Chapter 4
Measuring Scholarly Impact

Abstract The ability to measure scholarly impact, ranging from individual scientists to an institution of researchers, is crucial to both research assessment and the advance of science itself. In this chapter, we summarize an array of fundamental and widely used concepts and computational methods for measuring scholarly impact as well as identifying more generic properties such as semantic relatedness, burstness, clumping, and centrality. Most of these common ideas are applicable to a wide variety of needs as long as we can identify the profound issues that are in common across distinct phenomena. Normalizations of metrics across scientific fields and the year of publication are discussed with concrete examples.

Introduction

Quantitative measures of scholarly impact are rooted in the measurement of information, uncertainty, proximity, novelty, rarity, connectivity, and other numerous indicators of significance. Some of these indicators and domain independent, whereas others are domain specific (Piffer 2012, Shwed and Bearman 2010).

The pragmatic question to many of these diverse metrics is whether and to what extent we may learn something useful or something new from the input or signals we receive, including text and other types of messages. The value of information is that it brings changes to our knowledge or our belief. This property can be seen as the fitness of information (Chen 2014). Information entropy (Shannon 1948) can be seen as a measure of the potential of what we may learn. Equivalently, it can be seen as a measure of the amount of uncertainty that can be resolved. For example, a dialogue between a physician and a patient reduces the initial entropy as various uncertainties are progressively narrowed down. An assumption that has been commonly seen in the reasoning of many information metrics is that we are more likely to learn something from a relatively rare event or word than from a common one. We expect to find creative ideas in areas that have not been well studied. We expect that boundary spanning may inspire extraordinary ideas.

C. Chen and M. Song, *Representing Scientific Knowledge*,
https://doi.org/10.1007/978-3-319-62543-0_4

Another strategy is to measure the importance or saliency of something by comparing it to a baseline. The strategy has been used in novelty detection, intrusion detection, burst detection, measuring rarity, and identifying outliers. The importance can be also measured in terms of connectivity, such as degree centrality, betweenness centrality, or eigenvector centrality.

Semantic similarities are often measured with reference to an existing ontological structure or a taxonomy. Ontology-based semantic similarity measures include path-based such as Wu and Palmer (1994), information-content-based such as Resnik (1995), feature-based such as Tversky (1977), and other types. WordNet is one of the most popular resources of choice in defining semantic similarity measures.

The ultimate utility of an indicator is to make easy and simple comparisons. Normalization is essential when we need to assure different measurements are comparable. The examples included in this chapter are representative and influential because they are designed based on some of the most fundamental principles that have been used in the design of a wide variety of indicators.

Information Metrics

Information Content

The concept of information content (IC) is used in a wide variety of many information metrics as well as on its own. More importantly, the principles behind the quantitative measure are applicable to a broad range of scenarios. The idea is to measure how much we can learn from a source of information. When we receive a message, the message may tell us nothing that we don't already know. On the other hand, a message may turn what we believe or what we think we know upside down!

Given a transmitted message *m* of information, its information content IC(m) is defined as the negative of the log likelihood of the message.

$$\mathbf{IC}(\mathbf{m}) = -\mathbf{log}_2 \boldsymbol{p}(\mathbf{m})$$

As shown in Fig. 4.1, as the probability of an event increases, the value of IC decreases. In particular, the IC value is the lowest for very common events, whereas the IC values are larger for rare events.

Shannon entropy quantifies the information in a message as something that would be new to the recipient of the message. If a message brings nothing new to the recipient, then the message does not carry any information as far as the recipient is concerned. Shannon entropy, or information entropy, is defined in terms of information content across all the possible events of a random variable X:

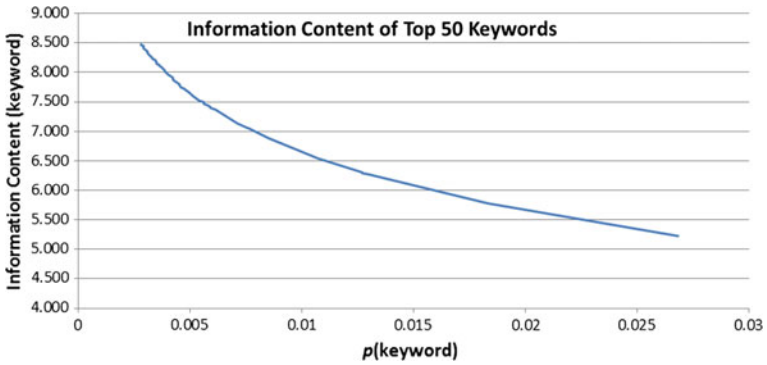

Fig. 4.1 Information content of top 50 most common keywords in 17,731 science mapping articles

$$\mathrm{H(X)} = -\sum_{i=1}^{n} p(x_i)\log(p(x_i)) = \sum_{i=1}^{n} p(x_i)\mathrm{IC}(x_i)$$

The value of the term $p(x_i) \times \mathrm{IC}(x_i)$ amplifies the small probability of a rare event with a large IC value but suppresses the large probability of a common event with a small IC value.

In general, we expect to learn a lot from rare events than a common event. Since we probably haven't experienced a rare event, it is likely that information associated with the rare event is new to our cognitive or belief system. Measuring interestingness or a degree of surprise often adopts similar principles.

Consider a dataset of science mapping publications we used in a systematic review (Chen 2017). The dataset contains 17,731 publications. These publications are indexed by 56,159 distinct keywords. From the relative frequency of a keyword, the information content of the keyword with respect to this particular dataset is calculated as $-log_2(f_k/f_N)$. We can use the following MySQL query to generate frequencies, relative frequencies, and information content of top 50 keywords to illustrate the concept of IC.

```
SELECT count(*), count(*)/56159, -log2(count(*)/56159), keyword
FROM keywords
WHERE project='sciencemapping17731' AND type!='sc'
GROUP BY keyword
ORDER BY count(*) DESC
LIMIT 50;
```

Table 4.1 list top 10 most common keywords. Keywords in this group have the lowest IC values because they occurred most frequently. Indeed, in the context of science mapping, keywords such as science, model, system, and impact do not tell us anything new, in part because they are field-independent words and in part they are almost applicable to any science mapping articles. Although keywords such as

Table 4.1 Information content of the most common keywords in a set of science mapping articles

Frequency (F)	Relative F (RF)	IC-log2(RF)	Keyword
1506	0.0268	5.221	Citation analysis
1026	0.0183	5.774	Science
724	0.0129	6.277	Model
716	0.0127	6.293	Information visualization
714	0.0127	6.297	System
603	0.0107	6.541	Time-domain analysis
477	0.0085	6.879	Impact
475	0.0085	6.885	Network
471	0.0084	6.898	Bibliometrics
425	0.0076	7.046	Journal

Table 4.2 Information contents of low-frequency keywords

Frequency	Relative frequency	Information content	Keyword
10	0.0002	12.4553	Latent semantic analysis
10	0.0002	12.4553	Explanation
10	0.0002	12.4553	Health policy
10	0.0002	12.4553	Randomized controlled trial
10	0.0002	12.4553	Nonlinear-system
5	0.0001	13.4553	Citation classic
5	0.0001	13.4553	Cross-section
5	0.0001	13.4553	Circuit modeling
5	0.0001	13.4553	Semantic network
5	0.0001	13.4553	Xylanase
1	0.0000	15.7773	Dysplastic nevus
1	0.0000	15.7773	Saturation time
1	0.0000	15.7773	Fiber-optics sensor
1	0.0000	15.7773	Ale metaanalysis
1	0.0000	15.7773	Terrorist

citation analysis and information visualization are field-dependent, their frequent occurrences serve little more than reinforce what we already know.

Table 4.2, generated by the MySQL query below, illustrates the information content scores of low-frequency keywords. In this dataset, keywords appear 10 times have a relevant frequency of 0.0002 and the information content of 12.4553. The ICs of keywords appeared for 5 times have even higher ICs of 13.4553. The ICs of 15.7773 are the highest possible for this particular dataset for keywords that appeared only once. The highest possible value depends on the total number of distinct keywords in the set.

```
SELECT *
FROM (
  SELECT
     count(*) AS c1,
     count(*)/56159 AS c2,
     -log2(count(*)/56159) AS c3,
     keyword AS c4
  FROM keywords
  WHERE project='sciencemapping17731' AND type!='sc'
  GROUP BY keyword
  ORDER BY count(*)
) AS a
WHERE c1=10
LIMIT 10;
```

Keywords such as latent semantic analysis and randomized controlled trial are less informative as keywords such as citation classic and semantic network, which in turn have lower information contents than dyplastic nevus, ale metaanalysis, and terrorist (Table 4.2).

Year-by-Year Labels of a Cluster

The evolution of a cluster may demonstrate various subthemes over time. CiteSpace supports a function to extract terms from each year's publications to characterize the nature of a cluster on a year-by-year basis (Fig. 4.2). The extraction is based on the LSI technique. We can select extracted terms from multiple dimensions of the latent semantic space so as to develop a good understanding of the major subthemes.

Selecting Noun Phrases with LSI

Figure 4.3 reveals further details of the biological terrorism cluster by extracting title terms from articles published in each year. Terms from the first two dimensions of the LSI latent semantic space are inspected here. Changes in these terms over time may give us additional insights into the evolution of the cluster.

The more detailed year-by-year terms are shown in Table 4.3. Top five terms for the largest three dimensions of the latent semantic space are listed for each year between 1999 and 2003, indicating that the cluster's research fronts started in 1999. The terms bioterrorism and biological terrorism appeared persistently in the first four years of the 5-year period. It seems that it reached its peak in 2001 because both the first and second dimensions are led by the semantically equivalent terms.

