Retrieval Interface

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Introduction

- Information Retrieval interaction: "… the interactive communication processes that occur during the retrieval of information by involving all the major participants in IR, i.e. the user, the intermediary, and the IR system." (Ingwersen, 1992)
- → What variables are involved?
  - models give lists
- → How do they affect the process? How to control?
  - experiments, experience, observation give answers
- → Do given interventions or communications improve or degrade the process?
  - e.g. searcher's (intermediaries or end-users) actions
- → Can systems be designed so that searcher's intervention improves performance?

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Interaction Points

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Covered topics

- Information needs
- IR basic models
- Interaction points
  - Query formulation
  - Selection
  - Examination
  - Source selection
- Appliance: Retrieval System and Mobile Search

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

1. Information Needs

Stefano Mizzaro. (1999) How Many Relevances in Information Retrieval? *Interacting With Computers*, 10(3), 305-322.

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

2. Interactive IR models

- Several models proposed
  - none as widely accepted as the traditional IR model
- They all try to incorporate
  - information objects ("texts"):
  - IR system & setting
  - interface
  - intermediary, if present
  - user's characteristics
    - cognitive aspects; task; problem; interest; goal; preferences …
  - social environment
  - variety of processes between them all.

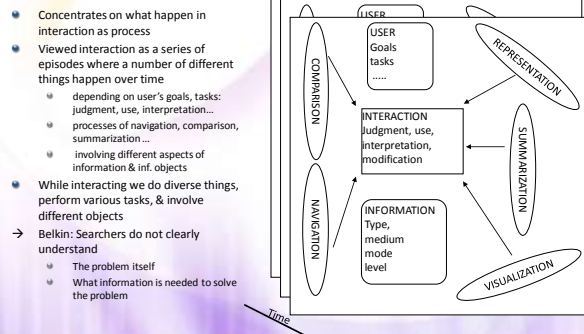Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Ingwersen's cognitive model

- Among the first to view IR differently from traditional model
- Included IR as system but also things outside system that interact
  - path of request from environment (problem) to query
  - path of cognitive changes
  - path of communication
  - various other paths of interactions

Information objects (doc, imgs)

Interface/ Intermediary (user)

User's cognitive space

Environment (use & general)

Query

Request

IR system setting

Cognitive transformations

Interactive communication

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Belkin's episodes model

- Concentrates on what happen in interaction as process
- Viewed interaction as a series of episodes where a number of different things happen over time
  - depending on user's goals, tasks: judgment, use, interpretation...
  - processes of navigation, comparison, summarization ...
  - involving different aspects of information & inf. objects
- While interacting we do diverse things, perform various tasks, & involve different objects
- Belkin: Searchers do not clearly understand
  - The problem itself
  - What information is needed to solve the problem

USER
USER
USER Goals tasks ......
COMPARISON
REPRESENTATION
INTERACTION Judgment, use, interpretation, modification
SUMMARIZATION
NAVIGATION
INFORMATION Type, medium mode level
VISUALIZATION
Time

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Stratified model

Users interact with IR systems to use information

INTERACTION LEVELS/STRATA

Situational tasks...
Affective intent..
Cognitive knowledge structure ...
Query characteristics
Surface level
INTERFACE
Engineering hardware, capacities...
Processing software, algorithms...
Content information resources - texts representation ...

USER
Adaptation
'COMPUTER'
Adaptation
Use of information
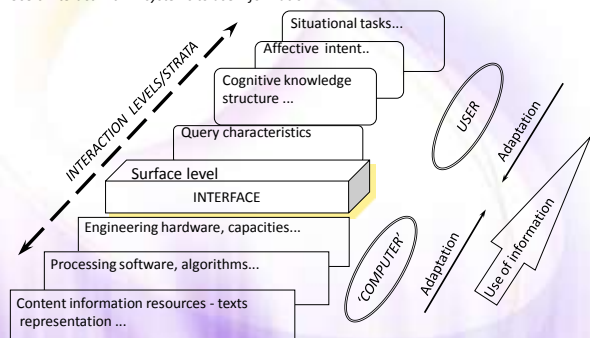
Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Broader View of Information Retrieval

