The IR Black Box

Search
Query
Ranked List

The Information Retrieval Cycle

Source
Selection
Resource

Query
Formulation

Search
Documents

Selection
Documents

Ranked List
Documents

Examination
Documents

Delivery

Overview

Query
Selection
Documents

Today

Documents

Documents

Last Week

Documents

Overview

Information Seeking Behavior

- Two parts of a process:
  - Search and retrieval
  - Analysis and synthesis of search results
- Standard (naïve?) assumptions:
  - Maximizing precision and recall simultaneously
  - The information need remains static
  - The value is in the resulting document set

Problematic Assumptions

- Users learn during the search process:
  - Scanning titles of retrieved documents
  - Reading retrieved documents
  - Browsing lists of related topics/thesaurus terms
  - Navigating hyperlinks
- The end goal isn’t necessary the document set
  - Recall Belkin’s Anomalous States of Knowledge
- Lists of documents may not be the best presentation method
An Alternative Model

- Standard IR model
  - Assumes the information need remains the same throughout the search process
- Berry-picking model (Marcia Bates)
  - Interesting information is scattered like berries among bushes
  - The query is continually shifting

Recall Mizzaro’s Model

- Real information needs (RIN) = visceral need
- Perceived information needs (PIN) = conscious need
- Request = formalized need
- Query = compromised need


Implications

- Interfaces should make exploration of the information space easy
- Interfaces should relate information objects with why they are retrieved
- Interfaces should highlight the relationships between multiple information objects
- Makes evaluation more difficult

Moore’s Law

- Transistors, speed, storage...

Human Cognition

- Computer performance vs. human performance

Moore’s Law Chart

- 1950, 1990, 2030

Human Cognition Chart

- 1990
Where is the bottleneck?

- system vs. human performance

Interface Support
- Starting points
  - How do I know where to start?
  - How do I know what to search?
- Query formulation
  - How do I specify a query?
- Results selection
  - How do I make sense of the results?
  - How do I grasp the content of retrieved items?
- Navigation
  - How do I move around in the information space?

The Role of Interfaces
- Source selection
- Search
- Query formulation
- Ranked list
- Examination
- Delivery

Ben’s Principles
- Informative feedback
- Easy reversal
- User in control
  - Anticipatable outcomes
  - Explainable results
  - Browseable content
- Limited working memory load
- Alternatives for novices and experts

Automated Source Selection
- A few ideas:
  - Compare the query against summaries of what is contained in the collection
  - Predict most likely source based on instances of the query terms in each collection
  - Use implicit feedback
- But today’s focus is on interfaces…

The Simplest “Starting Point”
- List the options:
Starting Points

- A tale of three sites:
  - Google http://www.google.com/
  - dmoz open directory project http://www.dmoz.org/
  - University of Maryland Libraries http://www.lib.umd.edu/

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

Cat-a-Cone Interface

Cat-a-Cone Architecture

ConeTree for Category Labels

- One can explore the category hierarchy
  - By searching on label names
  - By browsing the labels
  - By growing/shrinking subtrees
  - By spinning subtrees
- Affordances
  - Learn meaning via ancestors, siblings
  - Disambiguate meanings
  - View many categories simultaneously

Virtual Book for Result Sets

- Categories in retrieved documents linked to categories in tree
- Flipping through “book pages” causes some subtrees to expand and contract
  - Most subtrees remain unchanged
- Books can be stored for later reuse
Browsing Hierarchies

- How do you browse large hierarchies?

- MeSH Category Hierarchy = Medical Subject Headings
  - ~18,000 labels
  - ~8 labels/article on average, manually assigned
  - Top level categories:
    - anatomy
    - diagnosis
    - related disc
    - animals
    - psych
    - technology
    - disease
    - biology
    - humanities
    - drugs
    - physics

MeshBrowse

TaxonTree

HiBrowse

HiBrowse

HiBrowse
Query Formulation

- The role of the interface in query formulation
  - Can the interface help the user better express his or her query?
  - Does the user actually mean the query that’s entered?

