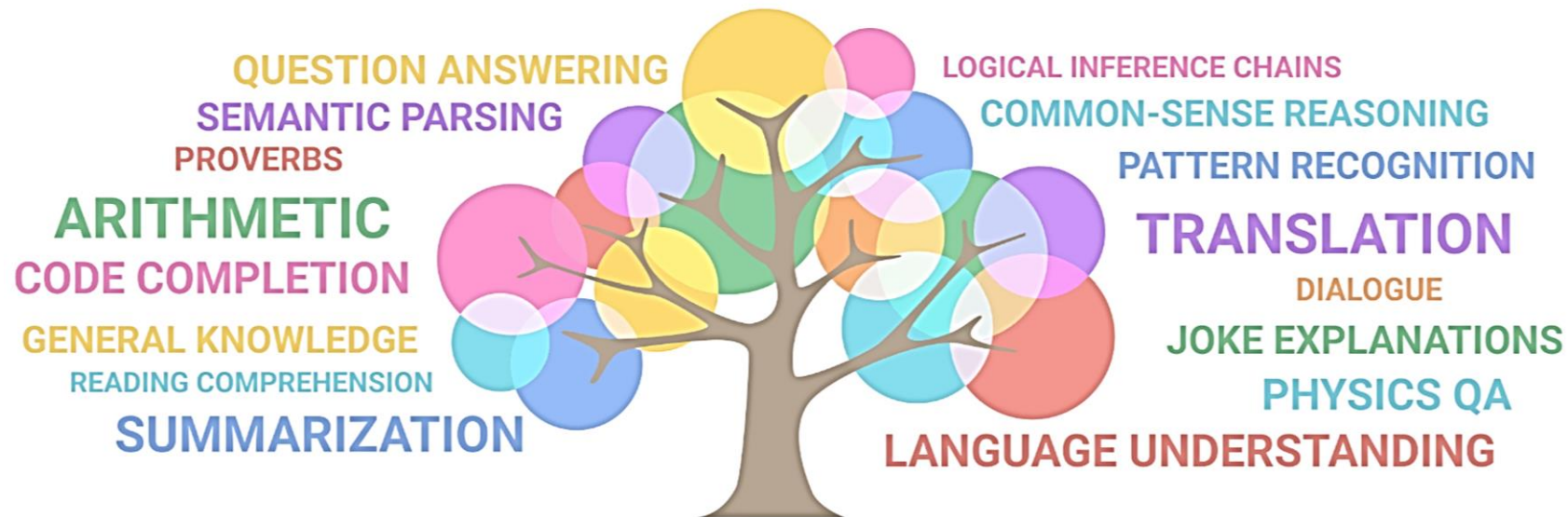
- “*wine death*” (query) to learn about reduced cardiovascular risks (information need)
may return information on car deaths under alcohol effect (ok for the query yet not the need)

Considering **searching** (document retrieval)...

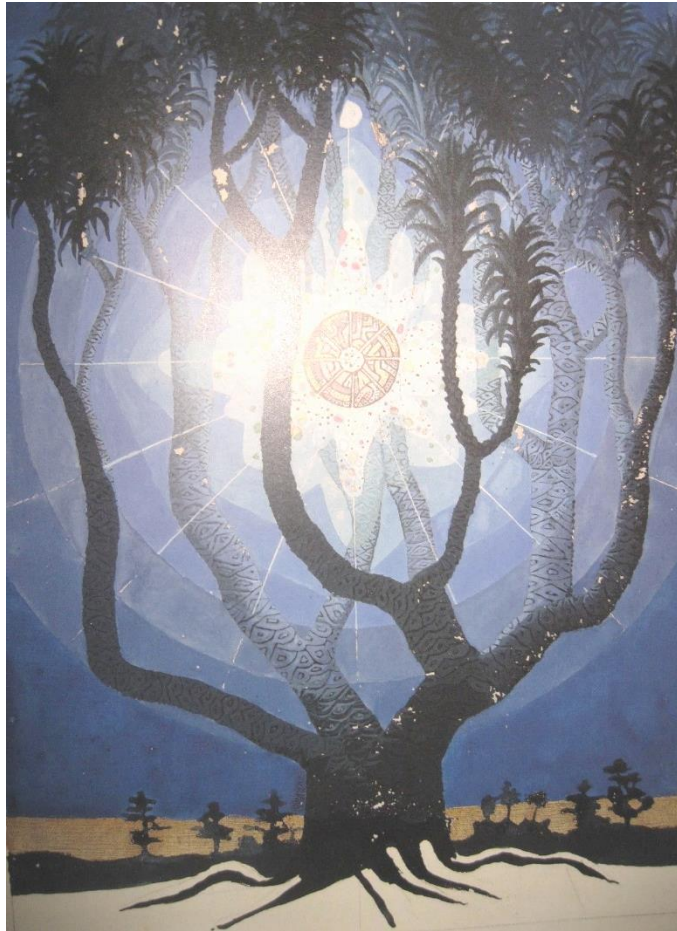
- **goal**: an IR system able to retrieve all relevant items to a user query while retrieving as few non-relevant items as possible
 - **precision**: fraction of retrieved docs that are relevant to the user’s information need
 - **recall**: fraction of relevant docs in collection that are retrieved
- simple retrieval *versus* **ranking**
 - ranking tasks order information items according to a degree of relevance to the user query
 - the notion of relevance is of central importance in IR!
 - evaluation of ranking outputs require additional metrics!