The query results from a clarification process

User Modeling

Query Words

Ranked List

Domain Modeling
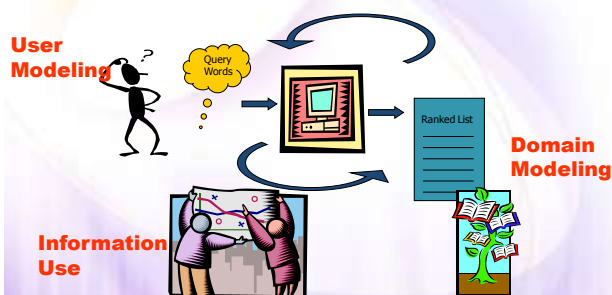
Information Use

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Bates' "Berry Picking" Model

A sketch of a searcher... "moving through many actions towards a general goal of satisfactory completion of research related to an information need."

Q2
Q4
Q1
Q3
Q5
Q0
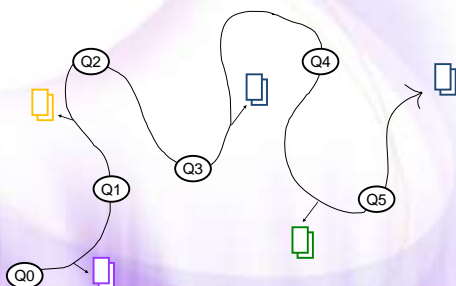
Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Broder's Web Query Taxonomy

- Navigational (~20%)
  - Reach a particular site ("known item")
- Informational (~50%)
  - Acquire static information ("topical")
- Transactional (~30%)
  - Perform a Web-mediated activity ("service")

Andrei Broder, *SIGIR Forum*, Fall 2002

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Some Desirable Features

- Make exploration easy

- Relate documents with why they are retrieved

- Highlight relationships between documents

Vu Thi Huong Giang, PhD   Human Interface   2009-2010
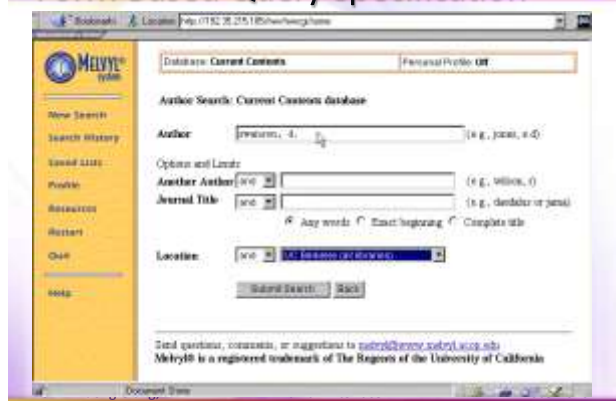
2. Query Formulation

- Command Language
- Form Fill-in
- Menu Selection
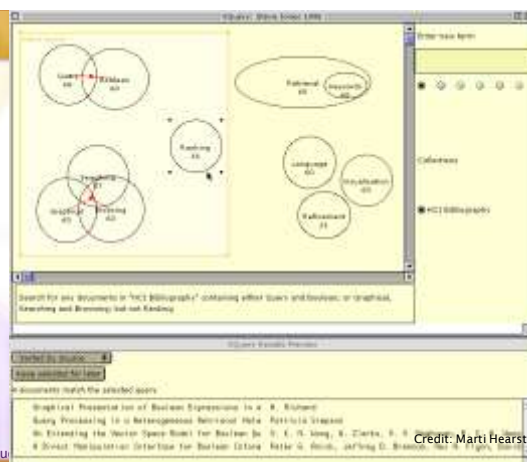- Direct Manipulation
- Natural Language

Ben Shneiderman, 1997
Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Form-Based Query Specification

Direct Manipulation Spec. VQUERY (Jones 98)

Credit: Marti Hearst

Vu Thi Hu

The "Back" Button

- Behavior is counterintuitive to many users

A

B

C

D
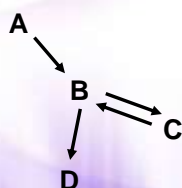
You hit "back" twice from page D.
Where do you end up?