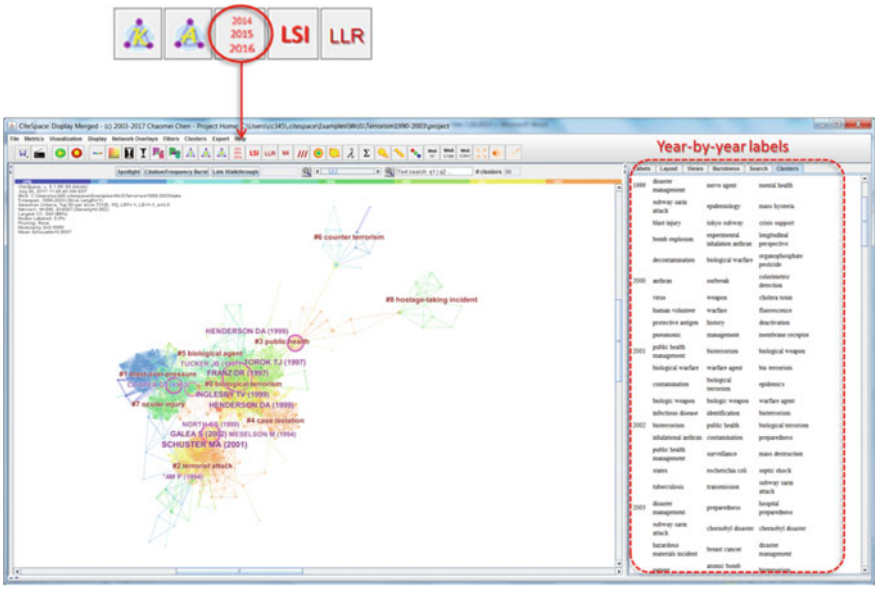

Fig. 4.2 Generating year-by-year labels of a cluster in CiteSpace

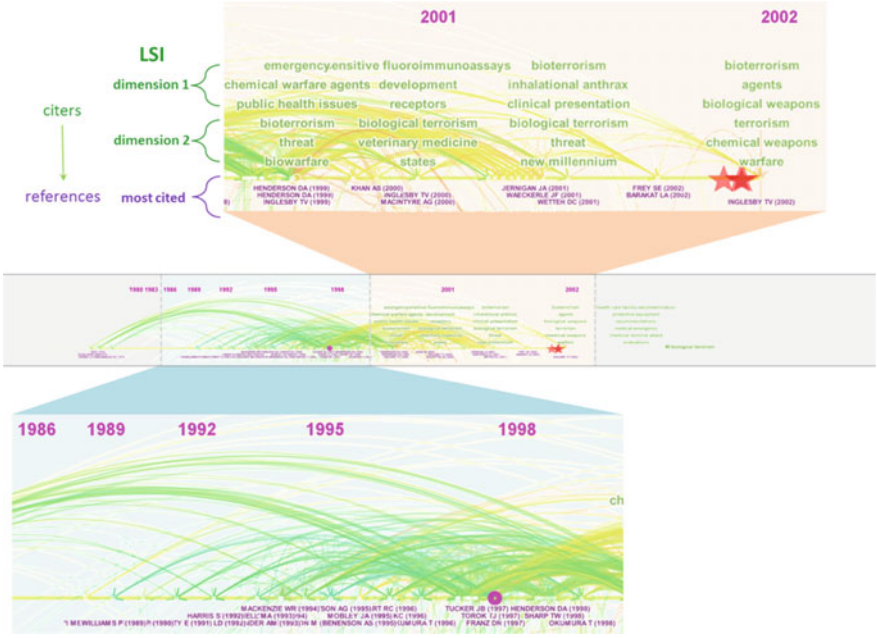

Fig. 4.3 Year-by-year labels of the biological terrorism cluster

Table 4.3 Year-by-year label terms of the biological terrorism cluster

Year	Dimension 1	Dimension 2	Dimension 3
1999	Emergency	**Bioterrorism**	Chemical
	Chemical warfare agents	Threat	Biological agents
	Public health issues	Biowarfare	Psychiatric aspects
	Hazmat	Emergency physicians	Domestic terrorism
	Anthrax	Reason	Emergency physicians
2000	Sensitive fluoroimmunoassays	**Biological terrorism**	Tetanus toxin
	Development	Veterinary medicine	Bind
	Receptors	States	Identification
	Using ganglioside-bearing liposomes	**Bioterrorism**	Novel
	Gangliosides	Tetanus toxin	Small molecule
2001	**Bioterrorism**	**Biological terrorism**	Emergency
	Inhalational anthrax	Threat	Ethics
	Clinical presentation	New millennium	Medical care
	Following **bioterrorism** exposure	Short-term safety experience	Chemical
	Surviving patients	Public health	Victims
2002	**Bioterrorism**	Terrorism	Food
	Agents	Chemical weapons	Thought
	Biological weapons	Warfare	Deployment locations
	Bacterial pathogens	Public health law	Vulnerability
	Terrorist attacks	Common goods	Terrorist attack
2003	Health care facility decontamination	Medical emergency	Report
	Protective equipment	Chemical terrorist attack	Drexel university emergency department
	Recommendations	Evaluations	Terrorism preparedness consensus panel
	Personnel	Teams	Radiation disasters
	Evaluations	Terrorism preparedness consensus panel	Children

Selecting Indexing Terms with LSI

Table 4.4 shows indexing terms extracted as the year-by-year labels for the biological terrorism cluster. Semantically equivalent terms such as biological warfare, bioterrorism, and biological terrorism appeared in the 1st, 3rd, 4th, and 5th years of the cluster. Terms such as disaster management, public health management, and management clearly identified the primary motivation of the research behind this cluster. Terms such as subway sarin attack, Tokyo subway, and Chernobyl disaster

Table 4.4 Year-by-year cluster labels extracted from indexing terms of the biological terrorism cluster

Year	Dimension 1	Dimension 2	Dimension 3
1999	Disaster management	Nerve agent	Mental health
	Subway sarin attack	Epidemiology	Mass hysteria
	Blast injury	Tokyo subway	Crisis support
	Bomb explosion	Experimental inhalation anthrax	Longitudinal perspective
	Decontamination	**Biological warfare**	Organophosphate pesticide
2000	Anthrax	Outbreak	Colorimetric detection
	Virus	Weapon	Cholera toxin
	Human volunteer	Warfare	Fluorescence
	Protective antigen	History	Deactivation
	Pneumonic	Management	Membrane receptor
2001	Public health management	**Bioterrorism**	Biological weapon
	Biological warfare	Warfare agent	**Bio terrorism**
	Contamination	**Biological terrorism**	Epidemics
	Biologic weapon	Biologic weapon	Warfare agent
	Infectious disease	Identification	**Bioterrorism**
2002	**Bioterrorism**	Public health	**Biological terrorism**
	Inhalational anthrax	Contamination	Preparedness
	Public health management	Surveillance	Mass destruction
	States	*Escherichia coli*	Septic shock
	Tuberculosis	Transmission	Subway sarin attack
2003	Disaster management	Preparedness	Hospital preparedness
	Subway sarin attack	Chernobyl disaster	Chernobyl disaster
	Hazardous materials incident	Breast cancer	Disaster management
	Patient	Atomic bomb survivor	**Bioterrorism**
	Breast cancer	Risk factor	Recommendation

indicate the influence of these attacks or disasters on research in bioterrorism over multiple years.

Semantic Relatedness

Two concepts are related if there is an incident or an event that involves both of them. A bank and a robber can be related by a bank robbery instance. Relatedness is a relation that connects two entities or abstract concepts. Association is commonly

used to describe a relationship. A semantic relation is defined between two entities. In natural language, a semantic relation is typically represented by a triple, namely, the subject, the object, and the relation. In the statement JOHN TEACHES CALCULUS, JOHN is the subject, CALCULUS is the object, and the verb TEACHES established the connection. JOHN is a teacher and CALCULUS is a course. At a higher level of abstraction, a teacher TEACHES a course. The two concepts of teacher and course are semantically related.

Scientific articles routinely include a section on related work. Authors often discuss previous studies that addressed the same problem in some ways, but they are not considered as similar studies. They are related to each other because they more or less addressed the same problem.

The similarity between two concepts implies that we are comparing the two concepts in terms one or more attributes. Two smartphones may be similar because of their appearance such as size or color or internal design such as apps or controls.

The concept of semantic similarity is typically defined based on an underlying ontology or taxonomy, where concepts are organized to reflect their semantic relations. Notable sources such as WordNet are widely used in related research.

Semantic relatedness between two concepts can be established in a given domain ontology. If the two concepts can be connected with a path in the ontological representation, then the semantic relatedness is evident.

Semantic similarity is a special case of semantic relatedness. Two semantically related concepts may not be semantically similar, whereas two semantically similar concepts must be semantically related. In the earlier example, a bank and a robber are semantically related, but it does not make much sense if we say that they are similar in terms of some attributes or aspects.

Resnik's Semantic Similarity

The most influential work on measuring semantic similarities is the work by Resnik (1995). His approach makes use of the IS-A semantic links in a taxonomy, namely the WordNet, and measure the semantic similarity based on the information content over the most relevant semantic structure. The results were very encouraging, with a correlation of 0.79 to the upper bound of 0.90 of human subjects.

Given a taxonomy of concepts, the semantic similarity between two nodes in the taxonomy can be estimated in many ways. Here we consider IS-A links only in the taxonomy. The most straightforward way is to measure the distance between the two concepts. The shorter the connecting path between them, the more similar the two concepts are. If there are multiple paths, the length of the shortest path should be used to represent the semantic similarity. In fact, this edge-counting approach was proposed by Rada and Bicknell (1989). However, each link in a taxonomy is usually considered to have a length of 1 unit. All the links have this property regardless which part of the taxonomy they belong to. In a taxonomy like the WordNet, the semantic strength of a link near to the top, i.e. the broadest possible

term may differ considerably from the semantic strength of a link near to the bottom of the taxonomy, where concepts are much more concrete and specific.

Intuitively, the edge-counting similarity is sensitive to the concepts' positions in the taxonomy. Such a sensitivity is not desirable because a similarity measure should not depend on additional factors. Resnik offered an alternative method to measure semantic similarity over a taxonomy of IS-A relations. His solution is based on the notion of information content. His approach also makes uses corpus-based statistics to estimate the probability of a concept. Connecting to the underlying data source makes it possible to use the same taxonomy with multiple contexts.

The semantic similarity between two concepts should reflect the extent to which they share information. In the context of an IS-A taxonomy, concepts are linked by IS-A relations. The extent to which two concepts share information is equivalent to finding a concept that subsumes both concepts. In WordNet, Coin subsumes both Nickel and Dime. The semantic similarity between Nickel and Dime is therefore reflected by the concept of Coin. Since Cash subsumes Coin, Cash indirectly subsumes both Nickel and Dime as well. Both Coin and Cash are called subsumers of Nickel and Dime. Which subsumer, Cash or Coin, makes the best candidate to represent the shared information content?

Coin is more specific than Cash. Coin has less irrelevant information than Cash. For example, Cash subsumes BILL as well as Coin. The information about BILL is irrelevant to the similarity between two Coins. Thus, the shared information content should be represented by the subsumer that has the lowest position on the taxonomy. The lower a concept on the taxonomy, the more specific it is.

The criteria discussed so far are applicable to the edge-counting method as well. The edge-counting method selects the shortest path that connects two concepts in question, for example, Nickel—Coin—DIME. If there is a longer path connecting the two concepts, then the longer path includes broader concepts rather than narrower concepts than the shortest path, for example, Nickel—Coin—Cash—Coin—Dime.

In order to avoid the unreliability issues with the edge-counting method, Resnik introduced probabilities of concepts in measuring semantic similarities. For each concept c in the underlying taxonomy, p(c) is the probability of encountering an instance of the concept. A concept positioned higher up in the taxonomy should have a higher probability than a concept positioned below it. If c1 IS-A c2 in the taxonomy, e.g. Dime IS-A Coin, then $p(c1) \leq p(c2)$. Thus, $p(\text{Dime}) \leq p(\text{Coin})$. The root concept r of the taxonomy should have $p(r) = 1$.

The information content IC of a concept c is: $-log_2\, p(c)$. Since the probability of the broadest concept is 1, the lowest value of information content is 0. All other values of information content would be positive. Theoretically, there is no upper limit.

