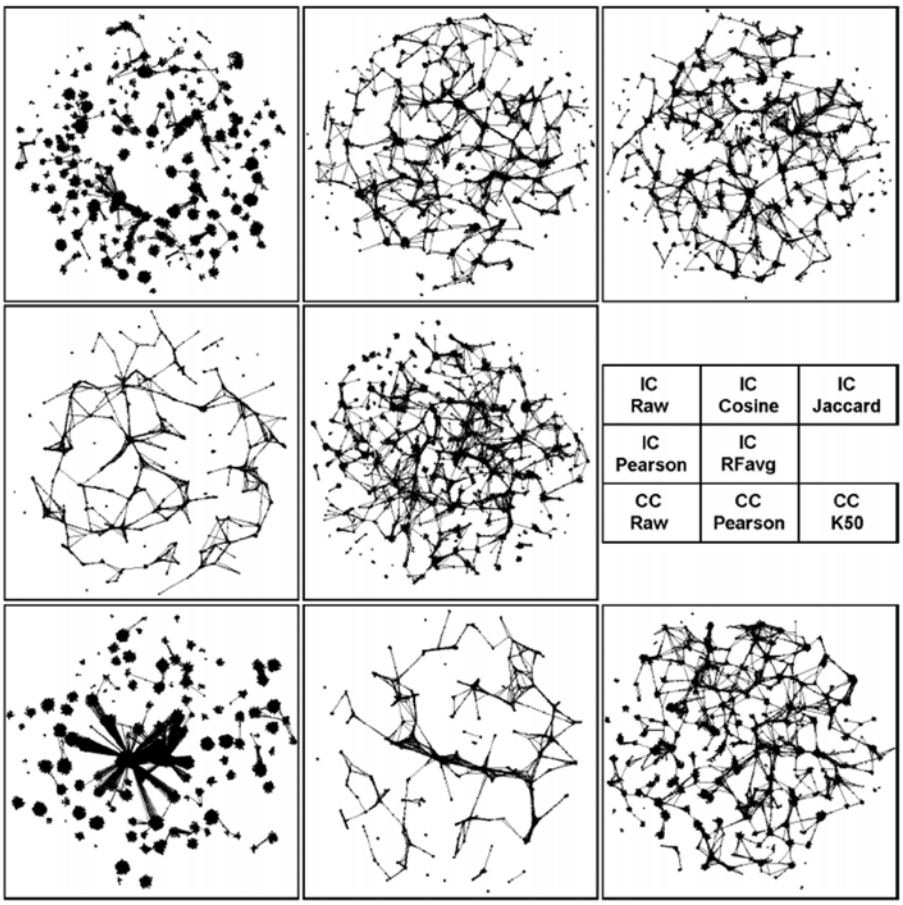


Information Visualization MOOC

Unit 4 – "What": Topical Data

Comparison of Text- and Linkage-Based Approaches

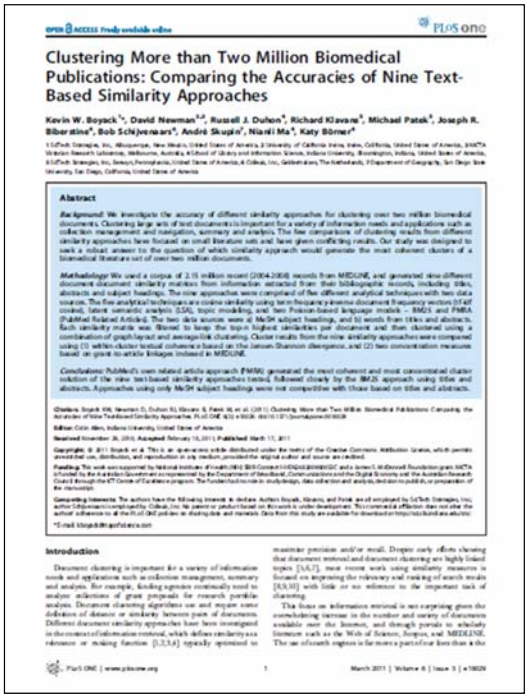
Relevant Research Disciplines:
Scientometrics, Computer Science

Comparing the Accuracy of Text-Based Similarity Measures Using Five Analytical Techniques

Example: document-document relatedness

- Cosine similarity using term frequency-inverse document frequency vectors (tf-idf cosine)
- Latent semantic analysis (LSA)
- Topic modeling
- Two Poisson-based language models:
  - BM25
  - PMRA (PubMed Related Articles).

Boyack, Kevin W., David Newman, Russell Jackson Duhon, Richard Klavans, Michael Patek, Joseph R. Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, and Katy Börner. 2011. "Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches." *PLoS ONE* 6 (3): 1-11.



