

Searching the Long Tail

Elizabeth D. Liddy

Center for Natural Language Processing
School of Information Studies
Syracuse University

2006 I-School Conference

Ann Arbor, Michigan

The Pendulum Swing of Search

- Future of search reflects many of the same notions discussed in “The Long Tail”
 - Web-based search exhibits a demand curve much like an online store's, with huge appeal of the top items, tailing off quickly for less popular ones
 - Challenge is to devise search that:
 1. Continues to provide excellent access to popular items
 2. Gets folks into the long tail on the open web
 3. Provides better search techniques in specialized collections
- History shows repeated swings between ‘*generic*’ and ‘*specialized*’ search over succeeding eras of search engine R & D
 - ‘*Specialized*’ = ‘s searching “The Long Tail”
 - Each swing has built on research of 1 or 2 specific disciplines

The Pendulum Swing of Search

- First commercial systems used Controlled Vocabularies originally developed for manual searching within specific disciplines – LS
- ← Early days of IR research systems took 1-size-fits-all approach – CS
 - Salton's SMART system was intended to search all subject areas, taking advantage of universals of word frequency distributions
- Era of expert systems used knowledge engineering for individual domains – CS, Business & Psychology
 - Their failure to scale & lack of portability limited their adoption
- ← Web-based searching by the masses with non-specialized tools – CS & IS
- Now, new methods and tools that improve generic techniques for specialized needs and collections – IS + CS + LS

Why Specialized Search is Needed Now

- Specialized Populations with specialized needs
 - Aerospace engineers
 - Public health practitioners
 - Math and science teachers
 - Customer-facing CRM sites
 - Humanities scholars
- Generic searching of full web does not work well
 - Providers & users of search engines want Long Tail documents
 - NASA wants Aerospace Engineering students to use information for Reusable Launch Vehicle projects **only** from approved sites, not the most popular information according to the general public's behavior
 - CRM for each company is done on internally focused collection

Problems of Generic Search Technologies

- Winning generic search strategies do not work on smaller collections
 - In generic search, redundancy plays major role
 - Multiple formulations of answer exist in large collections
 - Light et al ('98) and Dumais et al ('02) have shown high redundancy of answers is positively correlated with good answer provision
 - Web has billions of pages
 - For collections of interest to the Long Tail, there are many, many fewer pages
 - In closed world, there may actually be only 1 answer
- Information elements of interest may not be indexed by generic techniques
 - Need more than high-frequency terms or citation elements
 - Generic schema of metadata elements may not capture features of interest to provide best results
 - Unique, specific facets are vital for different domains

New Indexing / Searching Techniques

1. Ontologies that provide rich, specialized vocabularies for improving representation of documents & enriching queries
 - Much is being learned from early library science systems
 - Synonyms, BTs, NTs, RTs
 - NLP toolkits to automate ontology construction
2. Text processing techniques that produce domain-based metadata to better represent a domain's elements of importance
 - May require development of new metadata schema
 - System utilizes specialized Information Extraction technology
3. More precise access to electronic books
 - Using value-tested existing book features
 - Does not require specialized knowledge-base development for each new domain
 - Can provide specific, contextualized information

1. Ontology Construction Tools

- Enable self-service solutions
 - Collection providers can tailor the toolkit to the domain vocabulary and important concept types
 - Semi-automated review and authoring tools enable high quality extraction of information contained within diverse document collections
 - Process enables rapid domain specialization with minimal user interaction
- Knowledge Bases built with the tools are primarily lexical-semantic taxonomic resources used by the system to create improved representation of the text
- Using automatically harvested data, collection provider can:
 - Review and alter categorization of entities and relations
 - Expand the underlying category taxonomy to the domain of interest

Vanilla Extract for Exploring a Collection

The screenshot shows the Vanilla Extract application window. The title bar reads "Vanilla Extract". The menu bar includes "File", "Options", and "Help". The main window displays the following information:

- Collection:** authoring (/home/eileen/tt1.5/release/kbb/AUTO/authindex)
- # Documents:** 24
- Phrases** | **Terms** | **Context** | **Documents**
- # Terms:** 10468 **Average Terms/doc:** 436.2
- # Unique Terms:** 2284 **Average Unique Terms/doc:** 95.2
- Lemmatize Words Frequency: # occurrences of term
- Use Stop List # docs containing term

The interface is divided into two main panes:

- Terms:** A list of terms with a "Frequency" column. The term "sport utility vehicle" is highlighted. An arrow points from this term to the "Show Context" button at the bottom.
- Co-occurring terms:** A list of terms that co-occur with the selected term, with a "# Docs" column. The top entries include "truck", "lineup", "full lineup", "half", "passenger_car", "decision", "crash", "other light-duty truck", "purchase", "automakers", "vehicle collision compatibility", "light-duty", "other manufacturer", "past decade and a half", "between", "offer", "bad", "particularly", "follow comment", "large", "weather", "collision", "comment", and "manufacturer".

