Information Visualization MOOC

Unit 4 – “What”: Topical Data

Design and Update of a Classification System: The UCSD Map of Science

Relevant Research Disciplines:
Scientometrics, Information Sciences

http://ivmooc.cns.iu.edu
Design and Update of a Classification System: The UCSD Map of Science

1. Original map
2. Initial Update Using Scopus
3. Final Updated Map
4. Validation
5. Applications


1. Original Map (5 years of Scopus and WoS)

Data: The original classification and map use 7.2 million papers and their references from Elsevier’s Scopus (about 15,000 source titles, 2001–2005) and Thomson Reuters’ Web of Science (WoS) Science, Social Science, and Arts & Humanities Citation Indexes (about 9,000 source titles, 2001–2004)–about 16,000 unique source titles.
**Similarity Metric:** Combination of bibliographic coupling and keyword vectors.

**Layout:** The 554 subdisciplines were laid out on the surface of a sphere; the spherical layout is then flattened using a Mercator projection to create a two-dimensional version of the map. Clusters are further aggregated into 13 main scientific disciplines that are labeled and color coded.

**Data Overlays:**
Each node is labeled and has an extensive list of key phrases as metadata, which can be used to “science locate” nonjournal data, such as patents or grants. That is, key phrases from each patent or grant (titles and abstracts) are extracted; fractional assignment to map nodes proceeds by matching the associated metadata. Thus, each grant or patent is fractionally assigned to multiple nodes. Adding the fractions allows for the number of grants, dollars by agency, or patents associated with each node to be computed.

**Problem:** As time passes, new journals are created (e.g., *PLoS*) that cannot be mapped.
2. Initial Update (5 years + Scopus) by Klavans & Boyack

**Data:** In June 2009, 7,464 new source titles (2006–2008) from Scopus were added to the existing category structure.

**Process:** Identified all new journals that were not in the existing classification system, and assigned each new journal to one of the existing categories by counting the numbers of times journals in each category were referenced by the articles in the new journals. Each journal was assigned to the category that it referenced the most, as long as it cited articles within that cluster a minimum of 10 times.

**Result:** Update increased the number of Scopus journals in the classification system by 47%; this only accounted for a 13% increase in the number of articles.
3. Full Update (10 years of Scopus and WoS)
by Börner, Klavans, Patek, Zoss, Biberstine, Light, Larivière, & Boyack

Desirable features for a map of science classification system:
1. Use highest quality/coverage paper-level data to generate the science map classification system. Using journal-level data or highly cited papers exclusively leads to distortions [22].
2. Employ advanced dimensionality reduction techniques to map a high-dimensional semantic space to a two-dimensional map that preserves the most important data structures [23].
3. Select a clustering and layout that has easy to read, distinct clusters (e.g., subdisciplines, which have about the same number of records), are disjoint (i.e., they do not overlap or occlude one other), and have meaningful labels to ease data interpretation and communication. The map must match the typical viewer’s mental model of the domain.
4. Use graphic design (color, shape, size coding) and legend that can be understood by a large audience—map must empower users to form new hypotheses and get new answers.

Desirable features for a map of science classification system (cont.):
5. Support interactivity—e.g., zoom, filter, details on demand [24]. Multi-level maps—e.g., two-levels comprising subdisciplines aggregated into disciplines, support multi-level studies.
6. Define a mapping process to classify new data and overlay it onto the map, e.g., journals based on journal names and other records, e.g., patents, funding data based on keywords. As users have a hard time with fractional associations/counting, each record should be associated with one or few subdisciplines.
7. The science map and classification system should be easy to update to capture the continuously evolving structure of science. Computational workflow should be well documented so that it is easy to understand in principle and can be replicated by other experts. Updates should preserve the main structure of the map as much as possible.
8. Alignment and comparison of any new science map and classification with commonly used science classifications (e.g., classifications used by Thomson Reuters’ databases, Elsevier’s Scopus, the Library of Congress, Universal Decimal Classification) and the translation of major ontologies into different languages (Science-Metrix, [25]).
Data: The updated map and classification adds six years (2005–2010) of WoS data and three years (2006–2008) from Scopus to the existing category structure—increasing the number of source titles to about 25,000.

Process:
For each of the 4,021 new journals, we counted the number of citations to/from papers published in that journal to/from each subdiscipline of the original map. This yielded for each journal an outgoing and incoming citation count for each subdiscipline of the original map. To account for the fact that some subdisciplines publish more papers than others and that, thus, the probability of citing and being cited by these subdisciplines is greater than for smaller ones, we normalized each of these citation counts by the total number of papers published among all journals assigned (even only fractionally) to that subdiscipline. The top subdiscipline citing/cited was then assigned to these new journals.

