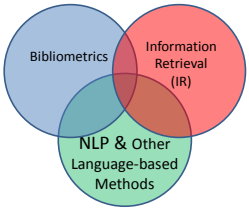
Bibliometrics, Information Retrieval & Natural Language Processing:
Natural Synergies to Support Digital Library Research

Dietmar Wolfram
University of Wisconsin-Milwaukee

BIRNDL 2016

Overview



Introduction

- The intersection of two key areas of information science offers many areas for research
- Recent BIR workshops demonstrate growing interest in the synergies between the two

Introduction

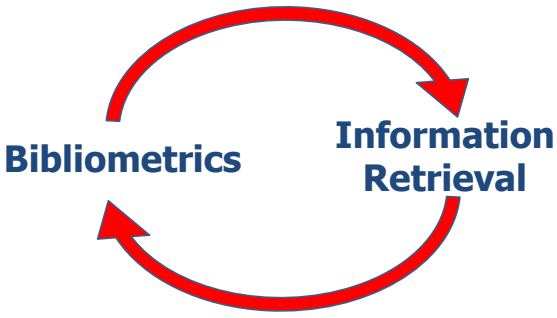
- Language-based methods have greatly benefitted IR and bibliometrics research
  - Natural Language Processing (NLP)
  - Text mining
  - Topic modeling
- Digital libraries (e.g., full text bib. records, heterogeneous collections) represent an ideal environment to study the intersection

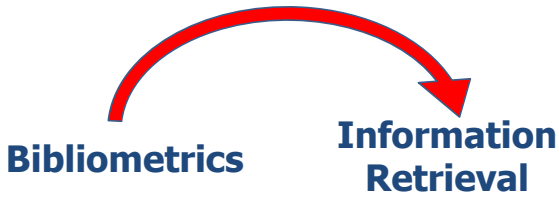
Language-based Methods & IR

- Beneficial for
  1. Content representation (NLP)
  2. Contending with large datasets & higher computational overhead (latent semantic analysis, topic modeling)
  3. More intuitive interface for users (NLP)

Language-based Methods & -metrics Research

- Citations & collaborations form the foundation of traditional comparative analysis
- Downside: No link ⇒ No relationship
- Language can expand relationship possibilities
  - Term co-occurrence
  - Topic modeling
  - Identifying hidden patterns with text mining

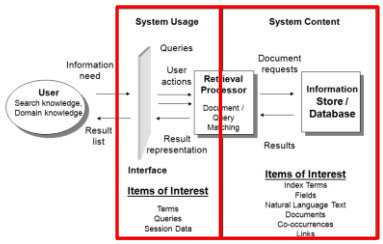




Areas of Application

- **Modeling IR processes**
  - System indexing & retrieval
  - IR system simulation
- **IR & allied system design & evaluation**
  - Using graph-based approaches / link analysis (co-authorship, citations, hyperlinks)
    - Ranking results
    - Supporting browsing & expanding results

IR Processes & Associated Data

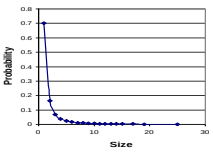


Adapted from Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, CT: Libraries Unlimited.

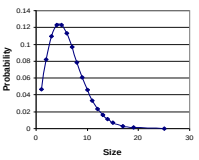
IR System Content Regularities

- Units: words/terms, fields, links, documents
- Indexing exhaustivity/specificity distributions
- Term co-occurrence relationships
- Growth of indexes and databases
- Persistence of documents

Observed Patterns in Content & Use Frequency

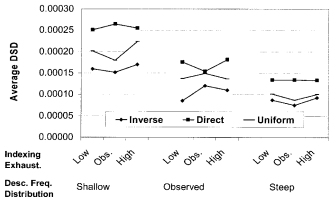


"Zipfian" or "Lotkaian" (Power Law)
Mode = 1, sometimes 0



"Unimodal"
Mode > 1

Effects of Indexing Decisions on Document Spaces

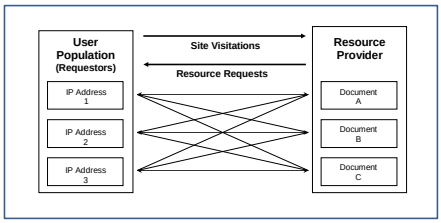


Wolfram, D., & Zhang, J. (2008). The influence of indexing practices and term weighting algorithms on document spaces. *Journal of the American Society for Information Science and Technology*, 59(1), 3-11.

IR System Usage

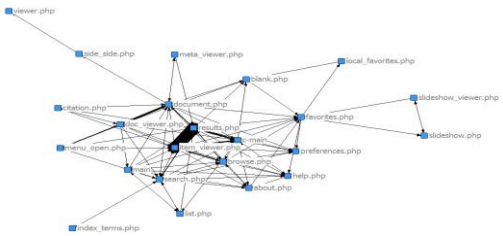
- Content Use
  - Website visitation
  - Document requests
- User search characteristics
  - Terms
  - Queries
  - Sessions (search and browsing actions)

Relationship Between Resources and Usage



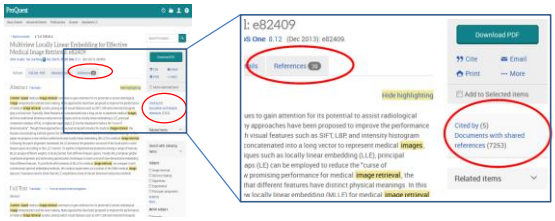
Ajiferuke, I., Wolfram, D., & Xie, H. (2004). Modelling website visitation and resource usage characteristics by IP address data. In H. Julien & S. Thompson (Eds.) *CAIS/ACSI 2004 - Access to Information: Technologies, Skills, and Socio-Political Context.*

Search Action Relationships



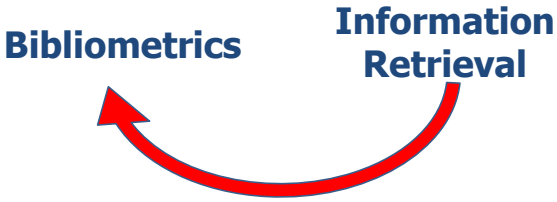
Han, H.J., Joo, S., & Wolfram, D. (2014). Using transaction logs to better understand user search session patterns in an image-based digital library. *Journal of the Korean Biblia Society for Library and Information Science*.

