Introduction to Bibliometrics

Applied Bibliometrics
KAN Min-Yen
What is Bibliometrics?

○ Statistical and other forms of quantitative analysis

○ Used to discover and chart the growth patterns of information
  ○ Production
  ○ Use
Outline

- What is bibliometrics?
- Bibliometric laws
- Properties of information and its production
Properties of Academic Literature

- Growth
- Fragmentation
- Obsolescence
- Linkage
Growth

- Exponential rate for several centuries: “information overload”
- 1st known scientific journal: ~1600
- Today:
  - LINC has about 15,000 in all libraries

Factors:
- Ease of publication
- Ease of use and increased availability
- Known reputation
Zipf-Yule-Pareto Law

\[ P_n \approx \frac{1}{n^a} \]

where \( P_n \) is the frequency of occurrence of the \( n^{th} \) ranked item and \( a \approx 1 \).

“The probability of occurrence of a value of some variable starts high and tapers off. Thus, a few values occur very often while many others occur rarely.”

- Pareto – for land ownership in the 1800’s
- Zipf – for word frequency
- Also known as the 80/20 rule and as Zipf-Mandelbrot
- Used to measure of citations per paper:
  \# of papers cited \( n \) times is about \( 1/n^a \) of those being cited once, where \( a \approx 1 \)
Random processes and Zipfian behavior

- Some random processes can also result in Zipfian behavior:
  - At the beginning there is one “seminal” paper.
  - Every sequential paper makes at most ten citations (or cites all preceding papers if their number does not exceed ten).
  - All preceding papers have an equal probability to be cited.

- Result: A Zipfian curve, with $a \approx 1$.
  What’s your conclusion?
Lotka’s Law

The number of authors making $n$ contributions is about $1/n^a$ of those making one contribution, where $a \approx 2$.

Implications:

- A small number of authors produce large number of papers, e.g., 10% of authors produce half of literature in a field
- Those who achieve success in writing papers are likely continue having it
Lotka’s Law in Action

White and McCain’s dataset (98): 14 K papers, 190 K citations

Y = 6518X^{2.3}
1981-2002 papers

Y = 625X^{3.0}
year 2000 papers
Bradford’s Law of Scattering

Journals in a field can be divided into three parts, each with about one-third of all articles:

1) a core of a few journals,
2) a second zone, with more journals, and
3) a third zone, with the bulk of journals.

The number of journals is 1:n:n^2

To think about: Why is this true?
Fragmentation

- Influenced by scientific method
  - Information is continuous, but discretized into standard chunks
    (e.g., conference papers, journal article, surveys, texts, Ph.D. thesis)

- One paper reports one experiment
- Scientists aim to publish in diverse places
Fragmentation

- Motivation from academia
  - The “popularity contest”
  - Getting others to use your intellectual property and credit you with it
    - Spread your knowledge wide across disciplines
  - Academic yardstick for tenure (and for hiring)
    - The more the better – fragment your results
    - The higher quality the better – chase best journals

To think about: what is fragmentation’s relation to the aforementioned bibliometric laws?
Obsolescence

Literature gets outdated fast!
- ½ references < 8 yrs. Chemistry
- ½ references < 5 yrs. Physics

- Textbooks out dated when published
- Practical implications in the digital library
- What about computer science?

To think about: Is it really outdated-ness that is measured or something else?
ISI Impact Factor

\( A \) = total cites in 1992
\( B \) = 1992 cites to articles published in 1990-91 (this is a subset of \( A \))
\( C \) = number of articles published in 1990-91
\( D \) = \( B/C \) = 1992 impact factor
Half Life Decay in Action

The half-life curve is getting **shorter**: What factors are at work here? Is this a good or bad thing?
Expected Citation Rates

- From a large sample can calculate expected rates of citations
  - For journals vs. conferences
  - For specific journals vs. other ones

- Can find a researcher’s productivities against this specific rate
  - Basis for promotion

To think about: what types of papers are cited most often? (Hint: what types of papers dominate the top ten in Citeseer?)


3. 2668 J R Quinlan. *Induction of decision trees*. Machine Learning, , 0.


Linkage

- Citations in scientific papers are important:
  - Demonstrate awareness of background
  - Prior work being built upon
  - Substantiate claims
  - Contrast to competing work

Any other reasons?

One of the main reasons # of citations by themselves not a good rationale for evaluation.
Non-trivial to unify citations

Citations have different styles:

- Citeseer tried edit distance, structured field recognition
  - Settled on word (unigram) + section n-gram matching after normalization
  - More work to be done here: OpCit GPL code

Non-trivial even for the web: Think URL redirects, domain names
Computational Analysis of Links

If we know what type of citations/links exist, that can help:

- In scientific articles:
  - In calculating impact
  - In relevance judgment (browsing → survey paper)
  - Checking whether paper author’s are informed

- In DL item retrieval:
  - In classifying items pointed by a link
  - In calculating an item’s importance (removal of self-citations)
Calculating citation types

- Teufel (00): creates Rhetorical Document Profiles
  - Capitalizes on fixed structure and argumentative goals in scientific articles (e.g. Related Work)
  - Uses discourse cue phrases and position of citation to classify (e.g., In contrast to [1], we ...) a zone
Using link text for classification

- The link text that describes a page in another page can be used for classification.