Going beyond document retrieval...

- How do we extend our IR systems to handle additional tasks?
How do we evaluate the success of IR systems for those tasks?
 - annotating documents (e.g., tags, spam e-mail, fake web content, hateful text)
 - manipulating documents (e.g., translation, restyling, summarization)
 - answering natural language questions



Outline



IR definition and origins

Data: structure levels

IR systems

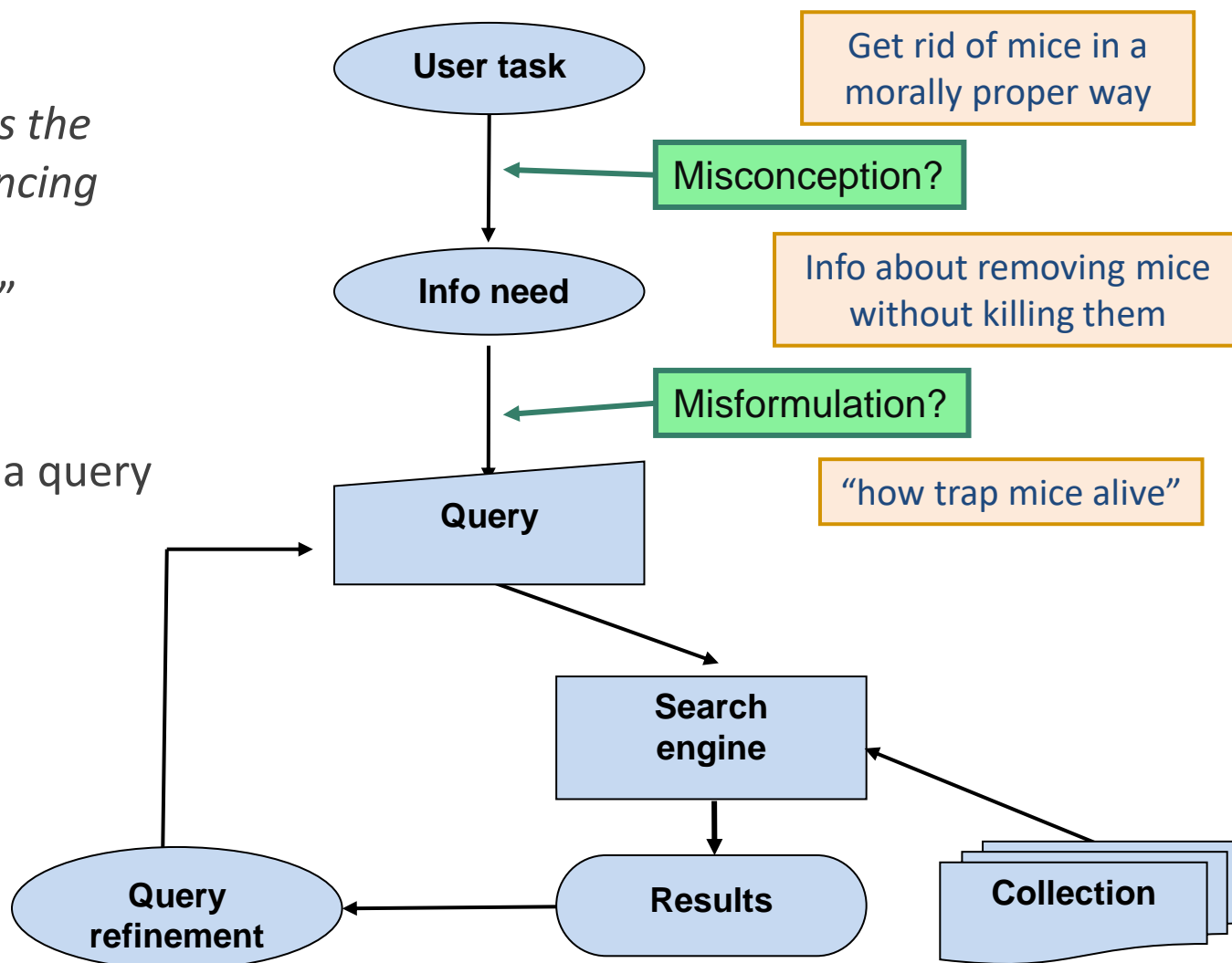
Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Classic search model

- **need:** *“Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)”*
- **querying** process
 - Translate the information need into a query
 - executing query on the IR system



Classic *versus* dynamic search model

- **Classic** information seeking process:
 1. problem identification
 2. articulation of information need(s)
 3. query formulation
 4. results assessment
- More recent models emphasize the **dynamic nature** of the search process:
 - users learn as they search
 - information needs adjust as they see retrieval results and document surrogates (**orienteering**)
- This dynamic process is sometimes referred to as the **berry picking** model of search
 - motivated by the rapid response times of today's web search engines
 - [Jansen *et al.*] search logs reveal that 52% of users commonly modify their queries

Types of searches

Marchionini makes a distinction between:

