Federated Search: Results Clustering

JR Jenkins, MLIS
Group Product Manager
Resource Discovery
Why Federated Search?

• The Web has changed how we deliver and consume information
  – The paradigm shift from physical to digital and electronic information has created new requirements and opportunities for delivering effective library experiences

• The Web has transformed the nature of library collections
  – The majority of new acquisitions are Web-based
  – Collections have increased dramatically and content is available anytime, anywhere
  – Web search engines compete with libraries
The Nature of Collections Has Changed!

• How do you define a library’s collection?
  – Physical location?
  – Monetary rights?
  – Access?
Global Connections, Global Collections!
Typical Library Subject Gateway

These pages provide access to resources - article databases, catalogs, reference books, web sites - organized by topic. These pages were created by the librarians responsible for the areas of study listed below.

- Aeronautics & Astronautics
- African Studies
- American Ethnic Studies
- American Indian Studies
- Anthropology
- Archaeology
- Architecture & Urban Planning
- Art
- Astronomy
- Atmospheric Sciences
- Bioengineering
- Biology
- Botany
- Business
- Canadian Studies
- Chemical Engineering
- Chemistry
- Children's Literature
- Chinese Studies
- Cinema Studies
- Civil & Environmental Engineering
- Classics
- Communication
- Comparative Literature
- Computer Science
- Dance
- Dentistry
- Drama
- East Asia (general)
- Earth & Space Sciences (Geology & Geophysics)
- Economics
- Education
- Electrical Engineering
- Engineering (general)
- English
- Environmental Sciences
- Fisheries
- Forest Resources
- French
- Gay & Lesbian Studies
- Geography
- German
- Government Publications
- Grants
- Health Sciences
- Hebrew Studies
- History
- History of Science
- Human Rights
- Industrial Engineering
- Information Science
- International Studies
- Italian
- Japanese Studies
- Jewish Studies
- Korean Studies
- Latin American Studies
- Law
- Linguistics
- Marine Studies
- Materials Science
- Mathematics & Applied Mathematics
- Mechanical Engineering
- Medicine
- Musicology
- Music
- Near East Studies
- News
- Nursing
- Oceanography
- Pharmacy
- Philosophy
- Physics
- Political Science & Public Affairs
- Psychology
- Public Health
- Religion
- Russian & East European Studies
- Scandinavian Studies
- Social Work
- Sociology
- South Asian Studies
- Southeast Asian Studies
- Speech & Language
- Statistics
- Technical Communication
- Western European Studies
- Women Studies
- Zoology
Search Today = UI Overload
Why Google?
Why Google?

“Why will our students not get up and walk a hundred meters to access a key journal article in the library? The natural human tendency in information seeking is to fall back on passive and sampling and selecting behaviors derived from millions of years of [evolution] .... It is not surprising, then, that the methods of access designed by librarians by librarians are little used.”

– Marcia Bates

Federated Search = Simple Starting Point

[Image of a federated search interface]

E-Resource Access and Management Services
From UI Overload To Citation Overload

Serials Solutions
E-Resource Access and Management Services
UI Overload to Citation Overload

Average Search Result Set

• 1000+ Citations spanning 40 pages (25 per page)
• From a wide variety of resources – with a wide variety of metadata structures (read taxonomies)
• Problem now is not a lack of information – but rather too much information
• “The main purpose of indexing and abstracting is to construct representations of published items in a form suitable for inclusion in some type of database”

• The goal of a retrieval environment is to maximize recall (useful items) and precision (non-useful items)
Lancaster – Database Evaluation Criteria

- Coverage
- Retrievability – How much of the literature on the topic, included in the database, can be found using “reasonable” search strategies?
- Predictability – Using information in the database, how well can a user judge which items are useful or not?
- Timeliness
‘Magellan’: What does the searcher want?

- The Explorer? (History)
- The Mutual Fund? (Business)
- The Venus Probe? (Science)
- The Straight of Magellan? (Geography)
- The Global Positioning System? (Technology)
- The Health Services Company? (Medicine)
How do we deal with this “Avalanche”?

