Collection Development Insights Informed by Text Mining Virtual Reference

Michelle Polchow, Electronic Resources Librarian
Collection Development Insights Informed by Text Mining Virtual Reference
Washington Research Library Consortium Annual Meeting
March 6, 2018
Identify access to virtual reference data for text mining

Examine different types of patterns extracted by visualization software

Discover the inexact science of drawing conclusions from data regarding user behavior

Consider the ongoing impact of transforming data into actionable business analytics
**BACKGROUND**

Objective

- “to create a user experience that is high quality, consistent, and robust regardless of the user’s location, access method or objective”
- Summer 2015
  - created a team to examine the virtual reference chat transcripts as a potential data source to inform this objective

Library Web Tools Usability Committee

- Required periodic usability testing of patrons – quarterly
# DATA SOURCE AND MINING TOOLS

<table>
<thead>
<tr>
<th>Software</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibraryH3lp</td>
<td>Provided chat transcripts data</td>
</tr>
<tr>
<td>OpenRefine</td>
<td>Data cleanup</td>
</tr>
<tr>
<td>QDA Miner (WordStat)</td>
<td>Qualitative analysis (Categorization Dictionaries, Dendrograms, Mapping, etc.)</td>
</tr>
<tr>
<td>R</td>
<td>Data Cleaning (example: removal of personal information); Defining variables; Word cloud visualization</td>
</tr>
<tr>
<td>Notepad++</td>
<td>Spot-checked classification terms; Find compelling stories</td>
</tr>
</tbody>
</table>

[LibraryH3lp](https://libraryh3lp.com/)
[OpenRefine](https://about.infogr.am/wp-content/uploads/2016/01/Screen-Shot-2016-01-04-at-1.43.44-PM.png)
[Notepad++](http://image.freshshare.com/i0/id20/notepad++.png)
[R](https://developer.r-project.org/Logo/Rlogo-1.png)
[QDA Miner](http://www.tstat.it/tstat/wp-content/uploads/2016/04/qdalogo.jpg)
[WordStat](http://www.kovcomp.co.uk/wordstat/index.html)
Data Examination Objectives:

- Analyze the text of 15,441 virtual reference chat conversations from 2012-2015
- Applying text-mining analysis to identify patron issues pertaining to library resources and services
- Gain insight from text analytic patterns to reveal usability problems

Source: http://autowaredigital.com/analytics-beyond-the-value-of-information/
Examine different types of patterns extracted by visualization software
LIBRARY DATA: TRADITIONAL VS. BIG DATA

Count volumes, linear feet, compare institutions

What does this data reveal about how libraries are meeting user need?

http://www.aserl.org/about/stats/  https://library.gmu.edu/about/facts
**Literature Review – historic abundance of hand coding questions**

**QUESTION TYPE CATEGORIES AND SUBCATEGORIES**

<table>
<thead>
<tr>
<th>Circulation</th>
<th>Document Delivery</th>
<th>How to Find</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renew</td>
<td>Articles</td>
<td>Newspapers</td>
</tr>
<tr>
<td>Fines</td>
<td>Lumina to U</td>
<td>Articles by Author</td>
</tr>
<tr>
<td>Returning Books</td>
<td>Off Campus</td>
<td>Biographical Info</td>
</tr>
<tr>
<td>Lost U-Card</td>
<td>Point to Point</td>
<td>Book</td>
</tr>
<tr>
<td>Hold Service</td>
<td>ILL</td>
<td>Book Reviews</td>
</tr>
<tr>
<td>Journals</td>
<td>Interlibrary Loan–Service</td>
<td>Citations</td>
</tr>
<tr>
<td>Film/Video</td>
<td>Interlibrary Loan–Fees</td>
<td>Database</td>
</tr>
<tr>
<td>Book</td>
<td>Interlibrary Loan–Request Status</td>
<td>E-Journal</td>
</tr>
<tr>
<td>Loan Period</td>
<td></td>
<td>E-Reserves</td>
</tr>
<tr>
<td>MLAC</td>
<td></td>
<td>Films</td>
</tr>
<tr>
<td>In-Process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Order</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Circulating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rush Request</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**How to Use**

- Databases
- E-Resources
- Remote Access to E-Journals
- Remote Access to Databases
- NEILibrary
- ILL
- Export to RefWorks
- MNCAT

**Library Info**

- Hours
- Library Book Sale
- Internet Account

**Remote Access/Privileges (E-Resources)**

- Alumni
- CIS Student
- Graduate Student
- Undergraduate Student
- Distance Learner
- Unaffiliated Patron

**Style Assistance**

- APA Style
- Cover Page Format
- Citation Format
- MLA Style
- Turabian Style

**Subject-Specific (In-Depth) Research Question**

- E-Journal Access
- ILL Form
- ILL Article Access
- Internet Account

Conversations > High Frequency Terms > Visualizations
Team composed of multiple perspectives:
  • Science, Social Science, Humanities
  • Collection Development vs. Teaching & Reference Services

Longitudinal study about virtual reference users:
  • Text mining three years of virtual reference chats
  • Data findings encompassed all library services and resources