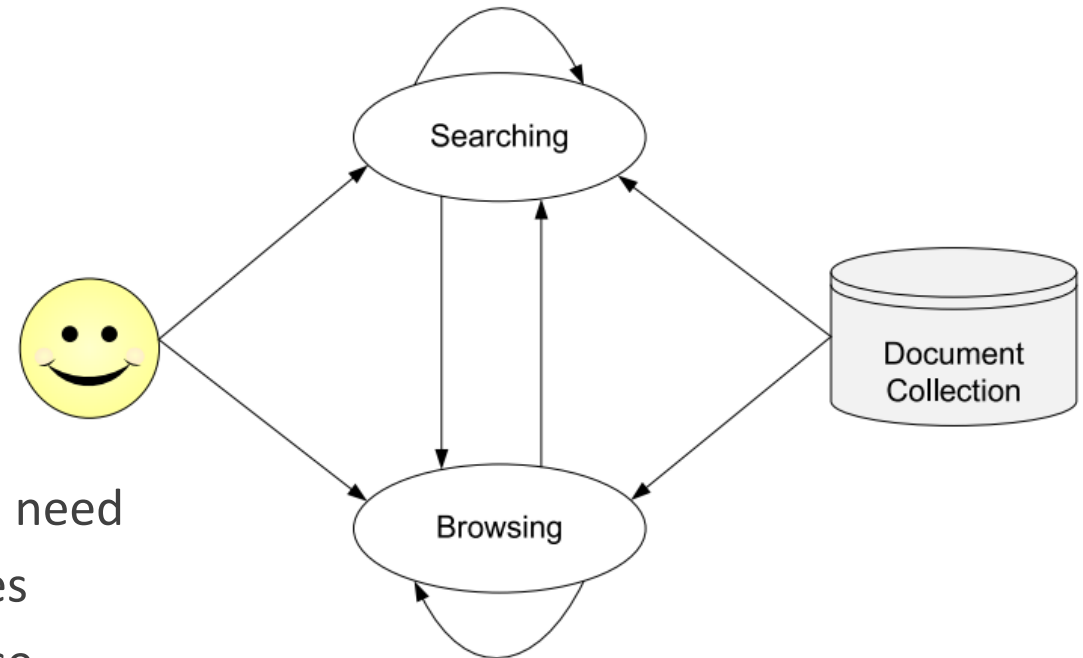
- **information lookup** tasks
 - can be satisfied by discrete pieces of information: numbers, dates, names or websites
 - *searching, fact retrieval or question answering*
- **exploratory** tasks
 - **learning**
 - requires more than a single query-response interaction
 - requires time scanning multiple items and synthesizing content to form new understanding
 - **investigation**
 - longer-term process requiring iterations that take place over perhaps long periods of time
 - example: finding a large portion of relevant information available

In the course, we further consider **document annotation** and **manipulation** tasks as part of IR

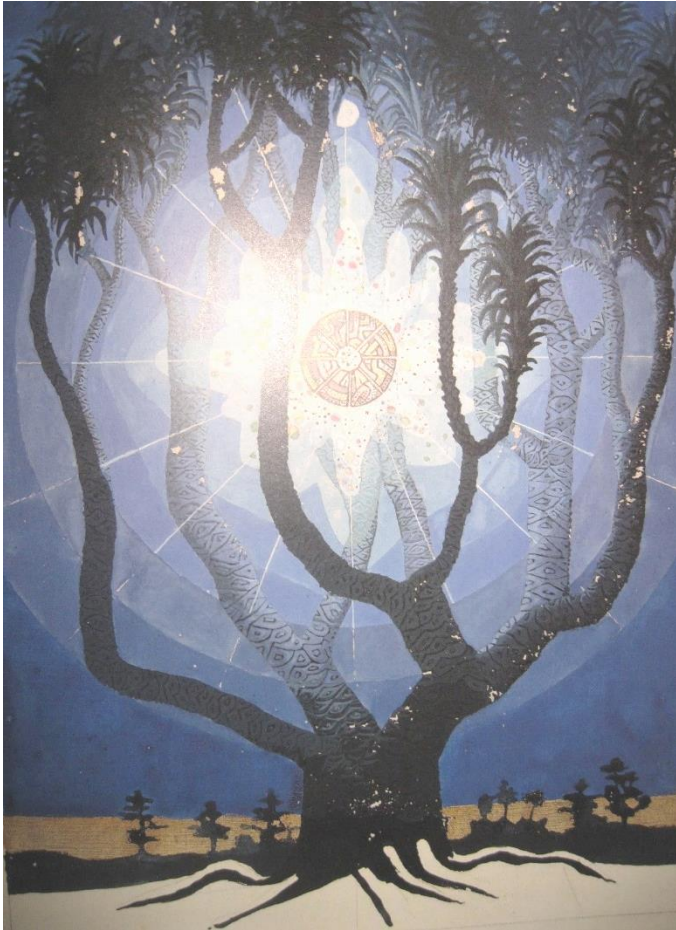
Hybrid searching and browsing

Lookup and exploratory searches can entail **browsing**

- navigating through documents of a given collection
- browsing strategy engaged when:
 - seeking information on a topic that is either poorly defined or inherently broad \Rightarrow *searching* and *browsing*
 - seeking information on a well-defined topic of interest \Rightarrow *searching*
 - the information structure well-matches the information need
 - for mentally less taxing recognition of information pieces
 - there are appropriate links among documents, otherwise browsing experience is frustrating



Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Aiding information retrieval

- **Aiding IR systems**
 - supportive **querying interfaces**
 - organization and **visualization** of IR outputs
 - guidance from **relevance feedback**: supervised IR
 - proactive **corpus exploration** (relations among documents, tagging): unsupervised IR
 - advanced **text processing**: natural language processing (e.g., query expansion, correction)
- IR systems are described by three major components
 - **search interface module** (facilities to input queries and output results) => **today!**
 - **indexing module** (facilities to crawl and store document collections) => **next** class!
 - **processing module** (the IR brain to answer queries) => our focus for the period!