At the bottom of the window, there are navigation controls:

- Left arrow, a slider, and a right arrow with the text "101-200/2260".
- A button labeled "Show Context" (circled in red).
- Left arrow, a slider, and a right arrow with the text "1-32/32".
- A button labeled "Show Docs".

On the right side of the "Co-occurring terms" pane, there are options for the "Co-occurrence Window":

- Contiguous
- 5 Terms
- 10 Terms
- Full Document

There is also a checkbox for "Use Stop List" which is checked.

2. Domain Based Metadata

- Robert Wood Johnson & NLM funded project
 - In conjunction with Anne Turner, MD, MPH, MLIS from UW
- Public Health professionals were asked to name 2 documents used in the last month that were important to their work
 - **59%** of the documents named were grey literature
 - Reported great difficulties in locating them on the web
- Interestingly, web search engines admit to only '*lightly indexing*' such lengthy, non-profitable documents
 - Take first X bytes of very long documents
 - What is indexed does not capture elements that matter to PH
- With PH experts, we developed a new metadata schema to produce summaries indexable by web search engines that reveal the essential elements of information in PH

Intervention Elements

PROBLEM

Description

Background Information

(Reports /Statistics /Guidelines/Protocols /Recommendations)

Description of Intervention

<p>Organizations</p> <p>Sponsoring /Funding /Affiliated</p> <p><i>Governmental</i></p> <ul style="list-style-type: none"> • Federal • State • County • Local <p><i>Non-governmental</i></p> <ul style="list-style-type: none"> • For-profit • Non-profit 	<p>Intervention Type</p> <ul style="list-style-type: none"> • Education • Prevention • Treatment • Surveillance 	<p>Methods</p> <p><i>Date/Duration</i></p> <p><i>Setting</i></p> <ul style="list-style-type: none"> • Individuals • Practitioners • Clinics • Hospitals • Institutions • Community <p><i>Target Population</i></p> <ul style="list-style-type: none"> • Age • Ethnicity • Gender • Employment • Geographic Location • Socio-Economic Status • Insurance Status <p><i>Evaluation</i></p>	<p>Outcomes</p> <ul style="list-style-type: none"> • Results / Findings • Knowledge Increase • Behavioral Change • Health Status Change • Guidelines / Recommendations
---	--	---	--

Information Produced

Type of Information

- Guidelines
- Newsletter
- Program Reports
- Meeting notes
- Policy Brief
- Statistics/Data
- Fact Sheet

Bibliographic Elements

- Title
- Creator
- Publishing agency
- Publication date
- URL
- Length of document

3. Searching Digital Book Segments

- More and more full length books in digital format are becoming important sources of information
 - But current bag-of-word approaches to searching digital books are not performing well
 - Frequently-occurring terms from a full book do not provide sufficiently localized access to specific topics contained in different segments of a book
 - Do not deal adequately with conceptually synonymous content
- Proposals to the I-Group and NEH
 - Utilize pre-existing book features to dramatically improve access to the content of electronic books

eBook Segment Indexing & Searching

- We're exploiting the intellectual effort already invested in:
 - Back-of-the-Book indexes
 - Tables of Contents
- Based on early work done at IST in the 70's & 80's
 - These guides to a book's content reflect solid intellectual effort that can be used by a system to create improved semantic indexing at book segment level
- System will integrate these sources using NLP-based extraction and organization algorithms to produce semantic indexing of each segment that will more accurately describe its content

Disciplines Involved in Search R & D

- Information Science
- Computer Science
- Library Science
- Linguistics
- Psychology
- Many of the social sciences
- Specialists from disciplines whom each system will be serving
 - Public Health
 - NASA engineers
 - CRM customers
 - Math & science teachers
 - Humanities scholars

Conclusion – Capitalize on Both

- The old techniques work well for the bulk of searches
 - Generic search on the largest collections, e.g. the web
 - Capture the distribution frequencies of new features
- But don't forget the Long Tail!
 1. More refined technologies for specific populations
 - Specialized search for smaller, focused collections where key words do not sufficiently distinguish between highly similar docs
 - Closer to a QA application
 2. Richer representation for crawling by standard web engines
- Inter-disciplinarity evident in I-Schools is vital!