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

PadPrints

- Tree-based history of recently visited Web pages
  - History map placed to left of browser window
  - Node = title + thumbnail
  - Visually shows navigation history
- Zoomable: ability to grow and shrink sub-trees

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Visual Browsing History in PadPrints
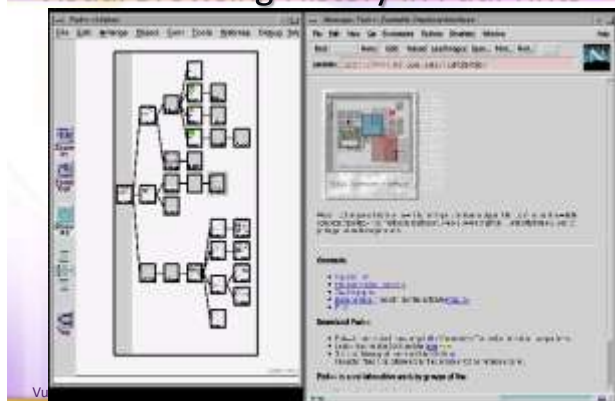
PadPrints Thumbnails

Vu Thi Huong Giang, Ph 009-2010

Alternate Query Modalities

- Spoken queries
  - Used for telephone and hands-free applications
  - Reasonable performance with limited vocabularies
    - But some error correction method must be included

- Handwritten queries
  - Palm pilot graffiti, touch-screens, …
  - Fairly effective if some form of shorthand is used
    - Ordinary handwriting often has too much ambiguity

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

3. Selection

- A Selection Interface Taxonomy
  - One dimensional lists
    - Content: title, source, date, summary, ratings, …
    - Order: retrieval status value, date, alphabetic, …
    - Size: scrolling, specified number, score threshold
  - Two dimensional displays
    - Construction: clustering, starfield, projection
    - Navigation: jump, pan, zoom
  - Three dimensional displays
    - Contour maps, fishtank VR, immersive VR

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Google: KeyWord In Context (KWIC)

- Query: truong dai hoc bach khoa ha noi

Vu Thi Huon   9-2010

Indicative vs. Informative

- Terms often applied to document abstracts
  - Indicative abstracts support selection
    - They describe the contents of a document
  - Informative abstracts support understanding
    - They summarize the contents of a document

- Applies to any information presentation
  - Presented for indicative or informative purposes

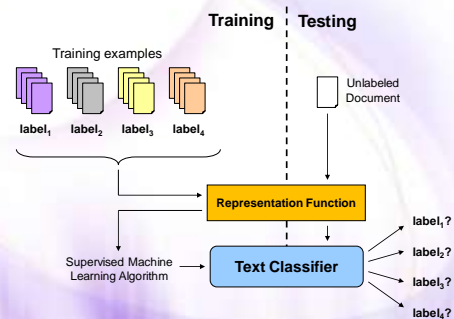Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Selection/Examination Tasks

- "Indicative" tasks
  - Recognizing what you are looking for
  - Determining that no answer exists in a source
  - Probing to refine mental models of system operation

- "Informative" tasks
  - Vocabulary acquisition
  - Concept learning
  - Information use

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Generated Summaries

- Fluent summaries for a specific domain
- Define a knowledge structure for the domain
  - Frames are commonly used
- Analysis: process documents to fill the structure
  - Studied separately as "information extraction"
- Compression: select which facts to retain
- Generation: create fluent summaries
  - Templates for initial candidates
  - Use language model to select an alternative

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Extraction-Based Summarization

- Robust technique for making disfluent summaries
- Four broad types:
  - Query-biased vs. generic
  - Term-oriented vs. sentence-oriented
- Combine evidence for selection:
  - Salience: similarity to the query
  - Specificity: IDF or chi-squared
  - Emphasis: title, first sentence

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Ask: Suggested Query Refinements

Selection principles

- Classification: Automatically assign labels to documents
  - Machine learning
  - K nearest neighbor
  - Cat-a-cone
- Clustering: Automatically group documents into clusters
  - Hierarchical Agglomerative Clustering
  - K-means clustering
  - Scatter/gather