Multidisciplinary journals: PLoS ONE and SCHWEIZERISCHE MEDIZINISCHE WOCHENSCHRIFT (Swiss Medical Weekly) have the highest combined relative importance across sub-disciplines, yet were assigned to one subdiscipline.
To further simplify the 2010 UCSD map, all multi-assigned journals were examined and only 34 were kept, among them Science, Nature, The Lancet, British Medical Journal, and Journal of the American Medical Association.

Results:
A comparison of the original 5-year and the new 10-year maps and classification system show: (i) an increase in the total number of journals that can be mapped by 9,409 journals (social sciences increased by 80%, humanities by 119%, medical by 32%, and natural science by 74%); (ii) a simplification of the map by assigning all but five highly interdisciplinary journals to exactly one discipline; (iii) a more even distribution of journals over the 554 subdisciplines and 13 disciplines when calculating the coefficient of variation; and (iv) a better reflection of journal clusters when compared with paper-level citation data. When evaluating the map with a listing of desirable features for maps of science, the updated map is shown to have higher mapping accuracy, easier understandability as fewer journals are multiply classified, and higher usability for the generation of data overlays, among others.

To our knowledge, this is the first time that a widely used map of science was updated.
Figure 4. Number of journals per discipline for 5-year (grey) and 10-year (black) UCSD science map. doi:10.1371/journal.pone.0039464.g004

Figure 5. Number of journals per subdiscipline for 5-year (grey/red circles) and 10-year (black line) UCSD science map. Inset: distribution of the gain in number of journals for each subdiscipline (a). Number of (fractionally assigned) terms per 554 subdisciplines (b). doi:10.1371/journal.pone.0039464.g005
Deployment:
The UCSD map of science data is available at http://sci.cns.iu.edu/ucsdmap/

Data:
The 2010 UCSD map of science and classification system covering 10 years (2001-2010) of Web of Science data and 8 years (2001-2008) of Scopus data with subdiscipline assignments by SciTech Strategies.

1. Data as MS AccessDB and as MS Excel file (identical info as MS AccessDB) as well as data dictionary and database schema.
2. Network .net file to visually render science map. Also provided as .net file with discipline nodes and names.

Usage Conditions
This map is shared under the Creative Commons, Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) license (http://creativecommons.org/licenses/by-nc-sa/3.0/). That is, you are free to share, e.g., to copy, distribute and transmit the work, and to remix, i.e., to adapt the work under the following conditions:

- Attribution — You must attribute the work in the following manner (but not in any way that suggests that they endorse you or your use of the work): Cite the above paper and use the following acknowledgment text: "The authors wish to acknowledge The Regents of the University of California, SciTech Strategies, Observatoire des Sciences et des Technologies, and the Cyberinfrastructure for Network Science Center for making the 2010 UCSD Map of Science and Classification System available for this work."

- Noncommercial — You may not use this work for commercial purposes.

Data:
The 2010 UCSD map of science and classification system covers ten years (2001-2010) of data from Thomson Reuters’ Web of Science and eight years (2001-2008) of Elsevier’s Scopus, specifically the fractional assignment of about 25,000 journal names to 554 subdisciplines grouped into 13 disciplines of science.

The counts for major record types are given here:
1. 13 disciplines with labels and color codes
2. 554 subdisciplines with x, y positions and size
3. 15,849 journals captured by 5-year map
4. 25,258 journals captured by 10-year map
5. 13,520 journal names used by Thomson Reuters
6. 22,253 journal names used by Scopus
7. 21,630 Scopus journal ID numbers
8. 19,988 ISSN numbers
9. 66,759 terms

UCSD map table schema

**Note:** There are no standards on how to render .net files!

Some define the zero point on the top left (e.g., GUESS), while others define the bottom-left point as 0,0 (e.g., Gephi, Pajek). This only becomes important when rendering a dataset that has a predefined left and right, top and bottom such as the UCSD map of science.

Simply multiply all node's y-position with -1 to solve this issue.
4. Validation: Consensus Map of Science


20 maps of science were examined and found to have a high level of correspondence.
M – Mathematics
CS – Computer science
P – Physics
PC – Physical chemistry
C – Chemistry
E – Engineering
G – Earth sciences (geoscience)
BC – Biochemistry
B – Biology
I – Infectious disease
MD – Medical specialties
HS – Health services
N – Brain research (neuroscience)
PS – Psychology/psychiatry
SS – Social sciences
H – Humanities

Line size/color = % of maps

90% or more
70% - 90%
50% - 70%

FIG. 6. One-dimensional consensus maps of science, Euclidean (top) and Riemannian (bottom).
5. UCSD Map Applications

- Illuminated Diagram Display
- VIVO National Researcher Network
- MapSustain Interactive Online Interface
- Sci2 Tool
Topical Analysis (What) Science map overlays will show where a person, department, or university publishes most in the world of science. (in work)
The science map at 554 sub-disciplines level.