Linking Citing & Cited Documents



Ranking Documents

- HITS (Kleinberg, 1997)
- PageRank (Page et al., 1999)
- Hw-rank (Bar-Ilan & Levene, 2015)
- Bradfordizing & author centrality (Mutschke & Mayr, 2015)
- Article-level Eigenfactor (Wesley-Smith, Bergstrom, & West, 2016)



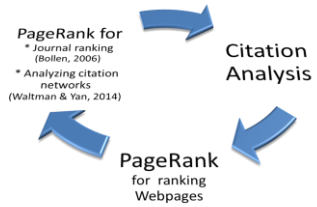
Reciprocal Contributions

- With growing datasets, new ways to store, process and display data are needed
- IR frameworks provide tools & approaches for -metrics researchers
  - Database design for bibliographic datasets
    - Relational & graph-based DBMSs, IR software & toolkits
  - Application of vector space & probabilistic IR models to compare data

Some Examples

- White (2007) – applied IR measures of term weighting (tf*idf) to bibliometric data
- Applications of Web link analysis
  - Research by Thelwall, Vaughan (many examples)
  - Use of PageRank for bibliometric ranking

PageRank Comes Full Circle



Using Language-based Relationships to Complement Link-based Relationships

Language expands studied relationships

1. Co-word analysis / Term co-occurrence
2. Topic modeling
3. Text mining

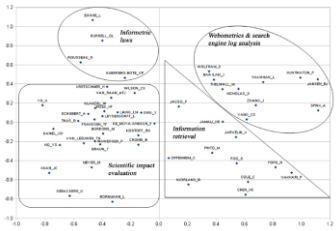
1) Co-word Analysis

- Longstanding use in metrics research (e.g., Braam & Moed, 1991; Ding, Chowdhury & Foo, 1997)
- Simple to use
- Independence assumption limitations
- IR matching methods can be used

2) Topic Modeling

- Applications of topic modeling
  - Tang et al. (2008) – applied Latent Dirichlet Allocation to academic search
  - Lu & Wolfram (2012) – compared author research similarity using topic modeling, co-authorship & co-citation
  - Ding & Song (2014) – measuring scholarly impact

Author-Topic Modeling for Author Research Relatedness



An A-T model produced more coherent groupings of prolific authors in information science than co-citation analysis

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based and author co-citation approaches. *Journal of the American Society for Information Science and Technology*, 63(10), 1973-1986.

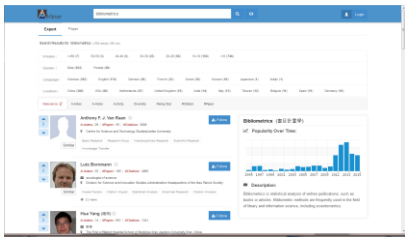
3) Text Mining

- Can be combined with bibliometric methods
  - Citation mining for user research profiling (Kostoff et al., 2001)
  - Clustering of scientific fields (Janssens, 2007)
  - Knowledge structure of bioinformatics (Song & Kim, 2013)
- Text mining techniques are integrated into some bibliometric mapping software, including
  - VOSviewer - http://www.vosviewer.com/
  - CiteSpace - http://cluster.cis.drexel.edu/~cchen/citespace/

Bibliometric-Enhanced Prototype & System Examples

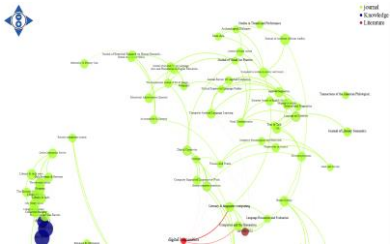
- I^3R (Croft & Thompson, 1987)
- Bibliometric Information Retrieval System (BIRS) (Ding et al., 2001)
- BibNetMiner (Sun et al., 2007)
- Aminer (Tang et al., 2008)
- Ariadne context explorer (Koopman et al., 2015)

Aminer



Search results based on bibliometric networks (aminer.org)

Ariadne



Future Directions

- Complexities of bibliometric datasets lend themselves to IR techniques
  - Resulting "big data" require data and text processing or mining techniques to identify overt & hidden patterns
- Topic modeling and other text-based methods show great promise in providing complementary approaches to citation & co-authorship data
  - Computational overhead to train models is still high
- Need for better evaluation methods for visualization outcomes

For More Information

- BIR Workshop Proceedings
  - 2014 - Mayr, Scharnhorst, Larsen, Schaer, & Mutschke
  - 2015 - Mayr, Frommholz, Scharnhorst, & Mutschke
  - 2016 - Mayr, Frommholz, & Cabanac
- Wolfram, D. (2015). The symbiotic relationship between information retrieval and informetrics. *Scientometrics*, 102(3), 2201-2214.
- Ding, Y., Rousseau, R., & Wolfram, D. (Eds.). (2014). *Measuring scholarly impact: Methods and practice*. Berlin: Springer.
- Wolfram, D. (2003). *Applied informetrics for information retrieval research.* Libraries Unlimited.