The semantic similarity between concepts c1 and c2 is the information content shared by the two concepts, which is in turn represented by the information content of the concepts that subsume the two concepts in the taxonomy

$$sim(c_1, c_2) = \max_{c \in S(c1,c2)}(-\log p(c))$$

where S(c1, c2) is the set of concepts that subsume both c1 and c2. Since the probability of a concept on the taxonomy is a monotonic along the IS-A links, the information content of a parent concept is less than the information content of its child concept, e.g. I(CASH) = −log p(CASH) ≤ log p(COIN) = I(COIN). Thus, the concept that reaches the maximum information content must be the subsumer that has the lowest position in S(c1, c2), or equivalently, the most specific concept that subsumes c1 and c2.

Resnik (1995) estimated the probability of a concept based on the Brown Corpus of American English, which is a collection of 1 million words of various genres of text, including news articles and scientific fictions. The occurrences of a word are counted towards all its parent concepts as well as its own concept in the taxonomy because an occurrence of DIME is also an occurrence of COIN and that of CASH. The probability of a concept is then defined as the relative frequency of the corresponding noun to the total number of nouns in the corpus.

Resnik validated his information content-based semantic similarity measure based on the assumption that a good similarity measure should agree with similarity ratings made by human subjects. Computational similarity measures should be consistent with similarity ratings based on our intuitions. He replicated an experiment designed by Miller and Charles. In Miller and Charles' original experiment, 30 pairs of nouns were given to 38 undergraduate subjects to rate "similarity of meaning" on a scale from 0, which means no similarity, to 4, which means perfect synonymy. These nouns were selected based on a previous study so that various degrees of similarity are covered by the set. Resnik gave the same 30 pairs of nouns to 10 computer science students or postdocs at the University of Pennsylvania and used exactly the same instructions. The average rating for each pair provides an estimate of the semantic similarity of the pair as judged by human.

Resnik found a correlation of 0.96 between the mean ratings in his experiment and in Miller and Charles' one. In terms of correlations with human judgements in Miller and Charles' experiments, the new human ratings are the nearest (r = 0.9015), followed by the information content (r = 0.7911), then by probability (r = 0.6671), with the edge counting the lowest (r = 0.6645).

Resnik's work is influential. Researchers have developed a number of variations based on Resnik's original work.

Other Measures of Semantic Similarity

WordNet Similarity for Java (WS4J)[1] is a Java library developed by Hideki Shim when he was a doctoral student at Carnegie Mellon University. It implements

[1] http://code.google.com/p/ws4j/.

several algorithms to compute semantic relatedness or similarity algorithms based on semantic relations in WordNet. An online demo is available at http://ws4jdemo.appspot.com. It appears that the demo version is somewhat better than the Java library. The examples below are based on the online version.

One can enter two words to the WS4J Demo and if they are found in WordNet, then the demo will report eight types of similarity measures for the pair of words. For instance, we can enter dime and nickel to the WS4J demo interface (Fig. 4.4). Note that nickel has multiple meanings, or senses in WordNet. Its meaning as a coin is the second sense.

WS4J Demo reports the structural details for each similarity, including common subsumers of concepts in WordNet. Figure 4.5 illustrates information that can be reconstructed from WS4J's outputs. Information of the local structure is useful for understanding basic concepts used in this group of algorithms. For instance, the Lowest Common Subsumers (LCS) of dime and nickel is currency. The shortest path connecting dime and nickel has a length of 3. Both dime and nickel have the depths of 11. IC(c) is the information content of the concept c. Thus, the subsumer coin has a lower information content, IC(coin) of 9.0577, than that of dime, which has IC(dime) of 11.0726. In this example, using the third sense of the word nickel, nickle3 in WordNet, WUP(dime, nickel2) = 0.9091 and RES(dime, nickel2) = 9.0577.

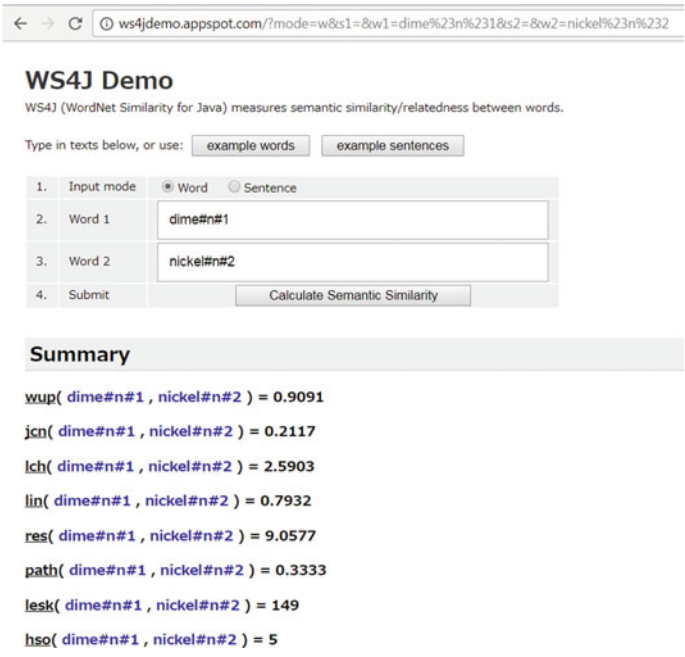

Fig. 4.4 WS4J Demo at http://ws4jdemo.appspot.com

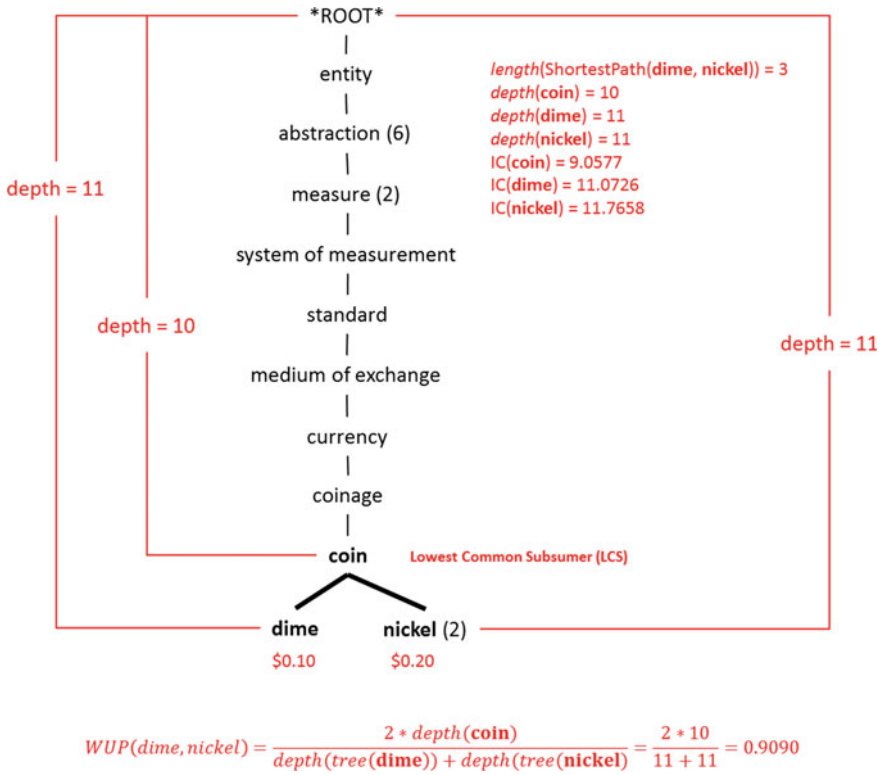

Fig. 4.5 The local structure of dime, nickel, and their LCS in WordNet and intermediate measures used in semantic similarity algorithms

Table 4.5 summarizes the algorithms for computing semantic similarities on WordNet. When applicable, we give sim(dime, nickel) as a concrete example to illustrate each algorithm.

Table 4.6 shows how various semantic similarity algorithms measure the semantic relatedness of 30 pairs of words and how they are correlated with ratings made by human subjects. The 30 pairs of words are the same set used by Miller and Charles (1991) in their experiment. They obtained similarity ratings from 38 human subjects on these pairs. Resnik duplicated the experiment in 1995 with 10 subjects. We consider the average rating from the Miller and Charles' experiment as the gold standard for the comparison. The comparison simply aims to see which algorithm behaves most like human raters.

Not surprisingly, human ratings in Resnik's experiment in 1995 have a strong correlation (r = 0.79) with human ratings obtained in Miller and Charles's experiment. This correlation is stronger than that from any of the computational algorithms. The algorithm that is the nearest to human ratings in Miller and Charles' experiment is the Resnik's similarity (RES), with a correlation of 0.61. RES is

Table 4.5 Semantic similarity algorithms with sim (dime, nickel) as an illustrative example

GS	Reference	Description
2904	Wu and Palmer (1994)	$WUP(s_1, s_2) = \frac{2*dLCS.d}{\min_{dlcs \in dLCS}(s1.d - dlcs.d) + \min_{dlcs \in dLCS}(s2.d - dlcs.d)}$ $WUP(dime, nickel) = \frac{2*10}{11+11} = 0.9090$ $wheredLCS(s_1, s_2) = \underset{lcs \in LCS}{\text{argmax}}(lcs.d)$ The Wu-Palmer similarity measures the semantic relatedness of two synsets s1 and s2 in WordNet with respect to the LCS—the least common subsumer of s1 and s2. For a synset s, s.d is its depth in WordNet The range of the WUP is [0, 1]
3146	Jiang and Conrath (1997)	$JCN(s_1, s_2) = \frac{1}{IC(s1)+IC(s2)-2*IC(LCS(s_1,s_2))}$ $JCN(dime, nickel) = \frac{1}{11.0726+11.7658-2*9.0577} = 0.2117$ The range of JCN is [0, +∞)
1891	Leacock and Chodorow (1998)	$LCH(s_1, s_2) = -\ln\left(\frac{length(LCS(s_1,s_2))}{2(MaxDepth(n))}\right)$ $LCH(dime, nickel) = -\ln\left(\frac{3}{2*20}\right) = 2.5903$ LCH is defined based on the shortest path between the two synsets and scale the path length by the maximum depth of the taxonomy The range of JCN is [0, +∞)
4312	Lin (1998)	$LIN(s_1, s_2) = \frac{2*IC(LCS(s_1,s_2))}{IC(s1)+IC(s2)}$ $LIN(dime, nickel) = \frac{2*9.0577}{11.0726+11.7658} = 0.7932$ Similar to JCN, but the range of LIN is scaled to [0, 1]
3602	Resnik (1995)	$RES(s_1, s_2) = IC(LCS(s_1, s_2)$ $RES(dime, nickel) = IC(coin) = 9.0577$ RES defined the similarity between two synsets to be the information content of their lowest super-ordinate (most specific common subsumer) The range of RES is [0, +∞)
PATH	Rada and Bicknell (1989)	$PATH(s_1, s_2) = \frac{1}{length(shortestpath(s_1,s_2))}$ $PATH(dime, nickel) = \frac{1}{3} = 0.3333$ PATH counts the number of nodes along the shortest path between the senses in the IS-A hierarchies of WordNet The range of Path is [0, +∞)
854	Banerjee and Pedersen (2002), Lesk (1986)	$LESK(s_1, s_2) = sum(dictionary\ definition\ overlaps)$ $LESK(dime, nickel) = 149.0$ LESK computes the relatedness of two words in terms of the extent to which their dictionary definitions overlap. Banerjee and Pedersen (2002) extended this notion to use WordNet as the dictionary for the word definitions The range of Path is [0, +∞)
1087	Hirst and St-Onge (1998)	$HSO(s_1, s_2) = 8 - distance - change\ Of\ Direction$ $HSO(dime, nickel) = 8 - 2 - 1 = 5.0$ HSO(s1, s2) = c − length(path(s1, s2)) − k * changes of directions (s1, s2) Links to be considered include 2 horizontal links, upward links, downward links The range of RES is [0,16]