PLoS one

Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches

Kevin W. Boyack, David Newman, Russell J. Duhon, Richard Klavans, Michael Patek, Joseph R. Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, Katy Börner

Abstract

Introduction

http://ivl.cns.iu.edu/km/pub/2011-boyack-clustering-more-plosone.pdf

Table 1. Listing of text-based similarity approaches and locations where the similarity calculations

Similarity approach	Data source	
	MeSH terms	Title/abstract words
tf-idf cosine	tf-idf MeSH (Indiana U.)	tf-idf TA (Indiana U.)
Latent semantic analysis	LSA MeSH (Indiana U.)	LSA TA (Indiana U.)
Topic modeling		Topics TA (UC Irvine)
Self-organizing map	SOM MeSH (SDSU/Indiana U.)	
Poisson-based	BM25 MeSH (Collexis)	BM25 TA (Collexis) PMRA (UC Irvine/SciTech)

doi:10.1371/journal.pone.0018029.t001

Table 2. Characteristics of the cluster solutions for the nine similarity approaches.

Approach	# Articles covered	% Coverage	# Clusters	Max Cluster Size
tf-idf MeSH	2,062,642	95.77%	24,708	1517
LSA MeSH	2,115,440	98.22%	25,287	1021
BM25 MeSH	2,011,339	93.39%	26,864	1015
SOM MeSH	2,153,169	99.97%	29,941	3576
tf-idf TA	1,796,349	83.41%	21,388	657
LSA TA	1,958,125	90.92%	23,831	1827
BM25 TA	2,022,694	93.91%	28,858	764
Topics TA	2,033,221	94.40%	24,163	1422
PMRA	2,029,564	94.23%	28,963	921

doi:10.1371/journal.pone.0018029.t002

Table 3. Summary of concentration results for the nine similarity approaches.

Approach	Herfindahl	Max(F1)	Pr80
tf-idf MeSH	0.1631	0.3790	0.2216
LSA MeSH	0.1124	0.3662	0.2127
BM25 MeSH	0.1570	0.3791	0.2167
SOM MeSH	0.1106	0.3796	0.2203
tf-idf TA	0.1299	0.3344	0.1571
LSA TA	0.1255	0.3646	0.2003
BM25 TA	0.2393	0.4281	0.2578
Topics TA	0.1584	0.4011	0.2379
PMRA	0.2410	0.4350	0.2637

doi:10.1371/journal.pone.0018029.t003

Comparing Eight Linkage-Based Similarity Measures

Example: journal-journal relatedness

Boyack, Kevin W., Richard Klavans, and Katy Börner. 2005."Mapping the Backbone of Science." *Scientometrics* 64 (3): 351-374.

Mapping the backbone of science

KEVIN W. BOYACK,[a] RICHARD KLAVANS,[b] KATY BÖRNER[c]

Introduction

Citation Similarity Measures

Five **inter-citation frequencies**:

one unnormalized measure:

- raw frequency (IC-Raw)

four normalized measures:

- Cosine (IC-Cosine)
- Jaccard (IC-Jaccard)
- Pearson's r (IC-Pearson)
- average relatedness factor of Pudovkin and Garfield25 (IC-RFavg)

Three **co-citation frequencies:**

one unnormalized measure

- raw frequency (CC-Raw)

two normalized measures:

- vector-based Pearson's r (CC-Pearson)
- normalized frequency measure K50 (CC-K50)

IC-Raw $\quad RAW_{i,j} = RAW_{j,i} = C_{i,j} + C_{j,i}$,

IC-Cosine $\quad COS_{i,j} = COS_{j,i} = \dfrac{\left(RAW_{i,j}\right)}{\sqrt{\sum_{k=1}^{n} C_{i,k} \sum_{k=1}^{n} C_{j,k}}}$,

IC-Jaccard $\quad JAC_{i,j} = JAC_{j,i} = \dfrac{\left(RAW_{i,j}\right)}{\sum_{k=1}^{n} C_{i,k} + \sum_{k=1}^{n} C_{j,k} - \left(RAW_{i,j}\right)}$,

IC-Pearson $\quad r_{i,j} = \dfrac{\sum_{k=1}^{n} \left(RAW_{i,k} - \overline{RAW_i}\right)\left(RAW_{j,k} - \overline{RAW_j}\right)}{\sqrt{\sum_{k=1}^{n} \left(RAW_{i,k} - \overline{RAW_i}\right)^2 \sum_{k=1}^{n} \left(RAW_{j,k} - \overline{RAW_j}\right)^2}}$,

where $\overline{RAW_i} = \dfrac{1}{n} \sum_{k=1}^{n} RAW_{i,k}$, $k \neq i$,

IC-RFavg $\quad RFA_{i,j} = RFA_{j,i} = \left(RF_{i,j} + RF_{j,i}\right) / 2$,

where $RF_{i,j} = 10^6 * C_{i,j} \Big/ \left(N_j \sum_{k=1}^{n} C_{i,k} \right)$.