- Amitay (98) extended this concept by ranking nearby text fragments using (among other things) positional information.
  
  - **xxxx**: .... .... .... ...
  
  - .... .... .... **xxx**, .... .... .... ...
  
  - .... **xxxx**[ .... ] [ .... ] [ .... ]
Ranking related papers in retrieval

- CiteSeer uses two forms of relatedness to recommend “related articles”:
  - TF × IDF
    - If above a threshold, report it
  - CC (Common Citation) × IDF
    - CC = Bibliographic Coupling
    - If two papers share a rare citation, this is more important than if they share a common one.
Citation Analysis

Deciding which (web sites, authors) are most prominent
Citation Analysis

- Despite shortcomings, still useful
- Citation links viewed as a DAG
- Incoming and outgoing links have different treatments

Analysis types

- **Co-citation** analysis – A and B both cited by C
- **Bibliographic coupling** – A and B both have similar citations (e.g., D)
Sociometric experiment types

- Ego-centered: focal person and its alters
  (Wasserman and Faust, pg. 53)

- Small World: how many actors a respondent is away from a target
Prominence

Consider a node prominent if its ties make it particularly visible to other nodes in the network
(adapted from WF, pg 172)

- Centrality – no distinction on incoming or outgoing edges (thus directionality doesn’t matter. How involved is the node in the graph.

- Prestige – “Status”. Ranking the prestige of nodes among other nodes. In degree counts towards prestige.
Centrality

- How central is a particular
  - Graph?
  - Node?
- Graph-wide measures assist in comparing graphs, subgraphs
Node Degree Centrality

- Degree (In + Out)
- Normalized Degree (In+Out/Possible)
  - What’s max possible?
- Variance of Degrees
Distance Centrality

- Closeness = minimal distance
- Sum of shortest paths should be minimal in a central graph
- (Jordan) Center = subset of nodes that have minimal sum distance to all nodes.

What about disconnected components?
Betweenness Centrality

- A node is central iff it lies between other nodes on their shortest path.
- If there is more than one shortest path,
  - Treat each with equal weight
  - Use some weighting scheme
    - Inverse of path length
References (besides readings)

- Bollen and Luce (02) *Evaluation of Digital Library Impact and User Communities by Analysis of Usage Patterns*
  [http://www.dlib.org/dlib/june02/bollen/06bollen.html](http://www.dlib.org/dlib/june02/bollen/06bollen.html)

- Kaplan and Nelson (00) *Determining the publication impact of a digital library*

- Wasserman and Faust (94) *Social Network Analysis* (on reserve)
Things to think about

- What’s the relationship between these three laws (Bradford, Zipf-Yule-Pareto and Lotka)?
- How would you define the three zones in Bradford’s law?
Pagerank and HITS*

Applied Bibliometrics
KAN Min-Yen
*Part of these lecture notes come from Manning, Raghavan and Schütze @ Stanford CS
Connectivity analysis

- **Idea:** mine hyperlink information in the Web

- **Assumptions:**
  - Links often connect related pages
  - A link between pages is a recommendation
    - “people vote with their links”
Query-independent ordering

- Using link counts as simple measures of popularity

- Two basic suggestions:
  - Undirected popularity:
    - in-links plus out-links (3+2=5)
  - Directed popularity:
    - number of its in-links (3)
Algorithm

1. Retrieve all pages meeting the text query (say \textit{venture capital}), perhaps by using Boolean model

2. Order these by link popularity (either variant on the previous page)

\textit{Exercise}: How do you spam each of the following heuristics so your page gets a high score?

- score = \# in-links plus \# out-links
- score = \# in-links
Pagerank scoring

- Imagine a browser doing a random walk on web pages:
  - Start at a random page
  - At each step, follow one of the $n$ links on that page, each with $1/n$ probability
- Do this repeatedly. Use the “long-term visit rate” as the page’s score
Not quite enough

- The web is full of dead ends.
  - What sites have dead ends?
  - Our random walk can get stuck.
Teleporting

- At each step, with probability 10%, teleport to a random web page

- With remaining probability (90%), follow a random link on the page
  - If a dead-end, stay put in this case

This is lay explanation of the “damping factor” \((1-a)\) in the rank propagation algorithm.
Result of teleporting

- Now we cannot get stuck locally
- There is a long-term rate at which any page is visited (not obvious, will show this)
  - How do we compute this visit rate?
A Markov chain consists of \( n \) states, plus an \( n \times n \) transition probability matrix \( P \).

