An Exploratory Study on Co-word Network Simulation

Wanying Zhao¹, Jin Mao¹ and Kun Lu²

¹School of Information Management, Wuhan University
²School of Library and Information Studies, University of Oklahoma
Co-word networks
Motivation

• Co-word analysis, network-based metrics

• Empirical studies
  • Collect data from a handful disciplines from Web of Science
  • Observe the relationship among different metrics on empirical networks

• Problems
  • Small number of co-word networks
  • Empirical observation is posterior, little predictive power
  • Sample may not be representative, little generalizability
  • Little understanding of the innate mechanism of co-word network
Simulation method

- Consider the generative process of a co-word network
- Infinite number of samples can be generated once the process is established
- Parameters are adjustable to observe the behavior of co-word network under different situations
- Provide insights into the innate mechanisms of co-word networks
Generative process

M times for a collection with M papers

Generate a paper \((p_i)\)

Keyword_1
Keyword_2
...
Keyword_\mu

\[ P(\mu) \sim \]

For each keyword k:

\[ P = 1 - \alpha \]
From existing keywords

\[ P = \alpha \]
A new keyword

Random selection: \( P(k) = \frac{1}{n} \)

Preferential attachment:
\[ P(k) = \frac{n_k}{\sum_j n_j} \]
Research design

• Three empirical datasets
  • LIS, Sociology and Physics, Fluids & Plasma
  • Top 20 journals in WoS
  • Jan 2006 – Dec 2015

• Simulate co-word networks with comparable sizes as the empirical datasets.

• Compare simulated co-word networks with empirical networks.
Results – Keyword Frequency Distribution
Network comparison
Findings

• PA networks are more similar to empirical networks than RS networks

• RS creates many more edges than empirical data

• The rate of generating new keywords needs to be domain-dependent
Limitations & Future work

• Only a few factors are considered when simulating co-word networks
  • Damping factor controls creation of new keywords
  • PA, RS for keyword selection

• The simulation does not consider the maturity of a field, or other field dependent factors

• Current study only consider KeywordPlus field. Author keywords, title words could also be studied.
Thank you!