The GS column is the citation count on Google Scholar as of July 21, 2017

Table 4.6 Comparing algorithms with Miller and Charles' (1991) experiment and Resnik's 1995 experiment

Pairs of words	Miller and Charles means	Resnik 1995	RES	WUP	LCH	LIN	PATH	HSO	JCN	LESK
Car, Automobile	3.92	8.04	7.00	1.00	3.69	1.00	1.00	16.00	12876699.50	9519.00
Gem, Jewel	3.84	14.93	2.49	0.63	1.61	0.24	0.13	0.00	0.06	8.00
Journey, Voyage	3.84	6.75	6.80	0.87	2.30	0.83	0.25	4.00	0.35	35.00
Boy, Lad	3.76	8.42	4.65	0.90	2.59	0.64	0.33	5.00	0.19	152.00
Coast, Shore	3.70	10.81	8.10	0.92	3.00	0.96	0.50	4.00	1.62	330.00
Asylum, Madhouse	3.61	15.67	3.94	0.67	1.61	0.00	0.13	0.00	0.00	6.00
Magician, Wizard	3.50	13.67	1.90	0.76	1.90	0.20	0.17	2.00	0.06	25.00
Midday, Noon	3.42	12.39	9.57	1.00	3.69	1.00	1.00	16.00	12876699.50	152.00
Furnace, Stove	3.11	1.71	2.49	0.50	1.12	0.23	0.08	5.00	0.06	190.00
Food, Fruit	3.08	5.01	0.61	0.38	1.29	0.10	0.09	0.00	0.09	130.00
Bird, Cock	3.05	9.31	0.61	0.29	0.92	0.00	0.06	2.00	0.00	17.00
Bird, Crane	2.97	9.31	1.82	0.64	1.49	0.00	0.11	0.00	0.00	12.00
Tool, Implement	2.95	6.08	6.31	0.94	3.00	0.91	0.50	4.00	0.85	509.00
Brother, Monk	2.82	2.97	1.90	0.70	1.61	0.21	0.13	0.00	0.07	29.00
Crane, Implement	1.68	2.97	1.37	0.53	1.39	0.00	0.10	0.00	0.00	7.00
Lad, Brother	1.66	2.94	1.90	0.73	1.74	0.24	0.14	0.00	0.08	14.00
Journey, Car	1.16	0.00	0.00	0.17	0.69	0.00	0.05	0.00	0.07	179.00
Monk, Oracle	1.10	2.97	1.90	0.70	1.61	0.18	0.13	0.00	0.06	30.00
Food, Rooster	0.89	1.01	0.61	0.29	0.92	0.08	0.06	0.00	0.07	18.00
Coast, Hill	0.87	6.23	6.14	0.71	2.08	0.73	0.20	3.00	0.22	123.00
Forest, Graveyard	0.84	0.00	0.00	0.24	1.05	0.00	0.07	0.00	0.06	20.00
Monk, Slave	0.55	2.97	1.90	0.80	2.08	0.20	0.20	3.00	0.07	74.00
Coast, Forest	0.42	0.00	0.00	0.29	1.29	0.00	0.09	0.00	0.06	59.00
Lad, Wizard	0.42	2.97	1.90	0.80	2.08	0.22	0.20	3.00	0.08	13.00

(continued)

Table 4.6 (continued)

Pairs of words	Miller and Charles means	Resnik 1995	RES	WUP	LCH	LIN	PATH	HSO	JCN	LESK
Chord, Smile	0.13	2.35	0.78	0.38	1.29	0.08	0.09	0.00	0.06	11.00
Glass, Magician	0.11	1.01	0.61	0.43	1.20	0.07	0.11	0.00	0.06	25.00
Noon, String	0.08	0.00	0.00	0.20	0.86	0.00	0.06	0.00	0.05	14.00
Rooster, Voyage	0.08	0.00	0.00	0.15	0.51	0.00	0.04	0.00	0.05	2.00
	1.00		**0.61**	**0.59**	**0.55**	**0.52**	**0.47**	**0.45**	**0.32**	**0.27**
		0.79	**0.58**	**0.58**	**0.54**	**0.40**	**0.44**	**0.37**	**0.29**	**0.11**

closely followed by WUP (r = 0.59). At the other end of the scale, JCN and LESK yielded the lowest correlations with Miller and Charles. If we use human ratings in Resnik's 1995 experiment, RES and WUP would have a tie (r = 0.58).

To our knowledge the largest gold standard of human ratings of similarity is the RG-65 test collection, containing similarity ratings of 65 pairs of words by 51 subjects on a scale of 0–4. The study was published in 1965 by Rubenstein and Goodenough.[2] Good enough!

The ACL wiki page lists a series of algorithms tested with the RG-65. Algorithms are ranked by Spearman and Pearson correlation coefficients. The highest correlation is achieved by an algorithm of Pilehvar and Navigli in 2015 with the Spearson correlation of 0.92 and Pearson correlation of 0.91. Several algorithms we have discussed earlier are included in the list, including HSO (0.813/0.732 by Spearson/Pearson correlations), JCN (0.804/0.731), LIN (0.788/0.834), and RES (0.731/0.800), and LSI (0.609/0.644).

In summary, as computational linguistics advances and a wide variety of resources become accessible, measuring semantic similarities has become increasingly powerful and reliable. For instance, estimating the probability of a word with the large pool of documents on Google is much more reliable than estimating it using a smaller collection of documents. The basic principles for estimating the semantic relatedness of a pair of words have fostered a large number of algorithms. Each of them has unique strengths.

Concentration

Burstness

The burstness of a variable X measures abrupt increases of the value of X over a specific period of time. Although the majority of research on burstness has focused on X as a scalar variable, the concept is intuitive enough to be expanded to a variable of multiple dimensions. In the real world, tsunamis would be a good example of a burst in a three dimensional space.

An Automaton

Kleinberg (2002) proposed a burst detection approach at the 8th ACM international conference on Knowledge Discovery and Data Mining (KDD). He models bursts in streams of text such as streams of email, publications, and speeches. The gap between the consecutive arrivals of items or events in time measures the frequency

[2] https://aclweb.org/aclwiki/RG-65_Test_Collection_(State_of_the_art).

of the events. A burst in a stream of email would be a period of time in which one receives many emails with small gaps. In contrast, during a period without any burst one would receive emails with much larger gaps. Such changes of frequencies are common in everyday life, for instance, distances between cars in rush hours.

Kleinberg's approach is to model the stream using an automaton that has an infinite number of states. In each state of the automaton, events take place at a particular rate. The automaton has states that characterize slow and fast rates of emission, a signal, an email, or an event. Streams with different rates can exist in the same system through state transitions. For instance, a slow-moving stream may be interwoven with a fast-moving stream by transiting from the corresponding slow-moving state to the state with a faster rate.

More formally, each stream is generated by an exponential distribution. Items in a stream are emitted probabilistically based on the exponential distribution so that the gap between one item and the next item follows the exponential density function $f(x) = \alpha e^{-\alpha x}$, where α is the rate of the arrival of the next item. If the automaton has two states that are responsible for emitting items at two different rates, low and high, then each state is modeled by its own exponential density function with α_{low} and α_{high}, respectively. The state transition probability in the automaton is p and it will remain in the same state with the probability of $1 - p$. Modeling the sequences with such an automaton is equivalent to determining the conditional probability of a state sequence based on the exponential density functions. The optimal sequence tends to minimize the number of state transitions; plus, the sequences would conform well to the corresponding gaps. Transitions to a high-frequency state will cost in proportional to a parameter gamma, but moving to a low-frequency will incur no cost.

Kleinberg demonstrated a hierarchical structure of the emails he received. The hierarchical structure revealed some bursts related to some intensive periods of emails due to proposal writing activities. His 2002 paper also included an example of 30 bursts detected from titles of all papers from two conferences between 1975 and 2001, namely SIGMOD and VLDB.

Burst Detection in CiteSpace

CiteSpace supports burst detection of several types of events, including citations to references and occurrences of keywords and noun phrases. The user may fine-tune the automaton by adjusting a few parameters of the automaton, including the minimum duration of a burst episode, state transition costs (gamma), and the ratio of the emission rates between states (Fig. 4.6).

Table 4.7 illustrates the burst durations of top 48 title terms with the strongest bursts in terrorism research between 1990 and 2017. The term biological terrorism has the strongest burst between 1996 and 2004. In terms of the automaton model, the term belongs to the state that emits articles at the fastest rate. A group of burst title terms are apparently related to the September 11 terrorist attacks in New York

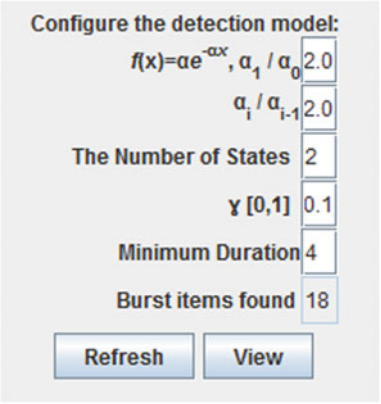

Fig. 4.6 The user can modify the automaton by adjusting a few parameters

Table 4.7 The burst durations of 48 title terms between 1990 and 2017 in terrorism research

Terms	Strength	Begin	End	1990–2017
Biological weapons	9.0056	**1990**	2003	
Terrorist bombing	5.0276	**1990**	2000	
Biological terrorism	12.6472	**1996**	2004	
Nuclear terrorism	6.0997	**2001**	2006	
World trade center attack	4.742	**2001**	2001	
Public health	4.0079	**2001**	2003	
New york city	9.3846	**2002**	2007	
Islamic terrorism	8.0579	**2002**	2005	
World trade center	5.1459	**2002**	2004	
Mass destruction	4.7395	**2002**	2006	
Military commissions	4.6649	**2002**	2003	
Terrorist attack	4.2836	**2002**	2005	
New York	4.0418	**2002**	2006	
11th terrorist attacks	3.9182	**2002**	2006	
Suicide terrorism	4.9632	**2003**	2010	
Mental health	5.4509	**2006**	2010	
Hurricane katrina	4.4913	**2006**	2010	
Global war	4.8608	**2007**	2009	
Southeast Asia	4.5094	**2007**	2009	
World trade center disaster	4.2206	**2008**	2011	
Northern Ireland	3.4748	**2009**	2013	
Intimate partner violence	6.4243	**2010**	2017	

(continued)

Table 4.7 (continued)

Terms	Strength	Begin	End	1990–2017
Economic growth	4.3858	**2010**	2013	
State terrorism	4.0271	**2010**	2011	
Systematic review	6.4222	**2011**	2017	
Comparative analysis	4.3312	**2011**	2014	
Terrorist threats	4.1981	**2011**	2013	
Public opinion	3.6887	**2011**	2015	
Domestic terrorism	3.5188	**2011**	2017	
Terrorist organization	6.2394	**2012**	2017	
Political violence	5.6434	**2012**	2014	
Terrorist group	5.2063	**2012**	2015	
Civil war	5.1036	**2012**	2017	
Posttraumatic stress symptoms	4.4322	**2013**	2017	
Risk perception	4.2900	**2013**	2015	
Social media	5.9780	**2014**	2017	
Terrorism research	3.8617	**2014**	2014	
Empirical analysis	3.5993	**2014**	2017	
Lone wolf	3.4993	**2014**	2015	
Islamic state	9.7410	**2015**	2017	
Armed conflict	7.5456	**2015**	2017	
Boko haram	7.1621	**2015**	2017	
European Union	5.1440	**2015**	2017	
Boston marathon bombing	4.5943	**2015**	2017	
National security	3.8458	**2015**	2017	
Violent extremism	5.1387	**2016**	2017	
Risk factor	4.6470	**2016**	2017	
Terror attacks	3.6236	**2016**	2017	

City, notably world trade center attack, new york city, and world trade center. The term world trade center disaster also has a burst between 2008 and 2011. Among terms with a period of burst within the last three years, three of them with the strongest bursts are associated with radical violent extremism, including islamic state, armed conflict, and boko haram. Boko Haram, for instance, is Nigeria's militant Islamist group responsible for a series of bombings, assassinations and abductions.