CC-Raw $\quad F_{i,j}$,

CC-Pearson $\quad r_{i,j} = \dfrac{\sum_{k=1}^{n} \left(F_{i,k} - \overline{F_i}\right)\left(F_{j,k} - \overline{F_j}\right)}{\sqrt{\sum_{k=1}^{n} \left(F_{i,k} - \overline{F_i}\right)^2 \sum_{k=1}^{n} \left(F_{j,k} - \overline{F_j}\right)^2}}$,

where $\overline{F_i} = \dfrac{1}{n} \sum_{k=1}^{n} F_{i,k}$, $k \neq i$,

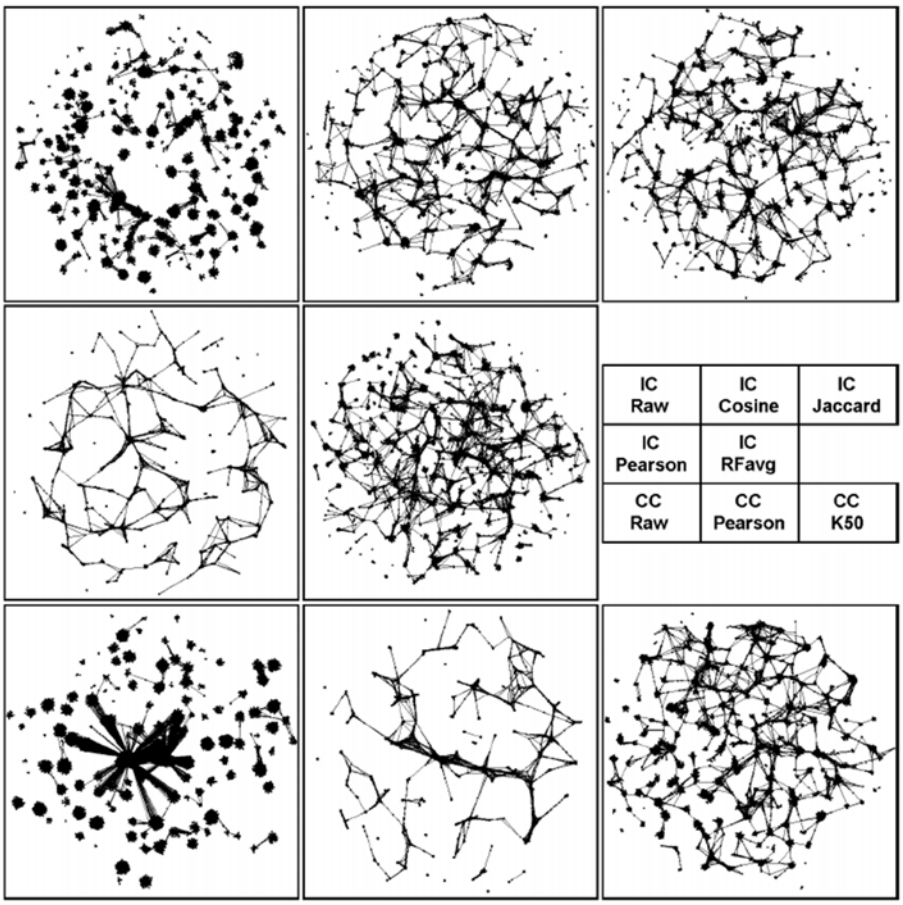
CC-K50 $\quad K50_{i,j} = K50_{j,i} = \max\left[\dfrac{\left(F_{i,j} - E_{i,j}\right)}{\sqrt{S_i S_j}} , \dfrac{\left(F_{j,i} - E_{j,i}\right)}{\sqrt{S_i S_j}} \right]$,

where the expected value of the cosine $E_{i,j} = \dfrac{S_i S_j}{(SS - S_i)}$,

$S_i = \sum_{j=1}^{n} F_{i,j}$, $j \neq i$,

and $SS = \sum_{i=1}^{n} S_i$.

In all three co-citation measures $F_{i,j}$ is the frequency of co-occurrences of journal i and journal j in reference documents (from the combined reference lists of the file year



Validation uses the ISI journal classifications to evaluate journal similarity measures and the corresponding maps. Journals in the same cluster of a journal mapping should have the same ISI category assignments.

The method of Gibbons and Roth (2002) requires a clustering of each of the maps. K-means clustering was applied (other clustering technique are possible).