Insights from user behavior

- Searchers typically...
 - **reformulate** their queries with slight modifications
 - look **only** to the **top-ranked** retrieved results or to the **highlights** of a text answer
 - biasedly think the **top documents** (or initial paragraphs) **are better** than those beneath
 - poorly estimate how much of the relevant material was found
 - **not** particularly good at **judging the relevance of information** (especially for unfamiliar topics)
 - search for information previously accessed
 - search strategies differ when searching over previously seen and new materials
 - search interfaces should support **query history** and **revisitation**

Search interfaces

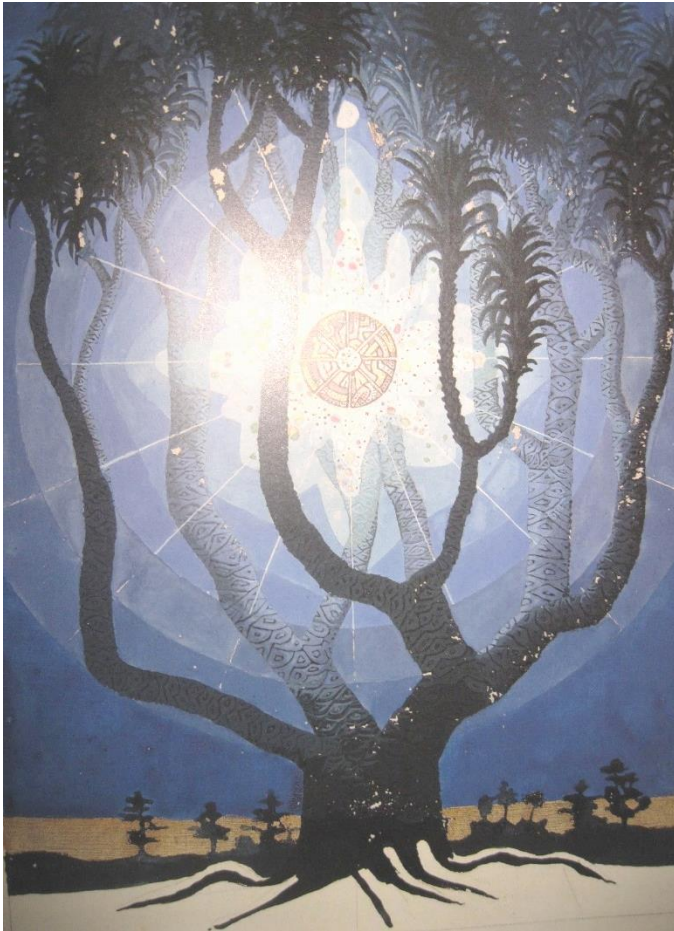
- User interaction with search interfaces differs depending on:

- the target **task**
- the **domain expertise** of the info seeker
- the **amount of time and effort** available to invest in the process



- This interaction – also known as ***information seeking*** – can be seen as being part of a larger process referred to as ***sensemaking***
 - **sensemaking**: iterative process of formulating a conceptual representation
 - examples: legal discovery process, epidemiology (disease tracking), studying customer complaints to improve service

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

– **query support**

– results display

– evaluation

Search interfaces

- *Classic* methods of information retrieval:
 - entering query into a *search entry form* ⇐
 - generally small queries: “*tests the waters*”
 - if results not relevant: query reformulation (orienteering)
 - selecting links from a directory (website, bookmarks)
- *Contemporary* methods of information retrieval
 - natural language: text or speech form
 - possibility to upload documents and further specify goals (doc retrieval, access facts annotation...)
- **Search entry forms** can be classified according to the supported syntax
 - Boolean operators and auxiliary operators
 - language query in written or audio format (e.g. Alexia)

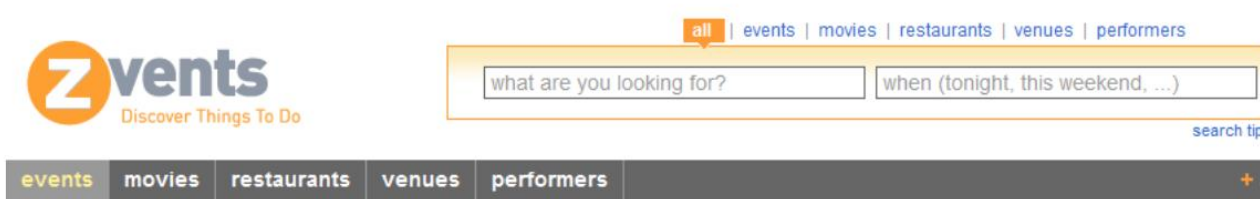
Search interfaces

- Can provide multiple forms (for **specialized inputs**)
 - offering hints on the kind of info (e.g., author, minimum #cites, reference document) to enter per form
- Typically support **auto-complete** and **auto-suggest** facilities (dynamic **query suggestions** and **reformulation**)
 - suggestions shown are those whose prefix matches the characters typed so far
 - suggestions may show **synonyms** of typed words
 - **spelling corrections**
 - [Anick *et al.*] users click on dynamic Yahoo suggestions more than one third of times
- **How to** implement query suggestions/reformulations?
 - available metadata and dictionaries (background knowledge)
 - user's own query history
 - querying behavior of similar users
 - querying behavior of all users