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Text Classification

- Problem: automatically sort items into bins
- Machine learning approach
  - Obtain a training set with ground truth labels
  - Use a machine learning algorithm to "train" a classifier
    - kNN, Bayesian classifier, SVMs, decision trees, etc.
  - Apply classifier to new documents
    - System assigns labels according to patterns learned in the training set

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Open Directory Project

http://www.dmoz.org/; keyword: solar

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Machine Learning

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

k Nearest Neighbor (kNN) Classifier

- Algorithm:
  - Select k most similar labeled documents
  - Have them "vote" on the best label:
    - Each document gets one vote, or
    - More similar documents get a larger vote
  - How can similarity be defined?
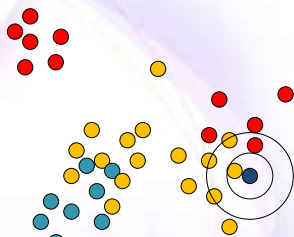
Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Cat-a-Cone

Vu Thi ...

Cat-a-Cone

- Key Ideas:
  - Separate documents from category labels
  - Show both simultaneously
  - Link the two for iterative feedback
  - Integrate searching and browsing
- Distinguish between:
  - Searching for documents
  - Searching for categories
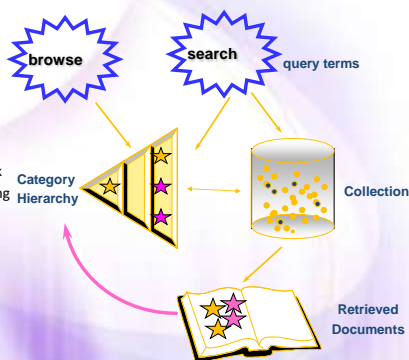
Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Clustering

- The Cluster Hypothesis: "Closely associated documents tend to be relevant to the same requests." (van Rijsbergen 1979)
- Advantages:
  - Provides an overview of main themes in search results
  - Helps overcome polysemy
- Disadvantages:
  - Documents can be clustered in many ways
  - Not always easy to understand the theme of a cluster
  - What is the correct level of granularity?
  - More information to present

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Vivisimo: Clustered Results

http://clusty.com/

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Clustering Result Sets

- Advantages:
  - Topically coherent document sets are presented together
  - User gets a sense for the themes in the result set
  - Supports browsing retrieved hits
- Disadvantages:
  - May be difficult to understand the theme of a cluster based on summary terms
  - Clusters themselves might not "make sense"
  - Computational cost

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Visualizing Clusters

△ Centroids

Vu Thi Huong Giang, PhD   Human Interface   2009-2010
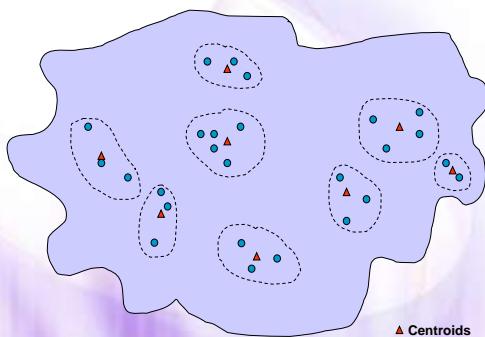
Hierarchical Agglomerative Clustering

- Algorithm:
  - Start with each document in its own cluster
  - Until there is only one cluster:
  - Determine the two most similar clusters c_i and c_j
  - Replace c_i and c_j with a single cluster $c_i \cup c_j$
  - The history of merging forms the hierarchy

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Cluster Similarity

- Assume a similarity function that determines the similarity of two instances: sim(x,y)
  - What's appropriate for documents?

- What's the similarity between two clusters?
  - Single Link: similarity of two most similar members
  - Complete Link: similarity of two least similar members
  - Group Average: average similarity between members

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

K-Means Clustering

Pick seeds
Reassign clusters
Compute centroids
Reasssign clusters
Compute centroids
Reassign clusters
Converged!

