Science Mapping and Applications: Choices and Trade-offs

Standards Workshop
August 11, 2011

Kevin W. Boyack, SciTech Strategies
Agenda

- Background
- Applications
- Choices and Trade-offs
- SciTech Choices
- SciTech Mapping Process
- Applications Enabled by SciTech Process
- Additional Findings Related to Choices
- Summary
Science Mapping

- Most common definition: visualizing or creating a visual map of scientific information
  - Documents, journals, authors, words …

- This requires
  - Data acquisition and cleaning
  - Defining elements and similarity
  - Partitioning and/or layout of the data

- And then one creates the visual
Why Do We Map Science?

- Basic understanding
  - Structure and dynamics of science
    - High level, all of science
    - Discipline, specialty
    - Researchers and groups

- Search and retrieval
  - “More like this”, “related articles”

- Metrics
  - Research
  - Evaluation

- Policy / Decision Making
Why Do We Map Science?

Academic View

- **Structure and dynamics**
  - High level, all of science
  - Discipline, specialty
  - Researchers and groups

- **Search and retrieval**
  - “More like this”, “related articles”

- **Metrics**
  - Research
    - Evaluation
  - Policy

World View

- **Structure and dynamics of science**
  - High level, all of science
  - Discipline, specialty
  - Researchers and groups

- **Search and retrieval**
  - “More like this”, “related articles”

- **Metrics**
  - Research
  - **Evaluation**

- **Policy**
Map of Science & Technology Indicators

2004 - 2008

METRICS

POLICY

MARKETS
**Choices**

- Data source
- Unit of analysis
  - Document, Journal, Author, Word (phrase)
- Sample size
  - Specialty, Discipline, All of science, Single year, Multiple years
- Similarity approach
  - Citation, Text, Thresholds
- Partitioning and/or layout
  - MDS, K-K, F-R, DrL (OpenOrd), Blondel, …
## Trade-offs

<table>
<thead>
<tr>
<th>Choice</th>
<th>Trade-offs</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data source</td>
<td>- Free vs. Costly</td>
<td>- Citation data isn’t free</td>
</tr>
<tr>
<td></td>
<td>- Single vs. Multiple Sources</td>
<td>- De-duplication isn’t trivial</td>
</tr>
<tr>
<td></td>
<td>- Content vs. Context</td>
<td>- Coverage comes at a cost</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>- Breadth vs. Detail</td>
<td>- Base on research question</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Don’t base on data available</td>
</tr>
<tr>
<td>Sample size</td>
<td>- Specialty vs. All of science</td>
<td>- Mapping ALL is costly</td>
</tr>
<tr>
<td></td>
<td>- Content vs. Context</td>
<td>- Specialty maps lack context</td>
</tr>
<tr>
<td></td>
<td>- Single year vs. Multiple years</td>
<td>- Stability or instability?</td>
</tr>
<tr>
<td>Similarity approach</td>
<td>- Simple vs. Complex</td>
<td>- Comp cost: Index &lt; Vector</td>
</tr>
<tr>
<td></td>
<td>- Citation vs. Text vs. Hybrid</td>
<td>- Hybrid costly, but likely best</td>
</tr>
<tr>
<td></td>
<td>- Threshold vs. Accuracy</td>
<td>- How much is really needed?</td>
</tr>
<tr>
<td>Partitioning / Layout</td>
<td>- Simple vs. Complex</td>
<td>- Simple often size limited</td>
</tr>
<tr>
<td></td>
<td>- Accuracy vs. Useability</td>
<td>- Is intuition satisfied?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Are distributions reasonable?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Useful levels of aggregation?</td>
</tr>
</tbody>
</table>
Mapping Choices

- Are often made based on what is available
  - Data, algorithms, expertise

- When they should be made based on
  - The research question or application
  - Balancing of the applicable trade-offs

- If the application is EVALUATION
  - Special care must be taken
  - Partition accuracy is critical
    - What is a discipline?
### Our Choices