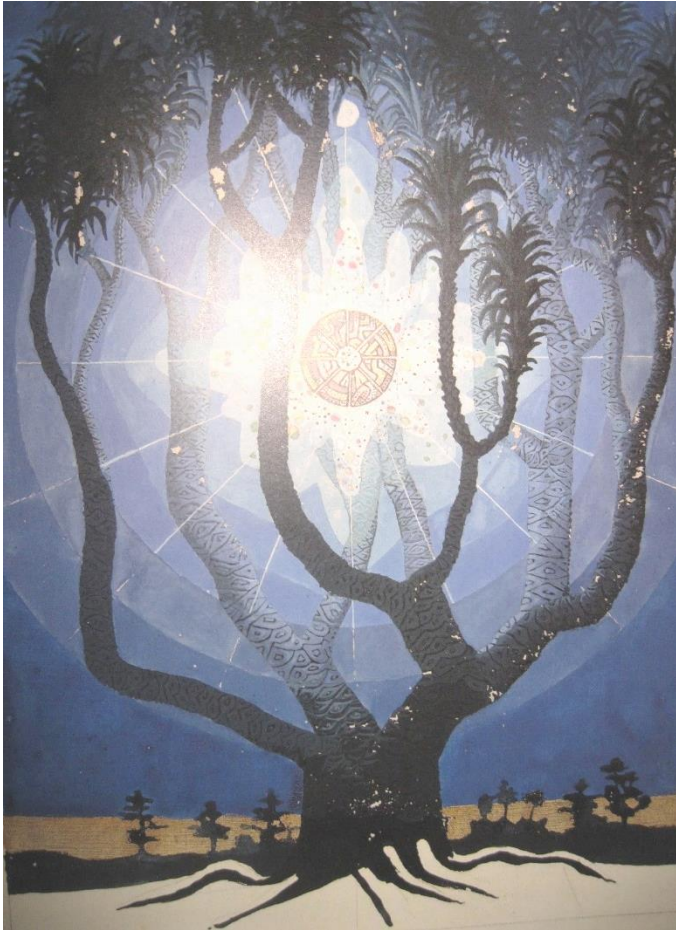
Search interfaces

Examples:

- google
- scholar
- yelp
- zevents
- netflix.com
- nextbio



Outline



IR definition and origins

Data structuring

IR system

Dynamic search model

Search interfaces

– query support

– **results display**

– evaluation

Visualizing results

Visualization differs depending on the task

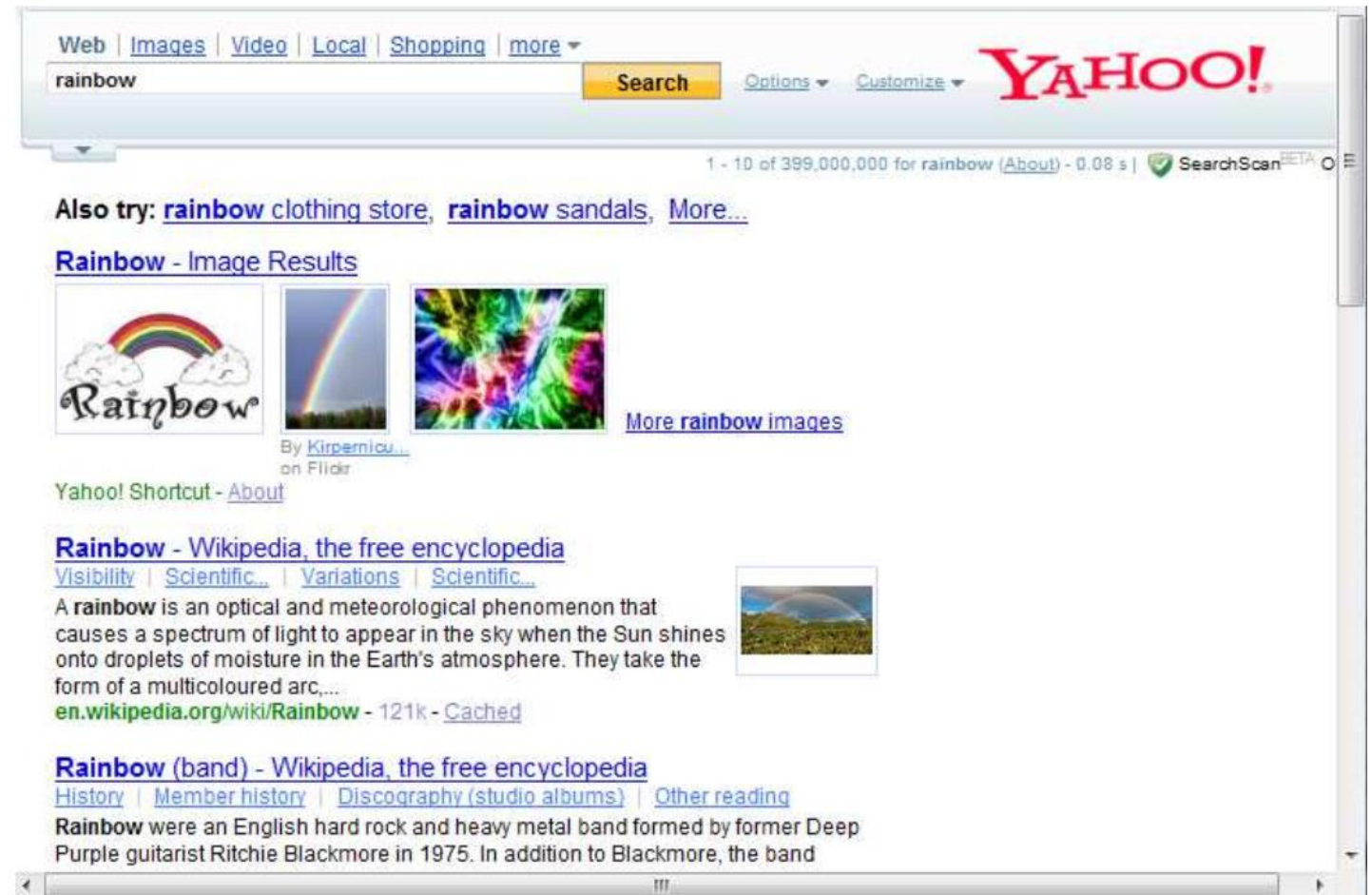
- **fact retrieval**: relevant pieces (single fact *versus* multiple facts with context and/or likelihood)
- **question answering**: simple plain text versus rich formatted text (*css, rft, markdown*)
- **document retrieval**: let us revisit *google* as a reference IR...