Gibbons, Francis D., and Frederick P. Roth. 2002. “Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation.” *Genome Research* 12:1574-1581.

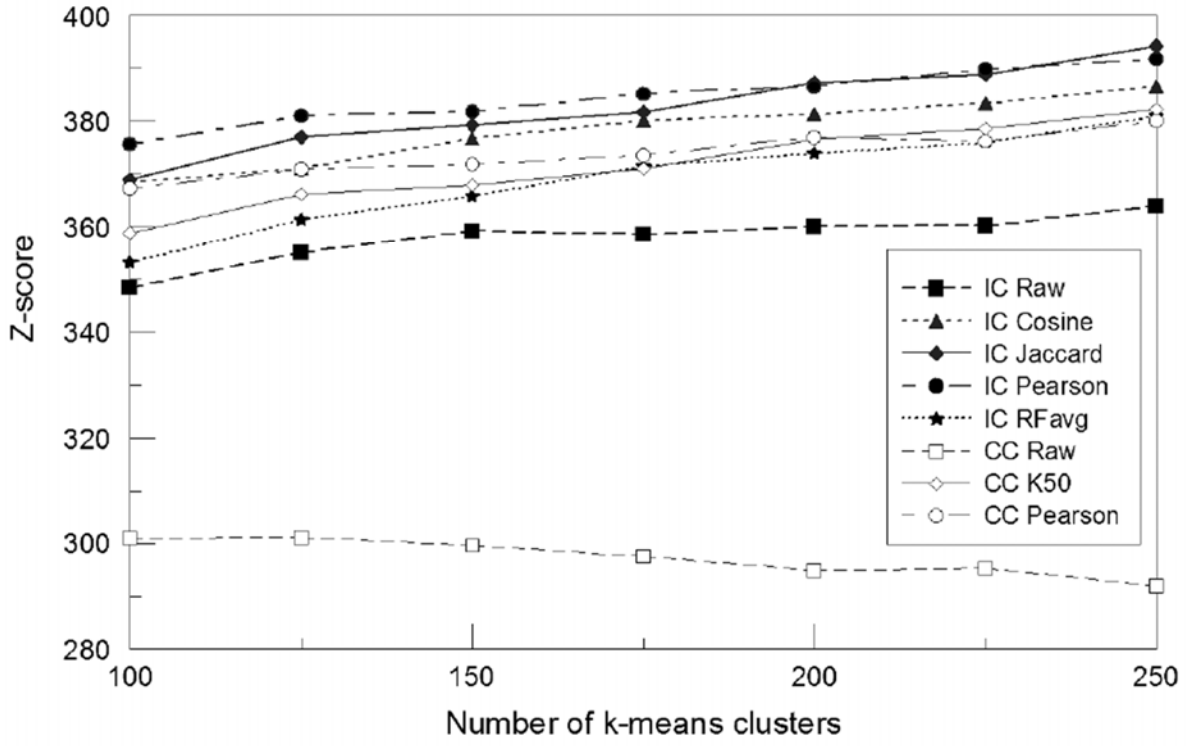


Table 1. Summary of validation results for maps based on eight similarity measures.

Measure	Local accuracy @ 95% coverage[1]	Scalability[1]	Z-score for 200 clusters	Clustering (qualitative)
IC-Raw	60.1%	High	360.0	Too few, loose
IC-Cosine	80.2%	High	381.3	Good balance
IC-Jaccard	79.5%	High	387.1	Good balance
IC-Pearson	71.7%	Low	386.5	Too tight
IC-RFavg	80.2%	High	373.3	Good balance
CC-Raw	25.6%	High	294.9	Too few, loose
CC-Pearson	65.3%	Low	377.0	Too tight
CC-K50	71.4%	High	376.6	Good balance

Comparing Eight Linkage-Based Similarity Measures

Example: journal-journal relatedness

Boyack, Kevin W., Richard Klavans, and Katy Börner. 2005. "Mapping the Backbone of Science." *Scientometrics* 64 (3): 351-374.

Acknowledgments

We would like to thank Miguel Lara and his colleagues at the Center for Innovative Teaching and Learning, University Information Technology Services at Indiana University, Bloomington.

The tool development work is supported in part by the Cyberinfrastructure for Network Science Center (http://cns.iu.edu) and Indiana University, the National Science Foundation under Grants No. SBE-0738111 and IIS-0513650, the US Department of Agriculture, the National Institutes of Health, and the James S. McDonnell Foundation.

Many visualizations used in the course come from the *Places & Spaces: Mapping Science* exhibit, online at http://scimaps.org, and from the *Atlas of Science: Visualizing What We Know*, MIT Press (2010).