<table>
<thead>
<tr>
<th>Choice</th>
<th>Trade-offs</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data source</strong></td>
<td>- Free vs. Costly</td>
<td>- Citation data is essential and worth paying for</td>
</tr>
<tr>
<td></td>
<td>- Single vs. Multiple Sources</td>
<td>- Future: full text?</td>
</tr>
<tr>
<td></td>
<td>- Content vs. Context</td>
<td></td>
</tr>
<tr>
<td><strong>Unit of analysis</strong></td>
<td>- Breadth vs. Detail</td>
<td>- Documents</td>
</tr>
<tr>
<td><strong>Sample size</strong></td>
<td>- Specialty vs. All of science</td>
<td>- Mapping ALL of SCIENCE</td>
</tr>
<tr>
<td></td>
<td>- Content vs. Context</td>
<td>- Context is very useful</td>
</tr>
<tr>
<td></td>
<td>- Single year vs. Multiple years</td>
<td>- Link single years</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Show instability</td>
</tr>
<tr>
<td><strong>Similarity approach</strong></td>
<td>- Simple vs. Complex</td>
<td>- Co-citation analysis</td>
</tr>
<tr>
<td></td>
<td>- Citation vs. Text vs. Hybrid</td>
<td>- Future: hybrid likely</td>
</tr>
<tr>
<td></td>
<td>- Threshold vs. Accuracy</td>
<td>- Top-n only, rest is noise</td>
</tr>
<tr>
<td><strong>Partitioning / Layout</strong></td>
<td>- Simple vs. Complex</td>
<td>- Complex (DrL x 10), robust</td>
</tr>
<tr>
<td></td>
<td>- Accuracy vs. Useability</td>
<td>- Small clusters</td>
</tr>
</tbody>
</table>

- We are continually testing, refining, and improving our processes
### Standards?

<table>
<thead>
<tr>
<th>Choice</th>
<th>Trade-offs</th>
<th>Standards?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data source</td>
<td>- Free vs. Costly</td>
<td>Bibliographic data</td>
</tr>
<tr>
<td></td>
<td>- Single vs. Multiple Sources</td>
<td>Cleaning to different levels</td>
</tr>
<tr>
<td></td>
<td>- Content vs. Context</td>
<td>Use what you can get!</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>- Breadth vs. Detail</td>
<td>Articles, journals, authors</td>
</tr>
<tr>
<td>Sample size</td>
<td>- Specialty vs. All of science</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Content vs. Context</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Single year vs. Multiple years</td>
<td></td>
</tr>
<tr>
<td>Similarity approach</td>
<td>- Simple vs. Complex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Citation vs. Text vs. Hybrid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Threshold vs. Accuracy</td>
<td></td>
</tr>
<tr>
<td>Partitioning / Layout</td>
<td>- Simple vs. Complex</td>
<td>Anecdotal validation in most cases</td>
</tr>
<tr>
<td></td>
<td>- Accuracy vs. Useability</td>
<td></td>
</tr>
</tbody>
</table>
Possible Standards

- Units
- Data sources
- Data cleanliness
- Workflows
- Validation
- Datasets
- Products (e.g. reference maps?)
- …
Our Process

Generate annual models, 2000-2008

- Select single year of source data
- Apply thresholds to select references
- Generate relatedness matrix (k50/cosine)
- Filter matrix to top-n values per reference

Research communities

- Cluster references → intellectual base
- Assign current papers → research front

Link annual models

Link communities from adjacent years using overlaps in their intellectual bases → threads
Process-Specific Applications

- Highly granular model of science gives detailed structure
Process-Specific Applications

- Detailed structure enables metrics for multidisciplinary sets of topics
- Competencies is the name we use for these sets of topics
Process-Specific Applications

- Research leadership (based on competencies) rather than research evaluation (based on rankings)
Process-Specific Applications

- Mapping all of science shows context for target literature
Process-Specific Applications

- Linking annual models shows detailed evolution and instability
- Consider a topic in a given year. It can be linked (or not) in time in the following ways:

  ![Diagram showing prior and following year linkages](image)
Process-Specific Applications

- Linkage types correlate with various properties (e.g. textual coherence, prolific authors)
- These can be used as predictors of future linkage
  » Identification of emerging topics at the micro-level

<table>
<thead>
<tr>
<th>Prior year linkage</th>
<th>Isolate 34%</th>
<th>Birth 14%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death 11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuation 41%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stasis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Following year linkage