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

7

K-Means

- Each cluster is characterized by its centroid (center of gravity):

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Reassignment of documents to clusters is based on distance to the current cluster centroids

- Let d be the distance measure between documents
- Select k random instances $\{s_1, s_2, \dots s_k\}$ as seeds
- Until clustering converges:
  - Assign each instance x_i to the cluster c_j such that $d(x_i, s_j)$ is minimal
  - Update the seeds to the centroid of each cluster
  - For each cluster c_j, $s_j = \mu(c_j)$

How do you select k?
Results can vary based on random seed selection
Some seeds can result in poor convergence rate, or convergence to sub-optimal clusters

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Scatter/Gather

- Clusters documents into "themes"
  - Displays clusters by showing:
    - Topical terms
    - Typical titles
- User chooses a subset of the clusters
- System re-clusters documents in selected cluster
  - New clusters have different, more refined, "themes"

Query = "star" on encyclopedic text

		sports	14 docs
symbols	8 docs	film, tv	47 docs
film, tv	68 docs	music	7 docs
astrophysics	97 docs		
astronomy	67 docs		
flora/fauna	10 docs		

	stellar phenomena	12 docs
	galaxies, stars	49 docs
	constellations	29 docs
	miscellaneous	7 docs

Clustering and re-clustering is entirely automated

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Vu Thi Huong Giang, PhD   Human Interface   2009-2010
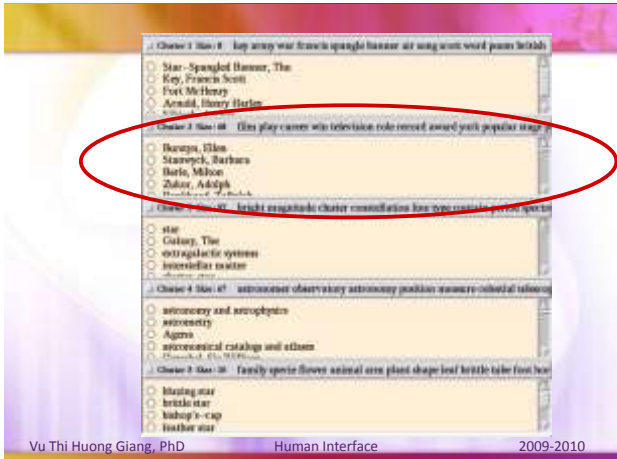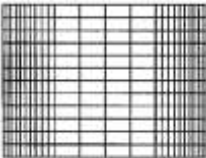
4. Examining Individual Documents

Document lens

Vu Thi Huong Giang, PhD   Human Interface   Robertson & Mackinlay, UIST'93, Atlanta, 1993.   2009-2010
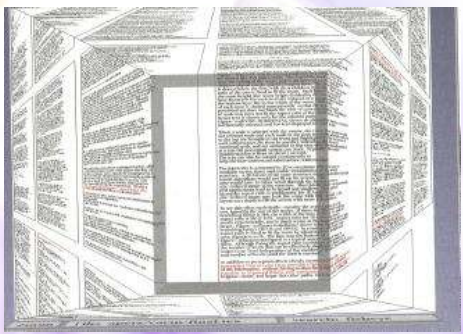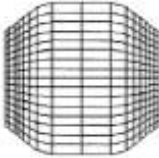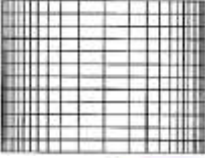
Distorting Reality
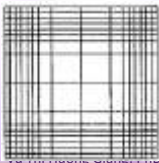
Bifocal **Perspective Wall** **Fisheye**

Human Interface

1-D Fisheye Menu

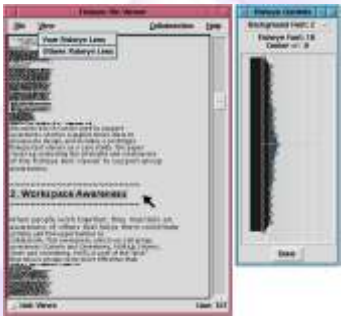http://www.cs.umd.edu/hcil/fisheyemenu/fisheyemenu-demo.shtml
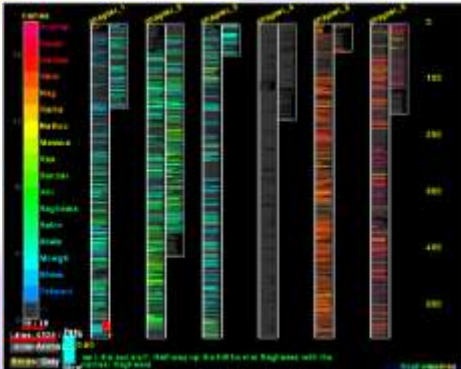
Vu Thi Huong Giang, PhD   Human Interface

