Evaluation of Library Services

Min-Yen KAN
Why Evaluation?

- Run as a business, need to justify costs and expenditure
- Quantitative data analysis necessitated by evolution into automated and digital libraries
- Need benchmarks to evaluate effectiveness of library
Quantitative metrics

- Circulation per capita
- Library visits per capita
- Program attendance per capita
- Turnover rate
- Registration as % of population

- *Output measures for public libraries*
  Zweizig and Rodger (1982)
Evaluation types

- Macroevaluation
  - Quantitative, comparable statistics
  - Degree of exposure

- Microevaluation
  - Diagnostic
  - Gives rationale for performance

- Materials-Based / Use-based
  - Evaluate the items’ suitability
Exposure

○ Axiom
  • The more a book in a library is exposed, the more effective the library.

○ Defining “an exposure” as a simple count
  • Pros
    ○ Easy; can handle different levels of granularity
  • Cons
    ○ $5 \times 1$ day borrowing is five times more exposure than $1 \times 5$ day borrowing
    ○ Shorter circulation would increase counts
More exact ways to quantify exposure

- **Item-use days: Meier (61)**
  - A book borrowed for five days may not be used at all

- **Effective user hours: De Prospo et al. (73)**
  - Sample users in library

What about ways to quantify exposure in the digital library?
Bang for the buck?

The greater the exposure.

- The more index methods available
- The more copies provided
- The more titles provided
- The more branch locations
- The more liberal the usage period
- The more assistance given
- The more aware the public is
Number of items

460 items acquired

415 items not in circulation

348 items in correct location on shelves

331 items correctly located on shelves by user

500 items requested

Acquisition barrier

\( P_A = 0.92 \)

40 items not acquired

Circulation barrier

\( P_C = 0.90 \)

45 items in circulation

Library barrier

\( P_L = 0.84 \)

67 items not in correct location on shelves

User barrier

\( P_U = 0.95 \)

17 items not located on shelves by user

\[
P_S = P_A \times P_C \times P_L \times P_U
\]

\( P_S = 0.66 \)

Synergistic factors – Materials availability

Adapted from Kantor (76)
Effectiveness as Circulation

- Need a minimal size to function at all
- The larger the collection the better... ... to a point

- From Hodowanec (78)
Macroevaluation

- In general, more exact measures require aggregating *sampling*, which tend towards microevaluation
  - So it’s a continuum after all

- Administrators use a battery of measures; not a single one, to measure effectiveness – Spray (76)
Microevaluation

Drilling down to the individual needs level

- The more concrete the need, the easier to evaluate
- Failure is harder to measure than success
  - Case 1: Got a sub-optimal resource
  - Case 2: Got some material but not all
Material-centered collection evaluation

What’s the purpose...

... of the collection
  - Who’s the readership – academic, public?

... of the evaluation
  - Document change in demand?
  - Justify funding?
  - Select areas to weed materials?
  - Adjust shelving/organization?
Material-based evaluations

- Checklist
  - Use standard reference bibliographies to check against

- Citation
  - Use an initial seed of resources to search for resources that cite and are cited by them

Are these methods really distinct?
  - How do people compile bibliographies in the first place?
Collection Mapping

- **Idea:** Build the collection in parts
  - Prioritize and budget specific subjects
    - Shrink, grow, keep constant
  - Evaluate subjects according to specific use
    - Which courses it serves, what are each courses’ needs

To think about:
- Which of these approaches are *micro* and which are *macro*?
Use Factors

- Age
- Language
- Subject
- Shelf Arrangement
- Quality
- Expected Use
  - Popularity
  - *Information Chain* placement
Use-based evaluation

- Physical Library
  - Slips
  - Circulation records
  - Table Counting

- Digital Library
  - Download counts
  - Citation counts (in scholarly works)

How do these two relate to each other?
MESUR project
Figure 5: Principal component analysis of Spearman rank-order correlations between 47 preliminary MESUR metrics.
Digital Libraries