Examined both uniform data and conversation transcripts
  • Map activity to behavior pattern within semester
  • Interpret overall high frequency terms and phrases
  • Librarians’ professional expertise critical to interpretation

Original goal compared with findings:
  • Investigating its applicability for usability,
  • Discovery that value, based on depth and breadth of interdepartmental information, held value for business analytics.
Discover the inexact science of drawing conclusions from data regarding user behavior
If Google and Facebook are the masters of leveraging big data, consider the impact that institutional specific data might have for libraries to uncover causal relationships and predictive behavior from their user population.
A GOOD PLACE TO START - VISUALIZATION OF GEORGE MASON DATA

Ferrance, C., Lam, M., Polchow, M., "Where to Start Clustering," Virtual Chat Big Data Research Project, George Mason University, ©2015
Distinct frequency pattern:
Academic Search Complete

Conversation
Source:
Librarian – Black
Patron – Blue
L&P - Red

Big Data Research Project, George Mason University, ©2015
VR Chat text analysis:

Dendrogram - word association shows pattern of word frequency as well as the interrelationship of words that co-occur.

- Words associated with ‘journal’ are in red.
- Frequency indicated by bar chart to left
Text Analysis Classification

- 2-5 word phrases,

WordStat results for software classification:

<table>
<thead>
<tr>
<th>Name</th>
<th>Keywords</th>
<th>FREQ</th>
<th>CASES</th>
<th>% CASES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NEWSPAPER_SOURCE, WASHINGTON_POST</strong></td>
<td>NEWSPAPERSOURCE; WASHINGTON_POST; YORK_TIMES; WALL STREET JOURNAL; NEWSPAPER_ARTICLES; LEXISNEXIS_ACADEMIC; OPPOSING_VIEWPOINTS; SPECIFIC_ARTICLE; SEARCH_ENGINE;</td>
<td>651</td>
<td>556</td>
<td>4.21</td>
</tr>
</tbody>
</table>
Consider the ongoing impact of transforming data into actionable business analytics.
Good place to start >> “Academic Search Complete”

- Reference librarians have limited time to answer questions
- VR staffed at varying levels of subject matter familiarity
- Opportunity for improved usability – integrating resources & services

From collection development lens:
Could the expedient nature of VR negatively impact collection use, bypassing specialized handbooks, encyclopedias, bibliographies and other reference materials, when VR suggests ‘a good place to start’?
MAP USER EXPERIENCE TO SEQUENCE OF LEARNING PROCESS – MAXIMIZE COLLECTION

1. INDEXES AND DATABASES: A

2. SUBJECT GUIDES

3. SELECTED RESOURCES FOR: AFRICAN STUDIES

https://www.lib.umn.edu/subjects/rqs/4
IDENTIFY MISALLOCATION OF CRITICAL RESOURCES

Peer Reviewed Journal Discovery

From collection development lens:
Is there a correct balance in resource allocation, in terms of economics, to assure journal discovery?
Harvest Authentic User Conversations -- Newspapers

- Pattern of 2-5 word phrases referencing newspapers

- Collection includes several newspaper databases
  - Authoritative sources
  - High quality metadata to support advance search functionality
  - Extensive coverage for historical content
  - Multiple sources supporting narrative comparison

Why all the questions?

http://photo.gmu.edu/MasonThroughtheYears/Stills-2011-Year-in-Photos/i-XF94nhc
The collection development stories in the transcripts:

- Newspaper reading assignment for purpose of staying current and monitor emerging issues on daily basis not intuitively accomplished by using keywords.

- Users want online newspaper reading experience including images and multimedia together with text.

- Database search does not intuitively support serendipitous discovery.

- Databases do not receive exclusive online content made available directly to newspaper subscribers.

From collection development lens: Although the libraries had an abundance of newspaper resources, were the access points and the format reflecting the expressed needs of the university?
Database keyword search is distinctly different from an anticipated experience of reading an online newspaper.

https://global.factiva.com/factivalogin/login.asp?productname=iworks
Newspapers

Database
- Keyword search
- Known item citation
- Compare news sources
- Historic record

Subscription
- Browse, serendipitous discovery
- Current events, emerging issues
- Images with text
- Exclusive online content
Traditional data management systems are not capable of providing real-time intelligence, but transformative possibilities exist using text and data mining processes:

- better integrate library resources and services,
- identify misallocation of critical resources, and
- harvest authentic user conversations with compelling stories to build collaborative campus-wide relationships.
Questions
For more information
Michelle Polchow, mpolchow@gmu.edu, 703-993-3714


**SOURCES**

Association of Southeastern Research Libraries (ASERL) Member statistics, Fall 2017 Statistical Index, http://www.aserl.org/about/stats/


George Mason University, University Libraries, https://library.gmu.edu/about/facts


SOURCES

Images
Cover: http://photo.gmu.edu/CampusBuildingsLandmarksandFac/Fairfax-Campus/i-Hk5jsbt
http://www.kovcomp.co.uk/wordstat/index.html
Slide 5: http://autowaredigital.com/analytics-beyond-the-value-of-information/
Slide 20: https://www.lib.umn.edu/subjects/rqs/4
Slide 25: nytimes.com