Figure 4.7 depicts the distributions of three title terms with the strongest bursts between 1990 and 2017. The term biological terrorism has the strongest burst between 1996 and 2004. The term Islamic state has the second strongest burst

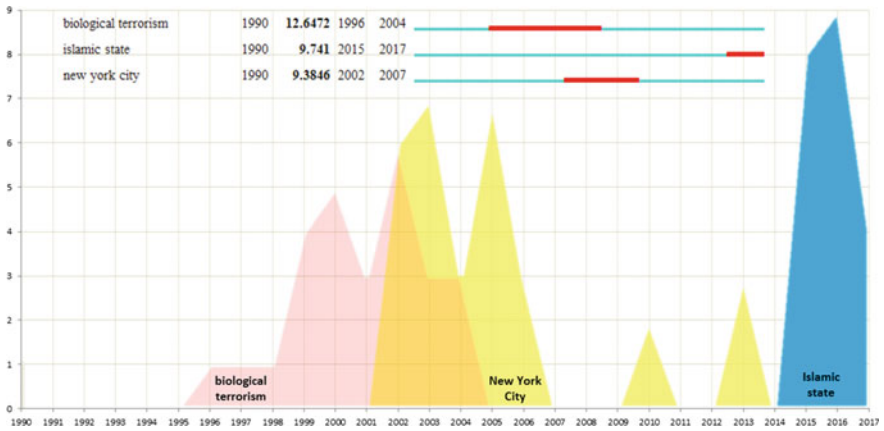

Fig. 4.7 The distributions of three title terms with the strongest bursts

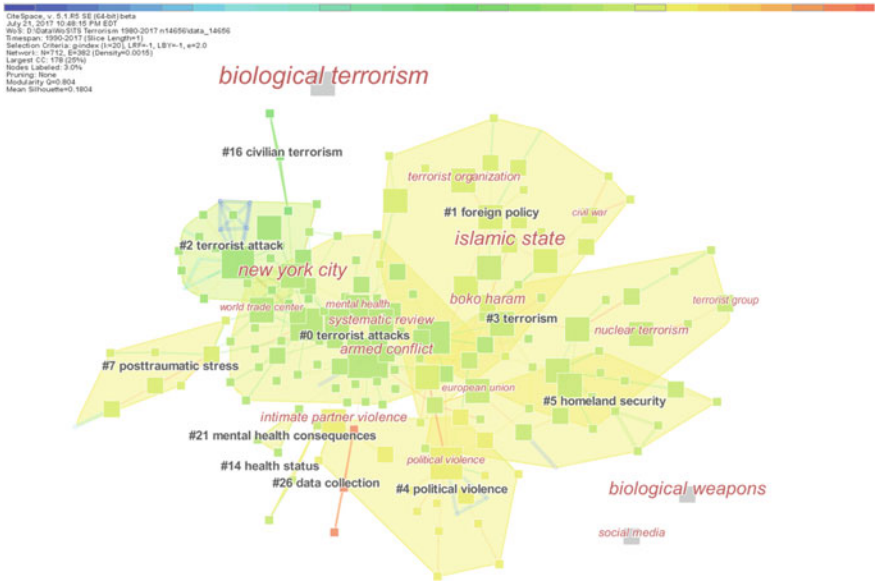

Fig. 4.8 A cluster view of title terms in terrorism research (1990–2017). Term labels are proportional to the strength of their burst. Labels starting with # are cluster labels, e.g. #0 terrorist attacks

between 2015 and 2017. The term New York City has the third strongest burst between 2002 and 2007. The term New York City appeared later on in titles in 2010 and 2013, but they are not bursts.

Figure 4.8 shows a network visualization of the title terms in terrorism research between 1990 and 2017. Publications on terrorism research in each year are selected

to form the network. The selection is based on the g-index, which is an index that quantifies scientific productivity. In fact, the g-index is an extension of the h-index (Hirsch 2005) such that, unlike the h-index, it takes into account citations of these publications. The g-index was proposed by Leo Egghe (2006). The g-index is defined such that the most cited g articles have at least g^2 citations.

$$\sum_{i=1}^{g} c_i \geq g^2$$

As a node selection criterion in CiteSpace, we modify the g-index with a constant k. When k = 1, the modified g-index is the same as Egghe's original g-index. When k > 1, the modified g-index would select more articles than the g-index because the actual citation count of an article is raised by k times.

$$k \sum_{i=1}^{g} c_i = \sum_{i=1}^{g} k \times c_i \geq g^2$$

Table 4.8 shows the selection process using the modified g-index. For instance, in 1999, there are eight articles in our dataset on terrorism research. The g-index for this group is 3, which means the three most cited articles together have 9 or more citations. If we set k as 1, then the most cited three papers will be selected out of the total of eight. If we would like to include more articles and set k to 20, then all eight articles meet the condition, i.e. 20 times the total citations of the eight articles are no less than 9 citations. As another example, our dataset includes 106 articles published in 1999. The citations of these articles yielded a g-index of 4, i.e. the subtotal of the four most cited articles is greater than or equal to 16. 25 articles become qualified based on k of 20 instead of 1. As the third example, our dataset has 2957 articles published in 2016 on topics relevant to terrorism research. The g-index is 11. By using k of 20, CiteSpace selected title terms from 79 articles instead of 11. Thus, using a k greater than 1 allows us to include more articles than using the original g-index.

Figure 4.8 shows a cluster view visualization of a network of co-occurring title terms between 1990 and 2017. The top level aggregates in the visualization are clusters. The label of each cluster starts with the character #, for example, #0 terrorist attacks. The size of a title term is proportional to the strength of a burst detected. The larger the node label size, the stronger a burst it has. Thus the one with the strongest burst is the term with the largest font size—biological terrorism. The second strongest burst is with Islamic state. The third one is with New York City.

Burst detection is a very valuable technique. It helps us to focus on the important development dynamically. It is also applicable to many types of events. In addition to detect bursts in title words, we can also apply the technique to identify bursts in

Table 4.8 The number of articles selected by the g-index each year to construct the network of title terms

Time slice	g-index	Articles	Selected articles	Links/all
1990–1990	g = 3, k = 20	8	8	3/3
1991–1991	g = 2, k = 20	12	12	8/8
1992–1992	g = 2, k = 20	15	15	14/14
1993–1993	g = 2, k = 20	11	11	5/5
1994–1994	g = 2, k = 20	21	21	23/23
1995–1995	g = 2, k = 20	21	21	11/11
1996–1996	g = 2, k = 20	42	21	6/6
1997–1997	g = 3, k = 20	68	22	4/4
1998–1998	g = 3, k = 20	49	22	7/7
1999–1999	g = 4, k = 20	106	25	4/4
2000–2000	g = 4, k = 20	95	27	5/5
2001–2001	g = 5, k = 20	163	35	5/5
2002–2002	g = 10, k = 20	895	65	37/37
2003–2003	g = 8, k = 20	1009	58	22/22
2004–2004	g = 9, k = 20	1197	64	35/35
2005–2005	g = 7, k = 20	1488	58	17/17
2006–2006	g = 9, k = 20	1595	66	18/18
2007–2007	g = 7, k = 20	1682	60	16/16
2008–2008	g = 9, k = 20	1717	66	20/20
2009–2009	g = 9, k = 20	1796	62	12/12
2010–2010	g = 7, k = 20	1830	57	17/17
2011–2011	g = 10, k = 20	2065	70	28/28
2012–2012	g = 7, k = 20	1848	59	7/7
2013–2013	g = 7, k = 20	1797	57	6/6
2014–2014	g = 8, k = 20	1800	59	21/21
2015–2015	g = 8, k = 20	2191	64	22/22
2016–2016	g = 11, k = 20	2957	79	28/28
2017–2017	g = 6, k = 20	931	46	10/10

citations and bursts in institutions and individuals that are particularly active on specific topics. Unlike many popular indices of scientific productivities such as the h-index and the g-index, burst detection can tell us much more about the dynamics of the underlying process so that one can better understand how the process pans out. Burst detection can help us answer many specific questions: does an individual researcher have a burst in terms of the number of articles he/she published? If so, when did the most recent episode of burst begin? How long did the period of burst last? Is the researcher still at a state with a high productivity?

Log-Likelihood Ratio

Many commonly used statistical methods such as z-standard scores assume that the data is normally distributed. When dealing with text analysis, however, it is most likely that the normal distribution assumption is no longer valid, especially when we focus on terms that represent emerging topics or novel concepts. Researchers have shown that statistics based on the normal distribution assumption in such cases often overestimate the occurrences of rare words and that much of the content bearing words, technical jargons, and domain-specific terminologies in scientific publications are rare in the pool of English words in general.

Ted Dunning is currently the Chief Application Architect at MapR. Nearly 25 years ago, in 1993, he wrote an influential paper on text analysis (Dunning 1993). In the paper, he demonstrated the advantages of log-likelihood ratio tests for identifying relatively rare but significant patterns in text, for example surprising and unexpected combinations of words. His article now has 2773 citations on Google Scholar. With this amount of citations, the paper would be very close to the peak of the Mount Kilimanjaro. As a reference the 2006 JASIST paper on CiteSpace (Chen 2006) now has 1716 citations on Google Scholar.

Likelihood Ratio

Parametric and nonparametric are two board classifications of statistical procedures. One way to differentiate one from another is whether a statistical procedure relies on any assumptions about a probability distribution from which the data were drawn. The bottom line is whether a statistic procedure makes any use of such an assumption. For instance, to calculate z-scores, or standard scores, we need to know the mean and standard deviations of the underlying distribution of the data. The mean and standard deviations only make sense if the data were normally distributed. Therefore, the statistical procedure regarding the z-scores is parametric. In contrast, nonparametric tests are also called distribution free because they do not rely on any assumptions about the underlying distributions.