IR Evaluation Metrics
Min-Yen KAN

* - Parts of this lecture come from Lilian Tang’s lecture material at the Univ. of Surrey
## Evaluation Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>System says is <strong>relevant</strong></th>
<th>System says is <strong>irrelevant</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document is actually relevant</strong></td>
<td>TP (True Positive)</td>
<td>FN (False Negative)</td>
</tr>
<tr>
<td><strong>Document is actually irrelevant</strong></td>
<td>FP (False Positive)</td>
<td>TN (True Negative)</td>
</tr>
</tbody>
</table>
# Sensitivity, specificity, positive and negative predictive value

<table>
<thead>
<tr>
<th>Test (System)</th>
<th>Relevant</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
<td>All with Positive Test TP+FP</td>
<td><strong>Positive Predictive Value</strong></td>
<td>= TP / (TP+FP)</td>
</tr>
<tr>
<td>-</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
<td>All withNegative Test FN+TN</td>
<td><strong>Negative Predictive Value</strong></td>
<td>= TN / (FN+TN)</td>
</tr>
<tr>
<td></td>
<td>All Relevant</td>
<td>All non-relevant</td>
<td>All documents = TP+FP+FN+TN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Sensitivity** = $\frac{TP}{TP+FN}$
- **Specificity** = $\frac{TN}{FP+TN}$
- **Pre-Test Probability of Relevance** = $\frac{(TP+FN)}{(TP+FP+FN+TN)}$ (in this case = prevalence)
Evaluation Metrics

- **Precision = Positive Predictive Value**
  - \[ \frac{TP}{TP + FP} \]
  - “ratio of the number of relevant documents retrieved over the total number of documents retrieved”
  - how much extra stuff did you get?

- **Recall = Sensitivity**
  - \[ \frac{TP}{TP + FN} \]
  - “ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database”
  - how much did you miss?
## P/R: an example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Decision</th>
<th>R@r (%)</th>
<th>P@r (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R</td>
<td>10%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>10%</td>
<td>33%</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>20%</td>
<td>50%</td>
</tr>
<tr>
<td>5</td>
<td>R</td>
<td>30%</td>
<td>60%</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>7</td>
<td>R</td>
<td>40%</td>
<td>57%</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>40%</td>
<td>44%</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>40%</td>
<td>36%</td>
</tr>
<tr>
<td>12</td>
<td>R</td>
<td>50%</td>
<td>42%</td>
</tr>
<tr>
<td>13</td>
<td>R</td>
<td>60%</td>
<td>46%</td>
</tr>
<tr>
<td>14</td>
<td>R</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>R</td>
<td>100%</td>
<td>45%</td>
</tr>
</tbody>
</table>

From: Managing Gigabytes
Interpolated precision gives a non-increasing curve.

But it doesn’t factor in the size of the corpus.

- Previous example on a corpus of 25 docs = 40% precision
- On a corpus of 2.5 M docs = also 40%
Factoring in size of a corpus

- Look at how P/R or Sn/Sp varies as a function of rank:

- Choose a number of different ranks and calculate P/R or Sn/Sp
  - Correspond to vertical lines on graphs at right
  - Plot Sn vs. 1-Sp to get points for ROC curve. Interpolate curve.

Which of these examples is which from the previous slide?
Look at the **probability** or rate of detection

- What does the diagonal represent?

- How do we compare ROC curves versus each other?
Getting a single number

- 11 pt average
  - Average precision at each .1 interval in recall

- Precision at recall point (% or absolute)

- F Measure
  - Ratio of precision to recall: \( F_b = \frac{(b^2+1) \text{ PR}}{b^2 \text{ P} + R} \) (e.g., \( F_3 \) = weight precision heavier)

- Area under ROC curve (Accuracy)
  - \( 1 = \) perfect, \( .9 \) excellent, \( .5 \) worthless

- What’s the difference between these measures?
- Which measures are best suited to which scenarios?