- At each step, we are in exactly one of the states.
- For \( 1 \leq i, k \leq n \), the matrix entry \( P_{ik} \) tells us the probability of \( k \) being the next state, given we are currently in state \( i \).

\[
P_{ik} > 0 \quad \text{is OK.}
\]
Clearly, for all $i$, \[ \sum_{k=1}^{n} P_{ik} = 1. \]

Markov chains are abstractions of random walks.

Try this: Calculate the matrix $P_{ik}$ using a 10% probability of uniform teleportation.
Ergodic Markov chains

- A Markov chain is ergodic if
  - you have a path from any state to any other
  - you can be in any state at every time step, with non-zero probability

- With teleportation, our Markov chain is ergodic
Steady State

- For any ergodic Markov chain, there is a unique long-term visit rate for each state
  - Over a long period, we’ll visit each state in proportion to this rate
  - It doesn’t matter where we start
Probability vectors

- A probability (row) vector \( \mathbf{x} = (x_1, \ldots, x_n) \) tells us where the walk is at any point.
- E.g., \((000 \ldots 1 \ldots 000)\) means we’re in state \(i\).

More generally, the vector \( \mathbf{x} = (x_1, \ldots, x_n) \) means the walk is in state \(i\) with probability \(x_i\).

\[
\sum_{i=1}^{n} x_i = 1.
\]
Change in probability vector

- If the probability vector is \( \mathbf{x} = (x_1, \ldots, x_n) \) at this step, what is it at the next step?
- Recall that row \( i \) of the transition probability matrix \( \mathbf{P} \) tells us where we go next from state \( i \).
- So from \( \mathbf{x} \), our next state is distributed as \( \mathbf{xP} \).
Regardless of where we start, we eventually reach the steady state $a$
- Start with any distribution (say $x=(10...0)$)
- After one step, we’re at $xP$
- After two steps at $xP^2$, then $xP^3$ and so on.
- “Eventually” means for “large” $k$, $xP^k = a$

Algorithm: multiply $x$ by increasing powers of $P$ until the product looks stable
Pre-processing:
- Given graph of links, build matrix $P$
- From it compute $a$
- The pagerank $a_i$ is a scaled number between 0 and 1

Query processing:
- Retrieve pages meeting query
- Rank them by their pagerank
- Order is query-independent
Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find two sets of inter-related pages:
  - Hub pages are good lists of links on a subject.
    - e.g., “Bob’s list of cancer-related links.”
  - Authority pages occur recurrently on good hubs for the subject.

- Best suited for “broad topic” browsing queries rather than for known-item queries.

- Gets at a broader slice of common opinion.
Hubs and Authorities

- Thus, a good hub page for a topic *points* to many authoritative pages for that topic.

- A good authority page for a topic is *pointed* to by many good hubs for that topic.

- Circular definition - will turn this into an iterative computation.
Hubs and Authorities

Hubs

- AsiawEEK
- USNWR

Authorities

- NUS
- Tsinghua
- NTU
High-level scheme

- Extract from the web a base set of pages that could be good hubs or authorities.

- From these, identify a small set of top hub and authority pages → iterative algorithm
Base set

1. Given text query (say university), use a text index to get all pages containing university.
   - Call this the root set of pages

2. Add in any page that either:
   - points to a page in the root set, or
   - is pointed to by a page in the root set

3. Call this the base set
Assembling the base set

- Root set typically 200-1000 nodes.
- Base set may have up to 5000 nodes.
- How do you find the base set nodes?
  - Follow out-links by parsing root set pages.
  - Get in-links (and out-links) from a connectivity server.
Distilling hubs and authorities

1. Compute, for each page \( x \) in the base set, a **hub score** \( h(x) \) and an **authority score** \( a(x) \).

2. Initialize: for all \( x \), \( h(x) \leftarrow 1; a(x) \leftarrow 1 \);

3. Iteratively update all \( h(x) \), \( a(x) \);

4. After iterations:
   - highest \( h() \) scores are hubs
   - highest \( a() \) scores are authorities
Iterative update

- Repeat the following updates, for all $x$:

\[ h(x) \leftarrow \sum_{y\leftarrow x} a(y) \]

\[ a(x) \leftarrow \sum_{y\leftarrow x} h(y) \]
How many iterations?

- Relative values of scores will converge after a few iterations.
- We only require the relative order of the $h()$ and $a()$ scores - not their absolute values.
- In practice, ~5 iterations needed.
Things to think about

- Use *only* link analysis *after* base set assembled
  - iterative scoring is query-independent
- Iterative computation *after* text index retrieval - significant overhead
Things to think about

- How does the selection of the base set influence computation of H & As?
- Can we embed the computation of H & A during the standard VS retrieval algorithm?
- A pagerank score is a global score. Can there be a fusion between H&A (which are query sensitive) and pagerank? How would you do it?
- How do you relate CCIDF in Citeseer to Pagerank?