- Isolate
- Birth
- Death
- Continuation
Process-Specific Applications

- Superstar authors (top 1%) are predictive
  - Superstars avoid isolates
  - Superstars move out of topics before they die
  - Superstars are heavily involved in splits and merges (non-stasis parts of continuation)

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>34%</td>
<td>14%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Death</th>
<th>Stasis</th>
<th>Continuation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11%</td>
<td></td>
<td>41%</td>
</tr>
</tbody>
</table>
Accuracy of Similarity Approaches

- Identified a large data corpus (2.15M articles)
  - Textual information (abstracts, MeSH) from Medline
  - Citation information from Scopus

- Compare 13 approaches – 3 citation, 9 text, 1 hybrid

- For each approach
  - Calculate article-article similarities
  - Filter similarities to reduce the number to a similarity file of reasonable size (2.15M articles, 20M sim pairs)
  - Cluster the articles
  - Measure accuracy of the cluster solution using multiple metrics
    - Coherence (Jensen-Shannon divergence)
    - Grant-to-article linkage concentration

- Analyze results
# Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Comp Cost</th>
<th>% Coverage</th>
<th>Coherence (combined)</th>
<th>Herfindahl</th>
<th>Pr80 (all grants)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-word MeSH</td>
<td>Medium</td>
<td>95.77%</td>
<td>0.0758</td>
<td>0.1631</td>
<td>0.2216</td>
</tr>
<tr>
<td>LSA MeSH</td>
<td>Very high</td>
<td><strong>98.22%</strong></td>
<td>0.0479</td>
<td>0.1124</td>
<td>0.2127</td>
</tr>
<tr>
<td>SOM MeSH</td>
<td>Very high</td>
<td><strong>99.97%</strong></td>
<td>0.0409</td>
<td>0.1106</td>
<td>0.2203</td>
</tr>
<tr>
<td>bm25 MeSH</td>
<td>Medium</td>
<td>93.39%</td>
<td>0.0759</td>
<td>0.1570</td>
<td>0.2167</td>
</tr>
<tr>
<td>Co-word TA</td>
<td>High</td>
<td>83.41%</td>
<td>0.0685</td>
<td>0.1299</td>
<td>0.1571</td>
</tr>
<tr>
<td>LSA TA</td>
<td>Very high</td>
<td>90.92%</td>
<td>0.0715</td>
<td>0.1255</td>
<td>0.2003</td>
</tr>
<tr>
<td>Topics TA</td>
<td>High</td>
<td>94.40%</td>
<td>0.0875</td>
<td>0.1584</td>
<td>0.2379</td>
</tr>
<tr>
<td>bm25 TA</td>
<td>High</td>
<td>93.91%</td>
<td>0.0979</td>
<td>0.2393</td>
<td>0.2578</td>
</tr>
<tr>
<td>pmra</td>
<td>Low</td>
<td>94.23%</td>
<td><strong>0.1055</strong></td>
<td>0.2410</td>
<td>0.2637</td>
</tr>
<tr>
<td>Bib coupling</td>
<td>Low</td>
<td>96.62%</td>
<td>0.1001</td>
<td><strong>0.2849</strong></td>
<td><strong>0.2706</strong></td>
</tr>
<tr>
<td>Co-citation</td>
<td>Low</td>
<td><strong>98.37%</strong></td>
<td>0.0947</td>
<td>0.2378</td>
<td>0.2621</td>
</tr>
<tr>
<td>Direct citation</td>
<td>Low</td>
<td>92.68%</td>
<td>0.0702</td>
<td>0.2037</td>
<td>0.2480</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Medium</td>
<td>96.83%</td>
<td><strong>0.1014</strong></td>
<td><strong>0.2893</strong></td>
<td><strong>0.2752</strong></td>
</tr>
</tbody>
</table>
Summary

- Science mapping involves a great number of choices
  » There are many trade-offs to consider
  » Choices constrain or enable analysis types

- SciTech has made a set of mapping choices, and developed a methodology that enables
  » Development of a detailed, multi-year model of science for structure and evolution
  » Metrics on competencies (sets of topics) that are important at the institutional level
  » Context to be mapped along with any target literature
  » Prediction of continuance
  » Identification of emerging topics

- We continue to explore accuracy and usefulness of maps
Relevant Literature


Thank you!