- Personal Reference Librarians (Reference Interview)
- Sorting / Filtering Tools
- Relevance Ranking
- Visualization
- Clustering
3 Requirements for Relevance Ranking

1. Ability to perform textual analysis on documents (i.e. text mining) to look for relationships between documents
2. Analysis of user preferences (i.e. which resources they select)
3. Authority rules (what are the preferred / most appropriate resources)

This is extremely difficult in Federated Search!
Current Relevance Rules in Fed Search

- Simple term count / scoring system (Example rules)
  - Term in title – 3 points
  - Term in abstract – 2 points
  - Frequency of term in citation – 1 point for each
  - Rank in the Native Database – variable multiple
  - Recency of Article – variable multiple

- Ultimately, this is a simple mathematical analysis
- It does not consider the most important factor – CONTEXT of the information need
WebFeat – Relevance Ranking

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Biography Resource Center</td>
</tr>
<tr>
<td>431</td>
<td>Health and Wellness Resource Center</td>
</tr>
<tr>
<td>44</td>
<td>Columbia Granger's World of Poetry</td>
</tr>
<tr>
<td>24</td>
<td>Business and Company Resource Center</td>
</tr>
<tr>
<td>95</td>
<td>Entranca Online</td>
</tr>
</tbody>
</table>

1. □ Magellan, Iowa partner on recovery program for Medicaid consumers with SMI: innovative program first of its kind in managed care arena.
   - Journal: Mental Health Weekly
   - Pub.: 2006-08-21
   - Volume: 16
   - Issue: 33
   - Pages: 1
   - Database: Health and Wellness Resource Center

2. □ Magellan, Ferdinand
   - Author: Ferdinand Magellan
   - Source: Encyclopedia of World Biography, 2nd ed
   - Pub.: 1998
   - Database: Biography Resource Center

3. □ Magellan Health Services Enhances Recovery and Improves Effectiveness of Care Through Innovative Iowa Program.
   - Journal: Business Wire
   - Pub.: 2006-08-14
   - Pages: NA
   - Database: Health and Wellness Resource Center

4. □ Magellan officials pleased with second-quarter earnings: officials cite 'robust' Medicaid market for behavioral services.
   - Journal: Mental Health Weekly
   - Pub.: 2006-08-07
   - Volume: 16
   - Issue: 31
   - Pages: 1
   - Database: Health and Wellness Resource Center
Visualiziation
Visualization – From Grokker

E-Resource Access and Management Services
Visualization – Drilling Down
Grokker in EBSCO
Is Visualization Useful?

• “Grokker .... presents web search results visually in automatically generated categories; it’s interesting and fun to play with, but it’s not useful.”
  – Peter Morville – Ambient Findability
What is Clustering?

• Clustering is a technology that *organizes* results in a way that makes it easy for patrons to find what they’re looking for.

• Clustering achieves this by “clustering” *(grouping)* similar results together. It is an automated classification tool that organizes results as they are presented.

• Cluster types available include subject clusters, which are results-specific taxonomies, and facet based clusters such as author, journal titles, and date.
Focusing the Results via Clustering
Topic View – “Magellan”
4 Tabs – Topics, Date, Journal, Author

Topics are auto-generated by the Clustering engine.

Date, Journal, Author Tabs are “aggregations” of the information in the citation metadata.
Dynamic Clustering

Results from ProQuest, Produces a unique set of clusters

Results from Gale, Produces a unique set of clusters
Clustering, Visualization, Thought Clouds
Key Benefits of Results Clustering

Faster Research, Improved Discovery

- Find answers faster
- Eliminate information overload
- Increase use of Library collection
Find Answers Faster

Spend less time searching – more time for analysis and interpretation

• Turns a long list of citations into browseable Subject clusters
• More effective than “relevancy” - the “right” answer is often found in many citations
• Journal, Date, and Author facets help patrons drill down to specific items – as well as discover new ones
Eliminate Information Overload

Brings clarity to an avalanche of results

- Results can be viewed as small, coherent groups
- Clusters/groups should be concise, accurate, and distinctive
- Knowledge discovery and browsing are faster and easier than ever before possible
Ease of Use is Critical

- Less training required the better – not a separate action
- Text based links embrace existing Web usability patterns
- Facets enable patrons to explore results from a variety of vantage points with one click
- Hierarchical clusters/groups indicate relevance of terms without relying on “magic bullet” relevance ranking
1st Impressions

• "Excellent, smart and, pretty -- very smart and very pretty, a perfect couple with Central Search. This is the best implementation of Vivisimo and will make users appreciate libraries again."
  – Peter Jacso – University of Hawaii

• I think this will increase the usefulness of Central Search manyfold for our users.
  – Kimberly Kane – Ford Foundation
Check Them Out on Your Own!

• Central Search Basic Demo
  – http://demo.cs.serialssolutions.com/demo
  – username: sersol  password: search

• Kartoo
  – www.kartoo.com

• Grokker
  – www.grokker.com
The Future of Search?
Thank you for your time!

JR Jenkins, MLIS
Group Product Manager
Resource Discovery
jr@serialssolutions.com
1- 866-737-4257 ext. 1031
Mobile: 1-206-851-0038