How to display an (ordered) set of retrieved documents?

- *adequate visualization* **per document**
 - page title, author, URL, date published, summary (snippet)
 - document **surrogate** refers to the information that summarizes the document
 - the quality of the surrogate greatly affects the perceived relevance of the search engine
 - query terms highlighted in the context in which they appear in the document
 - referred as **term highlighting** or keywords in context (**KWIC**)
 - improves user's ability to gauge document relevance
- *organizing the retrieved documents*: flat/list, group, hierarchy, faceted

Retrieval results display

- **blended surrogate** can combine:
 - text summaries
 - metadata (annotations)
 - media
- **text summaries**
 - can either follow **extractive** (copy of central sentences) or **generative** approaches (synthesized text)
 - either **query-independent** or **query-dependent** (also termed query-biased, query-oriented, or user-directed summaries)



Retrieval results display

- other *blended surrogates*:
 - figures from journal articles alongside the search results
 - speech assistance
 - audio and video snippets (e.g. pre-loading on hover)
 - going beyond doc retrieval:
 - multimodal answers, e.g. (generated) images and captioning

BioText SEARCH ENGINE [Home](#) | [About BioText](#) | [Contact Us](#)

Search:

Search Over: ☒ Full Text & Abstracts ☐ Figure Captions (List) ☐ Figure Captions (Grid) ☐ Tables Sort By: Results/Page:

Results 1-20 of 168 searching full text < 1 2 3 4 >

☒ ABSTRACTS ☒ FULL-TEXT EXCERPTS ☒ FIGURES

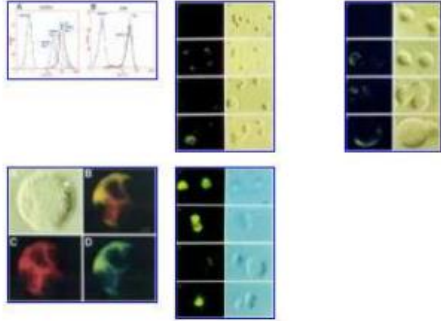
Down-regulation of cell surface CXCR4 by HIV-1
Choi, B., Gatti, P., Fermin, C., Vigh, S., Haislip, A., Garry, R. (2008) *Virology Journal*.

ABSTRACT
CXC 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](#)

FULL-TEXT EXCERPTS
...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...
...manner [29,30]. Chemokine receptors, including CCR5 and CXCR4, can be... [Show Full Excerpts](#)

VIEW FULL ARTICLE: [HTML](#) | [PDF](#)

FIGURES FROM ARTICLE:

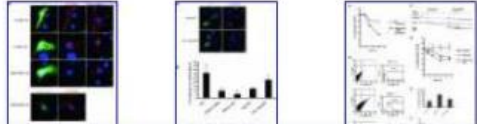


[View all figures \(5\) and tables from this article.](#)

Differential control of CXCR4 and CD4 downregulation by HIV-1 Gag
Valiathan, R., Resh, M. (2008) *Virology Journal*.

ABSTRACT
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

FIGURES FROM ARTICLE:



Retrieval results display

Organizing the retrieved documents:

- **category system:** meaningful labels reflecting the concepts relevant to a domain
 - good categories: coherent, complete per document, predictable across different searches
 - **type of categories**
 - **flat:** simple list of topics/subjects
 - **hierarchical:** in early web days, hierarchical systems such as Yahoo's were popular
 - **faceted** (tagging): assignment of multiple categories to a single item
 - can be used to group, filter (narrow) and sort documents
 - each category corresponds to a different facet
- **clustering system:** unsupervised grouping of documents