Interaction Styles

- Command language
- Form-based interfaces
- Menu-based interfaces
- Direct Manipulation
- Natural Language

WESTLAW® Query Examples

- What is the statute of limitations in cases involving the federal tort claims act?
  - LIMIT! /3 STATUTE ACTION /5 FEDERAL /2 TORT /3 CLAIM

- What factors are important in determining what constitutes a vessel for purposes of determining liability of a vessel owner for injuries to a seaman under the “Jones Act” (46 USC 688)?
  - (46 +3 688) “JONES ACT” /P INJUR /S SEAMAN CREWMAN WORKER

- Are there any cases which discuss negligent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?
  - NOT NEGLECT! FAIL! MAINT! REPAIR! NAVIGAT! AID EQUIP! LIGHT BUOY “CHANNEL MARKER”

- What cases have discussed the concept of excusable delay in the application of statutes of limitations or the doctrine of laches involving actions in admiralty or under the “Jones Act” or the “Death on the High Seas Act”?
  - EXCUS! /3 DELAY! (LIMIT! /3 STATUTE ACTION) LACHES! “JONES ACT” “DEATH ON THE HIGH SEAS ACT” (46 +3 781)

Google

- Search box as an example of a command language
- Advanced search as an example of a form-based interface
- What is the relationship between the two?

Menu-Based Interfaces

Direct Manipulation

Presenting Results

- I performed a search and got back a whole bunch of results… now what?
- Simplest interface: one-dimensional lists:
  - Content: title, source, date, summary, ratings, …
  - Order: retrieval status value, date, alphabetic, …
  - Size: scrolling, specified number, score threshold
Putting Results in Context

- Interfaces should
  - Give hints about the roles terms play within the result set and within the collection
  - Give hints about the relationship between different terms
  - Show explicitly why documents are retrieved in response to the query
  - Compact summary the result set
- This can be viewed as a problem in information visualization

Information Visualization

- What is it?
  - Information visualization for document retrieval:
    - Enhance a person’s ability to read, understand, and gain knowledge from a document or documents
    - Understand the contents of a document or collection of documents without reading them
  - We’re focusing on search results

More Specific Tasks

- Which documents contain text on the relevant topic?
- Which documents are of interest to me?
- Are there other documents that might be close enough to be worthwhile?
- What are the main themes of a document?
- How are certain words or themes distributed through a document?

Fundamental Problem

- Scale – there’s simply too much data to fit on one screen
  - Too many instances
  - Too many variables
  - Too many points of user focus
- Basic concepts
  - Overview + detail
  - Focus + context

Overview + Detail

- Providing an overview of the information is extremely valuable
  - Helps present overall patterns
  - Assists user with navigation and search
  - Orients activities
- Users also want to examine details: individual instances and variables
- How do we get the best of both worlds?

Case 1: Detail-only
Case 2: Zoom and Replace

Case 3: Overview + Detail

Case 4: Multiple Overviews

Case 5: Tiled Browser

Case 6: Bifocal Magnification

Case 7: Fish-eye view
**Focus + Context**
- Same basic idea as overview and detail, with one key difference:
  - Typically, the overview and the detail are combined into a single display
  - Shows information in its “natural surrounding”
- General paradigm seems to work well for text…

**Simple Taxonomy**
- Single document
- Collection of documents
- Today
- Next Week
- Concepts and relationships (semantics)
- Enhanced presentation (syntax)

**Too Much Text!**
- Text is too small to read!

**Presenting Results**
- KWIC and extensions
- TileBars
- Next time: clustering
  - Automatically group together “related” documents

**The Oldest Trick**
- KWIC = Key Word In Context
- Displaying query terms in the context of the retrieved documents

**Issues to Consider**
- What part of the document should the system extract?
  - A contiguous segment of text?
  - Segments extracted from different sections?
- How much context should you show?
- How should you mark the keywords?
How Much Context?

- How much context for question answering?
- Possibilities
  - Exact answer
  - Answer highlighted in sentence
  - Answer highlighted in paragraph
  - Answer highlighted in document

Interface Conditions

- Who was the first person to reach the south pole?

User study

- Independent variable: amount of context presented
- Test subjects: 32 MIT undergraduate/graduate computer science students
  - No previous experience with QA systems
- Actual question answering system was canned:
  - Isolate interface issues, assuming 100% accuracy in answering factoid questions
  - Answers taken from WorldBook encyclopedia

Question Scenarios

- User information needs are not isolated...
  - When researching a topic, multiple, related questions are often posed
  - How does the amount of context affect user behavior?
- Two types of questions:
  - Singleton questions
  - Scenarios with multiple questions
- When was the Battle of Shiloh?
- What state was the Battle of Shiloh in?
- Who won the Battle of Shiloh?