Given outcomes $\boldsymbol{k}$ as a point in the space of observations K, a set of model parameters ω as a point in the parameter space Ω, the likelihood $H(\omega; k)$ is the probability $P(k|\omega)$ that the outcome $\boldsymbol{k}$ would be observed given those parameter values at ω. $H(\omega; k)$ is the notation used by Dunning in his 1993 article.

$$H(\omega; k) = P(k|\omega)$$

For example, the likelihood function for repeated Bernoulli trials can be defined as follows:

$$H(\omega; k) = H(p; n, m) = p^m (1-p)^{n-m} \binom{n}{m}$$

In this case, the parameter space Ω is the set of all the probabilities p, i.e. [0, 1], whereas the subspace Ω_H for the hypothesis that $p = p_H$ is a singleton set $\{p_H\}$, which is a subset of [0, 1].

The likelihood ratio λ for a hypothesis is the ratio of two maxima of the likelihood function. One is the maximum value of the likelihood function over a subspace Ω_H on which the hypothesis applies. The other is the maximum value of the likelihood function over the entire parameter space Ω.

$$\lambda = \frac{\max\limits_{\omega \in \Omega_H} H(\omega; k)}{\max\limits_{\omega \in \Omega} H(\omega; k)}$$

For two binomial processes that are characterized by p_i,m_i, and n_i for $i = 1$ and 2, the maxima are reached when $p_1 = \frac{m_1}{n_1}$, $p_2 = \frac{m_2}{n_2}$, and $p = \frac{m_1 + m_2}{n_1 + n_2}$. Let

$$L(p, m, n) = p^m (1-p)^{n-m}$$

The log-likelihood ratio can be computed as follows:

$$-2log\lambda = 2(\log L(p_1, m_1, n_1) + \log L(p_2, m_2, n_2) - \log L(p, m_1, n_1) - \log L(p, m_1, n_1))$$

The value $-2\log\ \lambda$ is asymptotically distributed as χ^2 with the difference between the dimensions of Ω and Ω_H as the degree of freedom. Thus the log-likelihood ratio value is associated with a p-level, which indicates the statistical significance of the observed event. The 'oddness' measures how special the observation is.

Characterizing a Cluster

A major advantage of a likelihood ratio test helps us to identify events that are particularly more common in a subspace of the parameter space than the entire parameter space. A term that is particularly unique in one cluster but not in other clusters would have a very high likelihood ratio on the subspace associated with the matching cluster. For instance, the term post-traumatic stress disorder would stand out in terms of its likelihood ratio to differentiate a cluster on this topic from other topics in terrorism research.

Table 4.9 lists two sets of title terms selected from the three largest clusters. One set was selected by Latent Semantic Indexing (LSI) (Deerwester et al. 1990). The other was selected by Log-Likelihood Ratio (LLR). Two numbers are shown next to each term selected by LLR. The first number is the $-2\log\ \lambda$ value of the

Table 4.9 Representative terms selected by LSI and Log-Likelihood Ratio Tests for the largest three clusters in Project Demo 1 on terrorism research (1996–2003)

Cluster	Label (LSI)	Label (LLR)
0	Bioterrorism	Biological terrorism (8082.39, 1.0E-4)
	Reason	Front line (5684.68, 1.0E-4)
	Small molecule	New york city (5658.81, 1.0E-4)
	Family physicians	Emergency physician (5400.81, 1.0E-4)
	Thought	Blast over-pressure (4767.67, 1.0E-4)
	Nation	Terrorist attack (4541.68, 1.0E-4)
	Collaborative literature	11th terrorist attack (4210.69, 1.0E-4)
	Cure	Posttraumatic stress disorder (3605.89, 1.0E-4)
	Intentional poisoning	Chemical terrorism (3438.99, 1.0E-4)
	Bind|terrorism	Biological weapon (3269.72, 1.0E-4)
	Community-based model	Bioterrorism preparedness (3220.3, 1.0E-4)
	Protecting rural communities	Public health management (2887.06, 1.0E-4)
	Large-scale quarantine	Overpressure-induced injury (2811.21, 1.0E-4)
	Following biological terrorism	Involving hemoglobin (2811.21, 1.0E-4)
	Possible consequences	Biochemical mechanism (2811.21, 1.0E-4)
	Medical technicians	Oklahoma city bombing (2792.14, 1.0E-4)
	Panic	World trade center (2789.9, 1.0E-4)
	Common goods	Hospital preparedness (2768.94, 1.0E-4)
	Predictions	Medical response (2510.49, 1.0E-4)
		Psychological sequelae (2417.99, 1.0E-4)
1	Terrorism	Blast over-pressure (18729.92, 1.0E-4)
	Mental health responses	Overpressure-induced injury (11125.44,
	UCH experience	1.0E-4)
	Bomb blast	Involving hemoglobin (11125.44, 1.0E-4)
	Biochemical mechanism	Biochemical mechanism (11125.44, 1.0E-4)
	Blast lung injury	Conventional weapon threat (3893.35, 1.0E-4)
	Oklahoma city bombing	Medical consequence (3893.35, 1.0E-4)
	Pulmonary blast injury	Blast injury (3456.56, 1.0E-4)
	Explosion survivors	Exercise performance (3281.15, 1.0E-4)
	Sublethal blast overpressure| major	Sublethal blast overpressure (3281.15, 1.0E-4)
	incidents	Food intake (3281.15, 1.0E-4)
	Proposal	Social consequence (2223.28, 1.0E-4)
	Dissemination	Physical injury (1588, 1.0E-4)
	Manchester bombing	Soho nail bomb (1575.04, 1.0E-4)
	Casualty profiles	UCH experience (1575.04, 1.0E-4)
	Casualty profile	Terrorist bombing (1354.2, 1.0E-4)
	Construction	Evolving threat (1329.12, 1.0E-4)
	Hazmat	Biological terrorism (1316.57, 1.0E-4)
	Suicidal deaths	Terrorist attack (1076.2, 1.0E-4)
	Pathologic features	Open-air bombing (1074.82, 1.0E-4)

(continued)

Table 4.9 (continued)

Cluster	Label (LSI)	Label (LLR)
		Confined-space explosion (1074.82, 1.0E-4)
2	September Terrorist attacks Negative changes Following vicarious exposure Exposure New York city children Posttraumatic stress reactions Stress-related mental health Israel Coping behaviors|terrorism OPM-sang experience Risk assessment Functional impairment Supporting children Youth Television exposure International relations Oklahoma city Warfare	Terrorist attack (9718.45, 1.0E-4) New York city (8811.82, 1.0E-4) 11th terrorist attack (7733.12, 1.0E-4) Biological terrorism (7130.79, 1.0E-4) Posttraumatic stress disorder (6505.87, 1.0E-4) World trade center (5195.7, 1.0E-4) Psychological sequelae (4499.33, 1.0E-4) Blast over-pressure (3850.16, 1.0E-4) Biological weapon (3797.09, 1.0E-4) New York (3233.08, 1.0E-4) Front line (2935.05, 1.0E-4) Emergency physician (2882.3, 1.0E-4) Vulnerable population (2430.25, 1.0E-4) Drug user (2430.25, 1.0E-4) Bioterrorism preparedness (2339.45, 1.0E-4) Overpressure-induced injury (2270.61, 1.0E-4) Involving hemoglobin (2270.61, 1.0E-4) Biochemical mechanism (2270.61, 1.0E-4) Prior trauma (2247.83, 1.0E-4) Posttraumatic stress symptom (2247.83, 1.0E-4)

Up to top 20 terms are selected for each cluster

log-likelihood ratio. The larger this number is, the more special the term for the current cluster. The second number is the statistical significance of the $-2\log \lambda$ value from a χ^2 distribution. It is the p-level of the term.

The strongest LLR terms in the largest cluster #0 include biological terrorism (8082.39, 1.0E-4), front line (5684.68, 1.0E-4). The value 1.0E-4 is the statistical significance of the LLR value 8082.39 and 5684.68 according to a χ^2 distribution. The term biological terrorism is specific enough to give us a clear idea what the cluster is about. In contrast, the term front line is more ambiguous. Similarly, on the LSI list, terms such as reason, thought, and nature are usually too broad to be useful even within the specific context of a co-citation cluster.

Table 4.9 shows terms short selected by LSI and LLR as candidates for cluster labels. If the two lists match, then the decision would be easy. If they differ substantially, we need to investigate further. For the largest cluster (#0), it is relatively easy because bioterrorism and biological terrorism are semantically equivalent. Terms such as reason, thought, and front line are common on Google, suggesting that they are not good candidates for cluster labels because they are too broad and ambiguous to be informative. Although the LLR list includes 11th terrorist attack (LLR = 4210.69) and Oklahoma City bombing (LLR = 2792.14), their log-likelihood ratios are much lower than that of biological terrorism (LLR = 8082).

For the second largest cluster (#1), LSI identifies terms such as terrorism, mental health responses and uch experience, whereas LLR identifies blast over-pressure (LLR = 18,729.92), overpressure-induced injury (LLR = 11,125.44), and involving

hemoglobin (LLR = 11,125.44). The LLR terms seem to suggest a theme on physical injuries, but the LSI list includes mental health responses. Further down the LSI list, there are terms related to physical injuries such as blast lung injury, pulmonary blast injury, and sublethal blast overpressure.

For the third largest cluster (#2), the top three LLR terms are terrorist attack (LLR = 9718.45), New York city (LLR = 8811.82), and 11th terrorist attack (LLR = 7733.12). The LSI list is topped by terms such as September and terrorist attacks. These terms strongly suggest that this cluster is about the September 11th terrorist attacks in 2001 at the World Trade Center in New York.

More generally, it is useful to differentiate two kinds of words in text, depending on their role in a sentence: function words or content-bearing words. Function words organize different parts of a sentence together but they don't mean anything on their own. In contrast, content-bearing words are the ones that carry the meaning of a sentence. A sentence would be meaningless without such contents. Content-bearing words to a sentence would be similar to the wine to a bottle. An earlier description of the distinction can be found in Charles Carpenter Fries' work (Fries 1952). Function words are also called structure words, whereas content words are also called lexical words.

Function words include prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles. For instance, commonly seen function words include the, a, her, however, and otherwise. Content words are those that are not function words. Nouns, verbs, adjectives, and most adverbs are examples of content words. 99.9% of words in English are content words.

The following example is based on Project Demo 1: Terrorism Research (1996–2003) in CiteSpace. We first generated a network of co-cited references and then divided the network into several clusters. Each cluster is resulted from the citations made by a group of published articles. In order to understand what a cluster is about, one may inspect whether there are common reasons for these articles to cite the member references of the cluster together. CiteSpace implements a few functions to label a cluster based on terms selected from citing articles' titles, keywords, abstracts, or any combinations of terms from these fields. Figure 4.9 shows several clusters with automatically generated labels.

In addition to rank terms based on log-likelihood ratio tests with respect to their roles in a subspace of the underlying model, log-likelihood ratio tests can also measure associations between two terms so that one can generate an associative network of concepts or terms extracted from text. CiteSpace supports a function to compute the strengths of associations based on log-likelihood ratio tests (Fig. 4.10).