1-D Fisheye Document Viewer

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

SeeSoft

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

TileBars

Topic: reliability of DBMS (database systems)
Query terms: DBMS, reliability

DBMS
reliability
Mainly about both DBMS and reliability

DBMS
reliability
Mainly about DBMS, discusses reliability

DBMS
reliability
Mainly about, say, banking, with a subtopic discussion on DBMS/Reliability

DBMS
reliability
Mainly about high-tech layoffs

Vu Thi Huong Giang, PhD   Human Interface   2009-2010
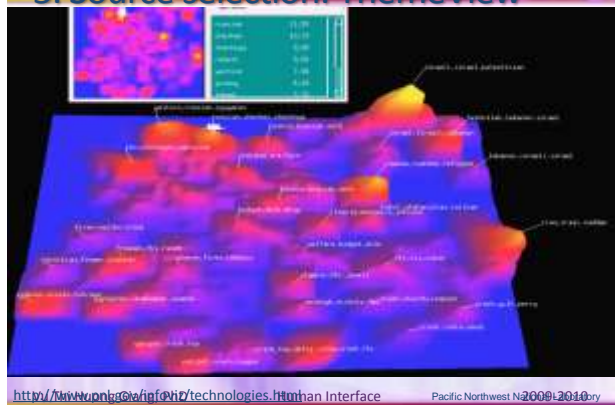
U Mass: Scrollbar-Tilebar
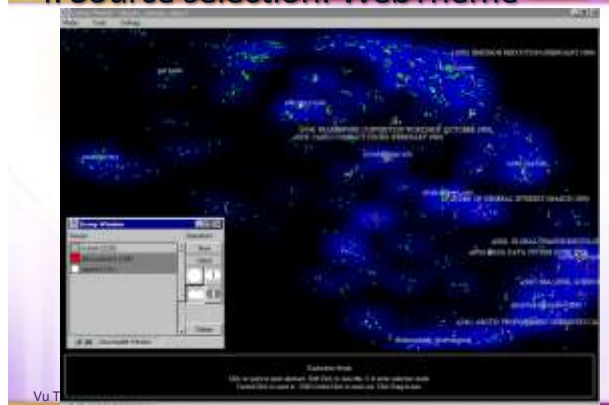
5. Source selection: ThemeView

http://www.pnl.gov/infoviz/technologies.html   Vu Thi Huong Giang, PhD   Human Interface   Pacific Northwest National Laboratory   2009-2010

4. Source selection: WebTheme

Vu T...

Ben S' 'Seamless Interface' Principles

- Informative feedback
- Easy reversal
- User in control
  - Anticipatable outcomes
  - Explainable results
  - Browsable content
- Limited working memory load
  - Query context
  - Path suspension
- Alternatives for novices and experts
  - Scaffolding

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Some 'Synergistic Interaction' Principles

- Interdependence with process ("interaction models")
  - Co-design with search strategy
  - Speed
- System initiative
  - Guided process
  - Exposing the structure of knowledge
- Support for reasoning
  - Representation of uncertainty
  - Meaningful dimensions
- Synergy with features used for search
  - Weakness of similarity, Strength of language
- Easily learned
  - Familiar metaphors (timelines, ranked lists, maps)

Vu Thi Huong Giang, PhD   Human Interface   2009-2010

Some Good Ideas

- Show the query in the selection interface
  - It provides context for the display
- Suggest options to the user
  - Query refinements, for example
- Explain what the system has done
  - Highlight query terms in the results, for example
- Complement what the system has done
  - Users add value by doing things the system can't
  - Expose the information users need to judge utility

Vu Thi Huong Giang, PhD   Human Interface   2009-2010