Retrieval results display

The screenshot displays the SuperBook IR system interface, which is organized into several panes and windows. The main window shows a hierarchical table of contents on the left, a search results list in the bottom left, and a pop-up graphic of the United States in the center. The table of contents is a dynamic "fisheye" view that automatically generates a dynamic "fisheye view" which helps preserve user's orientation. The search results list shows a query for "three dimension" and a list of found items. The pop-up graphic shows a map of the United States with the word "HIGH" repeated multiple times, indicating a search result. The interface also includes a "Dynamic 'Fisheye' Table of Contents - Automatically generates a dynamic 'fisheye view' which helps preserve user's orientation." section, a "Context-Guided Search - Automatically posts query 'hits' next to the topic headings in the Table of Contents - quickly directing searches." section, a "Rich Indexing - Automatically indexes every occurrence of every word in documents." section, a "Multimedia - Links to animations, video and other media and applications." section, a "Pop-Up Graphics" section, a "Tailored Text Displays - Dynamically formats and highlights text in response to user's search terms." section, a "Thumbnail Inline Graphics" section, an "Annotation- Add keywords or notes which are instantly indexed." section, and a "HyperText Functions - Shows graphics with a click; jumps to occurrences of search terms; links within and across documents" section.

Dynamic "Fisheye" Table of Contents - Automatically generates a dynamic "fisheye view" which helps preserve user's orientation.

Context-Guided Search - Automatically posts query "hits" next to the topic headings in the Table of Contents - quickly directing searches.

Rich Indexing - Automatically indexes every occurrence of every word in documents.

Multimedia - Links to animations, video and other media and applications.

Pop-Up Graphics

Tailored Text Displays - Dynamically formats and highlights text in response to user's search terms.

Thumbnail Inline Graphics

Annotation- Add keywords or notes which are instantly indexed.

HyperText Functions - Shows graphics with a click; jumps to occurrences of search terms; links within and across documents

SuperBook IR system: **hierarchical** organization and blended surrogates (after document *zoom-in*)

Retrieval results display

- *faceted navigation*
How to implement?
 - extracting **topics**
(topic modeling)
 - discovering **content patterns**
(concept analysis)

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

Powered by Flamenco

Save Search History and Settings Return to Search New Search Logout

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 (68)	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 (197)
Building (56)	Part of Building (44)
Built Open Space (14)	Road (21)

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

Birds (19)	Mammals, Hoofed (43)
Creatures and Beasts (1)	Mammals, Other (39)
Fish and Molluscs (6)	Parts of Plants (4)
Flowers (5)	Trees (33)

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)

 Caernavon 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 Moyle 19th - 20th century	 landscape with a ... 19th - 20th century
--	---	--	--

Retrieval results display

clustering: grouping documents according to some measure of similarity

- fully automatic yet not always predictable
- grouping can be led by specific doc properties (term similarity, temporal and style proximity)

The screenshot shows the Clusty search engine interface. At the top, there's a navigation bar with links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. A search bar contains the query 'senate' and a 'Search' button. Below the search bar, there are tabs for 'clusters', 'sources', and 'sites'. The 'clusters' tab is active, showing a list of clusters. The first cluster, 'Senate Committee', is highlighted in red and contains 29 documents. To the right of the clusters, a detailed view of the 'Senate Committee' cluster is shown, listing 6 results. Each result includes a title, a brief description, and a link to the source.

web news images wikipedia blogs jobs more »

senate Search

Clusty

clusters sources sites remix

All Results (199)

- Biography, Constituent services (57)
- Photos (34)
- Issues, news (8)
- Visiting Washington (6)
- Voting record (6)
- Virginia (4)
- Maine (3)
- Biography, Contact Details, And Constituent Services (2)
- Policy, Calendar (2)
- Other Topics (6)

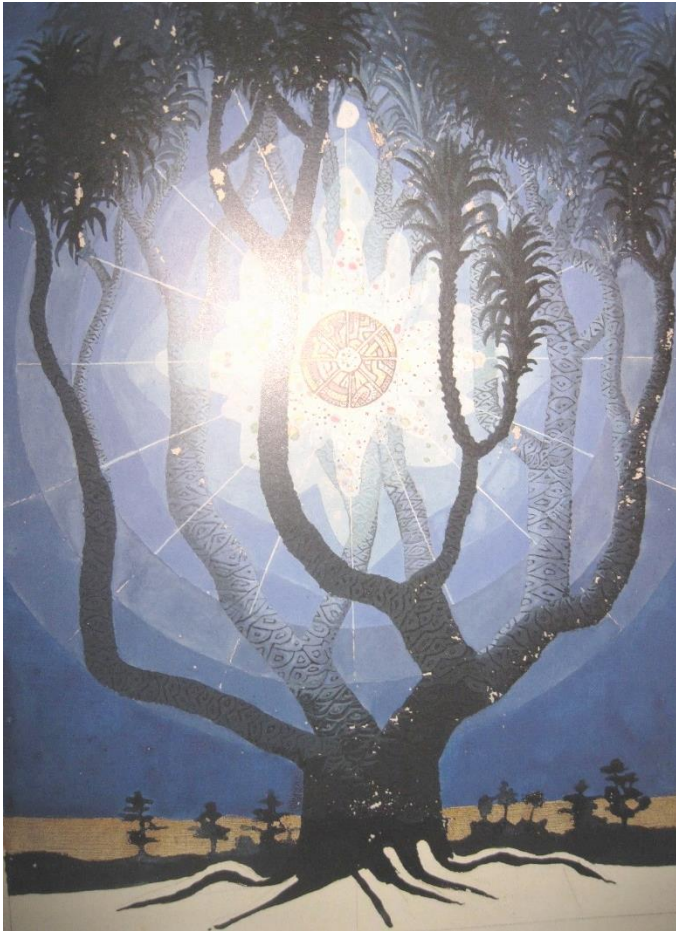
Senate Committee (29)

- State Senate (17)
- Votes (15)
- Constituent services (5)
- Obama Budget (2)
- Expand (2)

Cluster **Senate Committee** contains 29 documents.