Similarities between terms can be measured in terms of how often they appear together, i.e. co-occurrences, and how likely they appear given the fact that they are published in the same journal. Figure 4.11 illustrates some of the interrelationships between title terms from publications in the journal Scientometrics. The strength of a link is based on a log-likelihood ratio test that compares the probability of co-occurrences with probabilities of the entire parameter space, including other scenarios in which only one of them appears or none of them appears. The log-likelihood ratio (LLR) between terms publications and papers is 0.8390, which

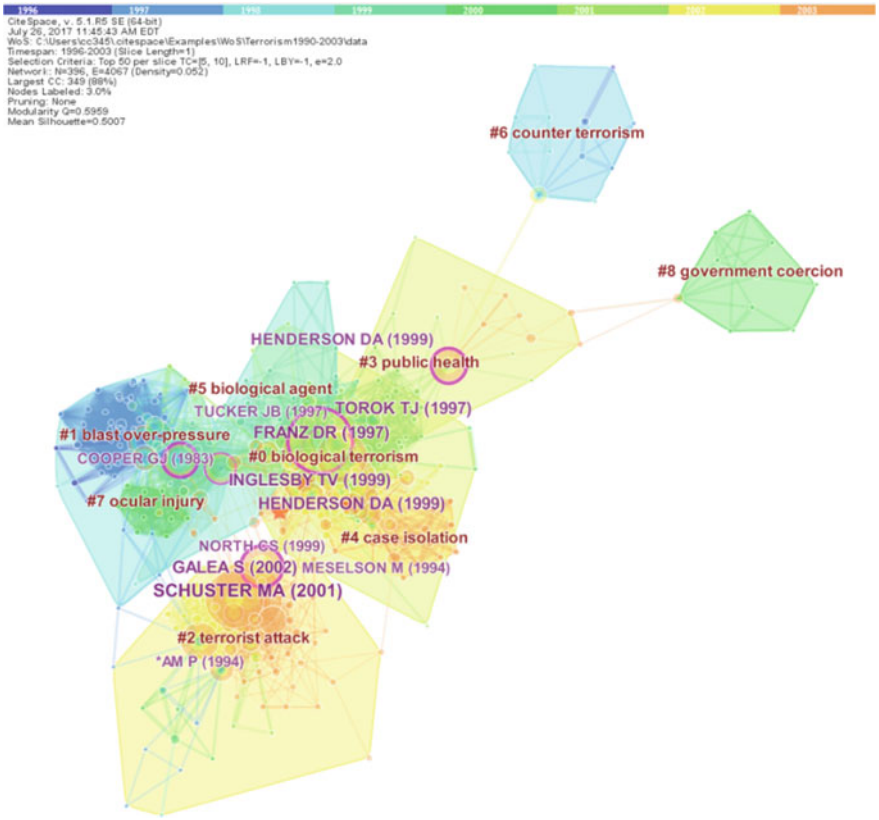

Fig. 4.9 Project Demo 1 in CiteSpace. Cluster labels are selected by LLR

is relatively low because the two terms are semantically equivalent so they are less likely to appear together. In contrast, the LLR between citation and impact is 8.1686 and the LLR between happiness and feelings is 10.6193. These relatively higher LLRs suggest some special connections between citation and impact and between happiness and feelings.

Entropy

Table 4.10 illustrates an approximate number of documents on Google that contain a term. The higher the number of instances on Google, the higher the probability of the term and the lower its information entropy is. For example, words such as 'the' and 'a' appeared most often on Google. Both are estimated to have appeared in approximately 25,270,000,000 documents on Google. In contrast, terms such as small molecule, bioterrorism, and posttraumatic stress disorder have much fewer

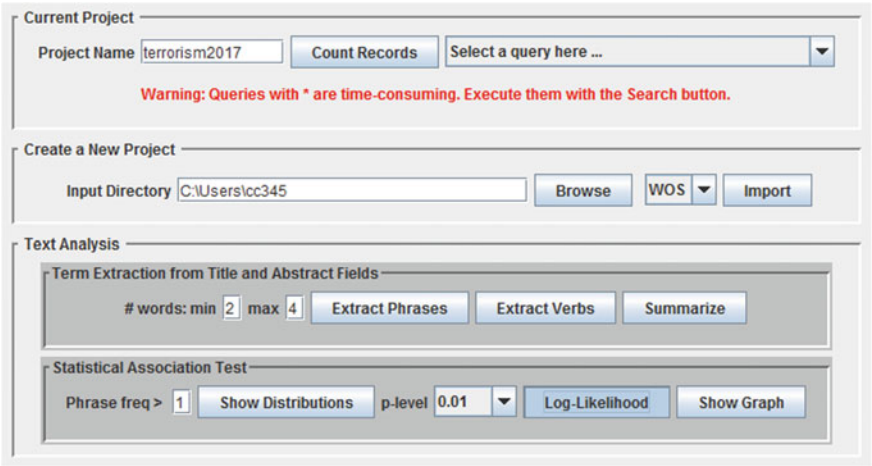

Fig. 4.10 Compute statistical associations with log-likelihood ratio tests in CiteSpace

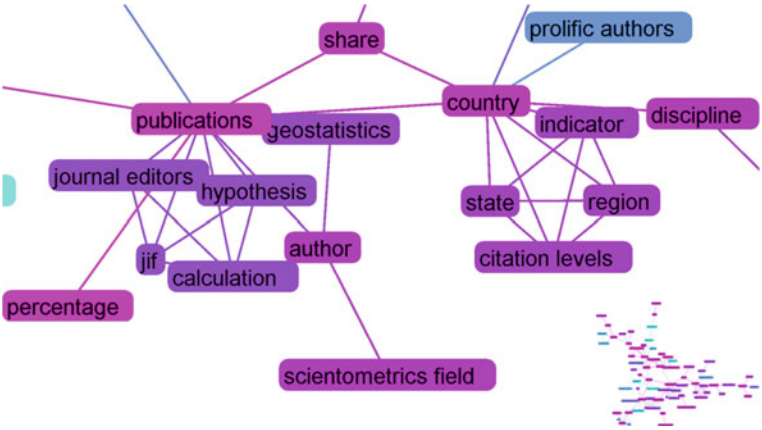

Fig. 4.11 Associations between title terms articles published in the Scientometrics

appearances, namely 22,100,000, 7,120,000, and 1,130,000, respectively. By the same method, Gone with the Wind has 90,200,000 hits. CiteSpace has 47,600 hits.

The information entropy of a term can be seen as a measure of its associated uncertainty. If we consider the appearance of a term as an event that transmits a message, then observing a rare event taking place is more information than observing a common event. Entropy is zero when we have nothing to learn from the occurrence of an event. The entropy reaches its maximum when the uncertainty is the highest, or, the occurrences of an event are completely random.

Figure 4.12 shows a plot of the information entropy of terms extracted from articles on terrorism each year. As new vocabularies are introduced into the latent

Table 4.10 The popularity of a few terms on Google as of July 26, 2017

Term	Instances on Google
The	25,270,000,000
a	25,270,000,000
It	19,730,000,000
Thought	1,770,000,000
Reason	1,540,000,000
Front line	291,000,000
Terrorism	147,000,000
Gone with the Wind	90,200,000
Small molecule	22,100,000
Bioterrorism	7,120,000
Posttraumatic stress disorder	1,130,000
UCH[a] experience	627,000
Blast over-pressure	278,000
CiteSpace	47,600

[a]UCH = University City Hospital

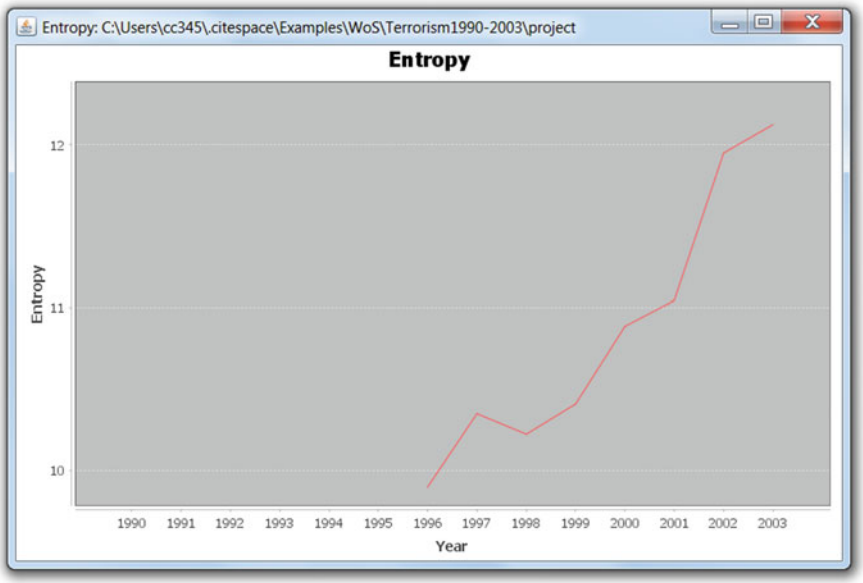

Fig. 4.12 The information entropy of terms extracted from articles on terrorism each year

semantic space, the information entropy would be higher in that year than before. The uncertainty of the latent semantic space increases due to the appearance of the additional terms. As shown in the plat, the largest increase is between 2001 and 2002 due to the September 11 terrorist attacks.

Table 4.11 illustrates top 20 terms based on their information entropy (Shannon 1948). The entropy of a term is calculated based on its distribution over the years between 1996 and 2003. The term explosion has the highest entropy. In other words, it is the least informative term in the context of terrorism. The standard deviation of the occurrences of a term can be used to measure the stability of a term. For instance, the term casualties has a standard deviation of 3.204, reflecting an uneven distribution of the term over the years. In comparison, the term explosion has a standard deviation of 1.309, reflecting a relatively stable distribution over this period of time.

Clumping Properties of Content-Bearing Words

Clumping was introduced in by Bookstein et al. (1998). They also introduced four ways to measure and identify clumping terms. Clumping metrics can be applied to an arbitrarily long text document or a chain of documents. The key assumption here is that content-bearing terms are more likely to clump than non-content-bearing ones. If non-content-bearing terms are randomly distributed throughout a text document, then one may focus on terms that their distributions deviate from the random distributions.

Condensation

The concept of clumping is similar to clustering except that clumping assumes a sequential order as an internal structure between items. Thus clumping can be seen as serial clustering of terms in text. The interest of studying clumping properties is to see whether a term appears unusually close together. The spatial closeness reminds us the temporal closeness associated with the topic of burst detection.

If we take sentences as units of observation, we would expect the number of sentences containing a given term to be less than the total number of occurrences of the term. If the term is clumping, the number of sentences containing the term should be even fewer. The degree of such condensation can be measured by the ratio of the actual number of sentences containing the term to the expected number of sentences of the term if it is randomly distributed. In practice, the unit can be a single sentence, a block of sentences, or a paragraph.