Science of Science (Sci2) Tool

Download Sci2 Tool v1.0 Alpha (June 13, 2012) from http://sci2.cns.iu.edu
Unpack into a /sci2 directory. Run /sci2/sci2.exe
Sci2 Manual is at http://sci2.wiki.cns.iu.edu

Load an ISI (*.isi), Bibtex (*.bib), Endnote Export Format (*.enw), Scopus csv (*.scopus) file such as /sci2/sampledata/scientometrics/isi/FourNetSciResearchers.isi

Run Visualization > Topical > Science Map via Journals using parameters given to the right.

Postscript file will appear in Data Manager. Save and open with a Postscript Viewer.
Topical Visualization
Generated from 361 Unique ISI Records
90 out of 112 publications were mapped to 182 subdisciplines and 13 disciplines.
June 24, 2012 | 04:04 PM EDT

Legend
Circle area: Fractional Journal Count
Undeclassified = 22
Minimum = 0
Maximum = 98
Color: Discipline
See end of PDF for color legend.

Area
29.09
15.10
2.8

How To Read This Map
The UCSD map of science depicts a network of 554 subdiscipline nodes that are aggregated to 13 main disciplines of science. Each discipline has a distinct color and is labeled. Overlaid are circles, each representing all records per unique subdiscipline. Circle area is proportional to the number of fractionally assigned records. Minimum and maximum data values are given in the legend.

Topical Visualization
Generated from 361 Unique ISI Records
90 out of 112 publications were mapped to 182 subdisciplines and 13 disciplines.
June 24, 2012 | 04:04 PM EDT

- **Biography**
  1. BMC EVOLUTIONARY BIOLOGY
  2. NATURE NEWS

- **Biotecnology**
  1. BMC BIOINFORMATICS
  2. FEBS JOURNAL
  3. GENOME RESEARCH
  4. INTERNATIONAL MICROBIOLOGY
  5. NATURE BIOTECHNOLOGY
  6. NATURE GENETICS
  7. NATURE Reviews Genetics
  8. NUCLEIC ACIDS RESEARCH
  9. PROTEOMICS

- **Brain Research**
  1. JOURNAL OF MATHEMATICAL PSYCHOLOGY

- **Chemical, Mechanical, & Civil Engineering**
  1. JOURNAL OF CERAMIC PROCESSING RESEARCH
  2. MATERIALS SCIENCE AND ENGINEERING A-STRUCTURAL MATERIALS
  3. PHYSICS WORLD
  4. SCIENTIFIC AMERICAN

- **Chemistry**
  1. COMPUTER PHYSICS COMMUNICATIONS
  2. JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES
  3. JOURNAL OF THE INDIAN INSTITUTE OF SCIENCE
  4. PURE AND APPLIED CHEMISTRY

- **Earth Sciences**
  1. CURRENT SCIENCE

- **Electrical Engineering & Computer Science**
  1. ASSIST 2003: PROCEEDINGS OF THE 8TH ASSIST ANNUAL MEETING....
  2. CANADIAN JOURNAL OF INFORMATION AND LIBRARY SCIENCE-REV...
  3. IEEE TRANSACTIONS ON PROFESSIONAL COMMUNICATION
  4. INFORMATION TECHNOLOGY AND LIBRARIES
  5. JOURNAL OF INFORMATION SCIENCE
  6. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE...
  7. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE...
  8. LIBRARY QUARTERLY
  9. LIBR.
  10. PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SC..

- **Health Professionals**
  1. ANNALS OF BIOMEDICAL ENGINEERING
  2. BULLETIN OF THE MEDICAL LIBRARY ASSOCIATION
  3. CROATIAN MEDICAL JOURNAL
  4. JOURNAL OF APPLIED PHYSIOLOGY
  5. JOURNAL OF PUBLIC HEALTH DENTISTRY
  6. METHODS OF INFORMATION IN MEDICINE
  7. PLASTIC AND RECONSTRUCTIVE SURGERY
  8. TEXAS MEDICINE
  9. UNFALLCHIRURG
  10. WIENER KLINISCHE WOCHENSCHRIFT

- **Humanities**
  1. BULLETIN OF THE ATMOSPHERIC SCIENTISTS

- **Infectious Diseases**
  1. FEMS MICROBIOLOGY LETTERS
  2. JOURNAL OF BACTERIOLOGY

- **Math & Physics**
  1. ADVANCES IN APPLIED PROBABILITY

CNS (cns.iu.edu)
In addition to using journal names to
- map career trajectories
- identify evolving expertise areas, and
- compare expertise profiles,
**existing classifications** can be aligned and used to generate science map overlays.

Run Visualization > Topical > Science Map via 554 Fields using parameters given to the right.
Postscript file will appear in Data Manager.
Save and open with a Postscript Viewer.
Sci2 Tool now supports Web services and serves as a visual interface to publically available NIH RePORT Expenditure and Results RePORTER data provided by NIH.