- U.S. Senate**
Official site of "the living symbol of our union of states." Connect with S **committees**, legislation, records, art, history, schedules, news, tours
www.senate.gov - [cache] - Live, Open Directory, Ask
- U.S. Senate Committee on Commerce, Science, & Trans**
Committee jurisdiction includes the Coast Guard, coastal managemen **waterways**, interstate commerce, maritime commerce, fisheries, mer **commerce.senate.gov** - [cache] - Live, Ask
- United States Senate Committee on Banking, Housing an**
United States **Senate Committee** on Banking, Housing and Urban Affi **banking.senate.gov** - [cache] - Live
- Senate of the Kingdom of Cambodia**
Information about legislative activities, laws, **committees**, **senators** at **www.senate.gov.kh** - [cache] - Open Directory, Ask
- Kansas Senate**
Senate Roster, ... Home > **Senate ... Senate Committees**
www.kslegislature.org/legsrn-senate/index.do - [cache] - Ask
- U.S. Senate Committee on Energy and Natural Resource**
Has jurisdiction over energy policy, regulation, and research. Also dea **conservation**, ports used for energy transport, irrigation, reclamation, r **energy.senate.gov** - [cache] - Live

Outline



IR definition and origins

Data structuring

IR system

Dynamic search model

Search interfaces

– query support

– results display

– **evaluation**

Evaluating search interfaces

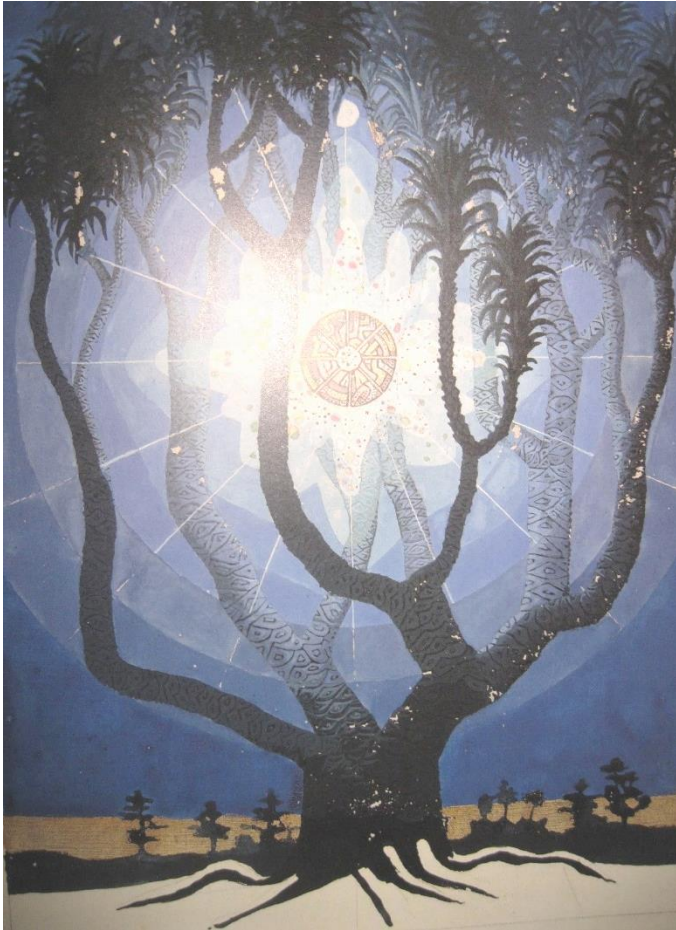
- Evaluation of IR systems entails different aspects:
 - **efficiency** and **efficacy** of their different components
 - the IR brain (**IR processing module**) to ensure proper answers in useful time
 - the **IR indexing module** to ensure updated and well-organized collections
 - what about **search interfaces**? Efficient and proper query support and result display also essential!
- User interface design: a field of **human-computer interaction** (HCI)
 - **user-centered design**: how people think about, respond to, and use technology
- Evaluation of user interface radically different from evaluating IR algorithms
 - a ranking algorithm can be evaluated by precision, recall and efficiency
 - user interface evaluation: subjective responses are as, if not more, important than quantitative measures
 - criteria: **speed, familiarity, aesthetics, preferred features, *perceived* ranking accuracy**
 - *“if a person has a choice between two systems, they will use the one they prefer”* (often the familiar one)

Evaluating user interfaces

How best to evaluate a user interface depends on the **IR system maturity**/stage:

- when starting with design and idea
 - **discount** usability methods: showing a few users different designs and asking pros and cons
 - **heuristic evaluation**: usability experts “walk through” a design and evaluate the functionality
 - *difficulties*: difficulty to mimic long-standing interactive search sessions
- advanced development stage
 - **longitudinal studies**
 - participants can test a new interface for an extended period of time
 - evaluation is based both on log analysis and questionnaires
- well-established and heavily-used user interfaces
 - **bucket testing (A/B testing)**
 - a randomly selected subset of the users is shown a new design their actions are logged and compared to another randomly control group that continues to use the existing interface

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

- **query support**
- **results display**
- **evaluation**

Thank You



rmch@tecnico.ulisboa.pt