Setup

- Materials:
  - 4 singleton questions
  - 2 scenarios with 3 questions
  - 1 scenarios with 4 questions
  - 1 scenarios with 5 questions
- Each question/scenario was paired with an interface condition
- Users asked to answer all questions as quickly as possible

Results: Completion Time

- Answering scenarios, users were fastest under the document interface condition
- Differences not statistically significant

**Results: Questions Posed**
- With scenarios, the more the context, the fewer the questions
  - Results were statistically significant
  - When presented with context, users read

![Questions Posed Chart]

**A Story of Goldilocks...**
- The entire document: too much!
- The exact answer: too little!
  - It occurred on July 4, 1776.
  - What does this pronoun refer to?
- The surrounding paragraph: just right...

**Lessons Learned**
- Keyword search culture is engrained
- Discourse processing is important
  - When was the Battle of Shiloh?
  - And where did it occur?
- Users most prefer a paragraph-sized response
- Context serves to
  - “Frame” and “situate” the answer within a larger textual environment
  - Provide answers to related information

**Other Ways to Show Context**
- Distortion-oriented techniques
- TileBars

**Everyday Life Example**

**Ways to Distort Reality**
- Bifocal Display
- Perspective Wall
More Ways...

**1D Fisheye**

**2D Fisheye**

---

A Real Fisheye

On I-285, another Perimeter mall

---

Perspective Wall

- 3D implementation of a bifocal display
- Presenting work charts:
  - x-axis is time
  - y-axis is project

---

Other Applications

---

Even More Applications
The Document Lens

Fisheye Menu

Fisheye Views of Documents

Multiple Focii

Does it always work?

TileBars

- All this zooming is making me dizzy
  - What about a more abstract representation of the document?
- Show a graphical representation of term distribution and overlap in search results
- Simultaneously Indicate:
  - Relative document length
  - Query term frequencies
  - Query term distributions
  - Query term overlap
**Technique**

Blocks indicate "chunks" of text, such as paragraphs. Blocks are darkened according to the frequency of the term in the document.

**Example**

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

<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBMS</td>
<td>Mainly about both DBMS and reliability</td>
</tr>
<tr>
<td>reliability</td>
<td>Mainly about DBMS, discusses reliability</td>
</tr>
<tr>
<td>DBMS</td>
<td>Mainly about, say, banking, with a subtopic discussion on DBMS/Reliability</td>
</tr>
<tr>
<td>DBMS</td>
<td>Mainly about high-tech layoffs</td>
</tr>
</tbody>
</table>

**TileBars Screenshot**

Compact, graphical representation of term distribution for full text retrieval results.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term frequency</td>
<td>Displayed</td>
</tr>
<tr>
<td>Distribution</td>
<td>Simultaneously display</td>
</tr>
<tr>
<td>Overlap</td>
<td></td>
</tr>
<tr>
<td>Doc length</td>
<td></td>
</tr>
</tbody>
</table>

However, does not provide the context in which query terms are used.

**TileBars Summary**

- Compact, graphical representation of term distribution for full text retrieval results.
- Simultaneously display term frequency, distribution, overlap, and doc length.
- However, does not provide the context in which query terms are used.
- Do they help?
  - Users intuitively understand them.
  - Lack of context sometimes causes problems in disambiguation.

**Scrollbar-Tilebar**

Compact, graphical representation of term distribution for full text retrieval results.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term frequency</td>
<td>Displayed</td>
</tr>
<tr>
<td>Distribution</td>
<td>Simultaneously display</td>
</tr>
<tr>
<td>Overlap</td>
<td></td>
</tr>
<tr>
<td>Doc length</td>
<td></td>
</tr>
</tbody>
</table>

However, does not provide the context in which query terms are used.

**Do they help?**

- Users intuitively understand them.
- Lack of context sometimes causes problems in disambiguation.

From U. Mass
Navigation

- The “berry-picking model”: interfaces should make it easy to follow trails with unanticipated results.
- Interfaces should support navigation of the information space.

The “Back” Button

- The “back” button isn’t enough!
- Behavior is counterintuitive to many users.

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.

PadPrints Screenshots

PadPrints Thumbnails

Zoomable History
### Does it work?
- Study involved CHI database and National Park Service website
- In tasks requiring return to prior pages, 40% savings in time when using PadPrints
- Users more satisfied with PadPrints

### The Importance of Interfaces
- The user interface is the part of the system that the user interacts with:
  - Interaction is an integral part of the information seeking process
  - Search experience is affected by the quality of the interface
- Interfaces should:
  - Help users get started
  - Help users keep track of what they have done
  - Help users make sense of what the system did
  - Suggest next choices
- It is very difficult to design good UIs
- It is very difficult to evaluate search UIs