Given a term t, suppose the document to be analyzed has D units, N of them contain t, and t occurs T times in total. The number of ways to have T occurrences in D units is D^T, i.e. for each of the T instances, select a unit from D. The next step is to calculate the probability that exactly N units contain one or more instances of t. There are $\binom{D}{N}$ ways to select the units with at least one hit. A Stirling number of

Table 4.11 The information entropy of a term in articles of terrorism research (1996–2003)

TERM	Entropy	1996	1997	1998	1999	2000	2001	2002	2003	Subtotal	S.D.
Explosion	**2.914**	2	2	4	6	3	3	4	4	28	1.309
Blast	**2.901**	3	6	2	5	3	3	3	6	31	1.553
Tokyo	**2.877**	1	5	2	2	4	3	3	3	23	1.246
Injuries	**2.872**	4	2	3	4	4	6	6	9	38	2.188
Injury	**2.869**	4	5	1	4	5	4	5	8	36	1.927
Nerve	**2.840**	1	1	2	2	1	2	2	4	15	0.991
Bombing	**2.807**	3	3	4	6	6	1	9	4	36	2.449
Mortality	**2.790**	1	1	2	4	1	2	3	1	15	1.126
Israel	**2.780**	3	2	4	2	4	3	4	10	32	2.563
Gas	**2.786**	1	2	1	2	3	5	4	5	23	1.642
Sarin	**2.780**	0	3	2	2	3	2	3	3	18	1.035
Casualties	**2.765**	3	1	3	5	5	7	8	11	43	3.204
Oklahoma	**2.757**	2	2	1	5	6	1	4	3	24	1.852
Hospital	**2.745**	1	3	1	6	2	4	6	6	29	2.200
Children	**2.738**	6	2	2	3	5	2	6	11	37	3.114
Physicians	**2.733**	2	1	2	2	2	7	6	4	26	2.188
France	**2.725**	1	0	1	1	2	1	2	1	9	0.641
Disruption	**2.722**	2	0	2	1	1	1	2	1	10	0.707
Terrorist bombings	**2.722**	1	1	1	2	0	2	2	1	10	0.707
Biological terrorism	**2.720**	1	1	3	5	5	9	8	5	37	2.925

the second kind is a partition of T terms into N classes. There are N! ways to order the components of the partition. The probability p(N, T) is therefore:

$$p(N,T) = \frac{N!\begin{pmatrix} D \\ N \end{pmatrix}\begin{Bmatrix} T \\ N \end{Bmatrix}}{D^T}$$

The expected number of units containing t is as follows. See Bookstein et al. (1998) for details.

$$E_{C1} = D\left[1 - \left(1 - \frac{1}{D}\right)^T\right]$$

If N units contain term t, then N/E_{c1} measures the strength of the condensation. For a clumping term, this ratio would be less than 1. As we will see shortly, this is the clumping measure implemented in CiteSpace.

The second condensation-based measurement is based on specific distributions of terms over units. The probability that m occurrences of a term appear in any given unit can be modeled by the binomial distribution:

$$p(m) = \begin{pmatrix} T \\ m \end{pmatrix}\left(\frac{1}{D}\right)^m\left(1 - \frac{1}{D}\right)^{T-m}$$

Thus one can expect $p(m)*D$ units to contain m occurrences.

The third measurement is the number of clumps. Here a clump is defined as a consecutive chain of units containing the term. The probability of K clumps of the term t is defined as follows. Again see Bookstein et al. (1998) for detailed reasoning.

$$p_K = \frac{\begin{pmatrix} N-1 \\ K-1 \end{pmatrix}\begin{pmatrix} D-N+1 \\ K \end{pmatrix}}{\begin{pmatrix} D \\ N \end{pmatrix}}$$

The expected number of clumps is defined by the following formula:

$$E_{L1} = (D - N + 1)\frac{N}{D} = N\left(1 - \frac{N-1}{D}\right)$$

The ratio K/E_{L1} measures linear-clustering clumping. Content-bearing terms are terms the ratio of which is substantially less than one.

Finally, the fourth measure of clumping is based on gap length between marked units. If N marked units are randomly distributed over D units of text, then the probability that a randomly chosen unit not be marked is $\gamma = 1 - N/D$.

The probability of r blank units between two marked units is given approximately by the geometric distribution.

*Clumping Versus TF*IDF*

Table 4.12 illustrates terms with strong condensation strengths (clumping) and top terms identified by term frequency (TF) by inverse document frequency (IDF). Perhaps more interestingly, these terms are associated with the largest co-citation cluster—the one on biological terrorism. Terms that appear on both lists are highlighted in the table.

The top clumping terms include radiation, virus, vaccine, plague, toxins, hemorrhagic, spores and toxin. These terms are clearly related to the central theme of biological terrorism. These terms are domain-dependent terms. Removing or ignoring the role of these terms will undermine the adequacy of a study. The clumping list also includes some terms that are not as tightly connected to biological terrorism as the first type of terms. The second type of terms include food, water, and protective. Yet another group of terms on the clumping list are domain independent terms. One can expect to see these terms in publications on any research topic, namely, evaluations, consensus, model, task, and final. The first three types are domain dependent. The fourth, sixth, are eighth and domain independent.

By applying the same classification heuristics to the list of terms ranked by their TF*IDF scores, the TF*IDF list has fewer Type 1 terms than the clumping list (4 vs. 10), more Type 2 terms (12 vs. 4), and about the same number of Type 3 terms (4 vs. 6).

Importance and Impact

Among the many types of importance metrics, two are particularly relevant to our understand how scientific knowledge is organized and diffused: eigenvector centrality and betweenness centrality. The eigenvector centrality is also called eigencentrality. Both of them measure the importance of a node in a network.

Degree Centrality and Eigenvector Centrality

The degree centrality of a node in a network is a simple measure of the node's importance in terms of how many nodes it connects to (Freeman 1977). Within the same network, a person with a lot of friends will have a higher degree centrality

Table 4.12 Top 20 content-bearing terms from the largest cluster of terrorism research (1990–2003) along with top terms identified by TF*IDF

TF	IDF	TF*IDF	Clumping	Term	Type	TF	IDF	TF*IDF	Clumping	Term	Type
21	3.53	74.05	**0.16**	*Radiation*	1	67	1.39	**92.88**	0.50	Chemical	2
15	3.26	48.87	**0.29**	Virus	1	127	0.69	**88.03**	0.67	Health	2
14	3.26	45.61	**0.30**	Vaccine	1	73	1.1	**80.2**	0.61	Bioterrorism	2
10	3.53	35.26	**0.31**	Evaluations	3	21	3.53	**74.05**	0.16	*Radiation*	1
28	2.56	71.82	**0.32**	*Food*	2	67	1.1	**73.61**	0.59	Emergency	2
13	3.26	42.36	**0.33**	Consensus	3	53	1.39	**73.47**	0.58	Response	3
20	2.83	56.66	**0.33**	*Anthracis*	1	28	2.56	**71.82**	0.32	*Food*	2
16	3	47.93	**0.34**	Model	3	49	1.39	**67.93**	0.59	Attack	2
16	3	47.93	**0.34**	Water	2	59	1.1	**64.82**	0.64	Care	2
9	3.53	31.74	**0.35**	Regional	2	25	2.56	**64.12**	0.36	*Training*	3
15	3	44.94	**0.36**	Plague	1	29	2.2	**63.72**	0.43	Physicians	2
25	2.56	64.12	**0.36**	*Training*	3	27	2.3	**62.17**	0.42	Smallpox	1
11	3.26	35.84	**0.38**	Protective	2	88	0.69	**61.00**	0.65	Medical	2
11	3.26	35.84	**0.38**	Task	3	23	2.56	**58.99**	0.39	*Toxins*	1
23	2.56	58.99	**0.39**	*Toxins*	1	36	1.61	**57.94**	0.62	Preparedness	2
8	3.53	28.21	**0.39**	Botulinum	1	83	0.69	**57.53**	0.66	Public	3
8	3.53	28.21	**0.39**	Final	3	20	2.83	**56.66**	0.33	*Anthracis*	1
8	3.53	28.21	**0.39**	Hemorrhagic	1	80	0.69	**55.45**	0.64	Agents	2
16	2.83	45.33	**0.40**	Spores	1	28	1.95	**54.49**	0.57	Detection	3
16	2.83	45.33	**0.40**	Toxin	1	49	1.1	**53.83**	0.71	Weapons	2

The cluster is labeled as bioterrorism. Terms appear on both lists are emphasized

than that of someone with fewer friends. Two people with the same number of friends will have the same degree centrality.

More realistically, friends may have their own friends. If person A has a friend who has many friends and person B has a friend who has no more friends, should A and B have the same centrality? Unlike degree centrality, the eigenvector centrality treats friends differently. Connecting to an important friend will increase your own importance. As many have put it, it is about who you know, at least sometimes. In a social network, having many friends is generally a good idea unless all your friends are antisocial except with you.

The most famous member of the eigenvector centrality family is probably Google's PageRank. Recent research in neuroscience found that the eigenvector centrality of a neuron in a neural network is correlated with its relative firing rate.[3] The eigenvector centrality has been used to measure the prestige of a scientific journal, notably, the SJR indicator developed by a group of researchers in Spain.

The original idea can be traced to the works of Leontief (1941) and that of Seeley (1949) on reciprocal influence in social metric networks in 1949. The work of Phillip Bonacich (1972) is also widely known in relevant literature. Here we use Bonacich's notation. Given a network, the e_i centrality of node n_i in a network reflects the centralities of its neighboring nodes.

$$\lambda e_i = \sum_j R_{ij} e_j$$

Or, equivalently,

$$\lambda e = Re$$

where R is a matrix representation of the network. The diagonal values of R are zeros, i.e. $r_{ij} = 0$. By definition, e is the an eigenvector of R and λ is the corresponding eigenvalue.

As illustrated in Fig. 4.13, the visualization on the right shows that Cluster #1 at the top level (Level 0) has a concentration of nodes with high eigenvector centrality scores. In the context of a co-citation network, a high eigenvector centrality node means that it is co-cited with some well-connected references. The density of Cluster #1 is considerably higher than the density of the network overall.

The visualization on the left is generated based on articles that cited references in Cluster #1. Articles that did not cite any members of Cluster #1 are omitted from this procedure. As a result, the new network not only preserves the essential

[3]Fletcher, Jack McKay and Wennekers, Thomas (2017). From Structure to Activity: Using Centrality Measures to Predict Neuronal Activity. *International Journal of Neural Systems*. **0** (0): 1750013. doi:10.1142/S0129065717500137.

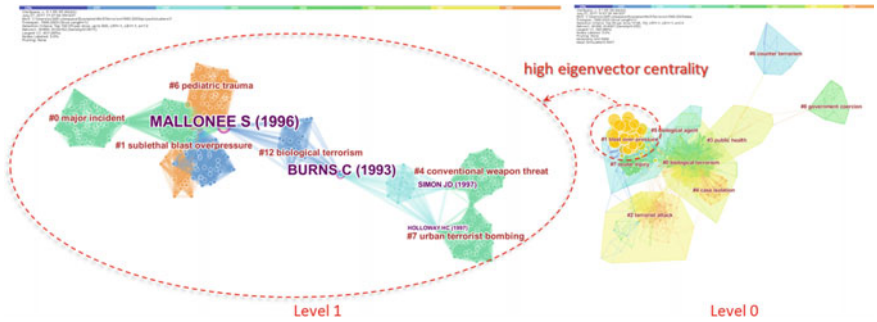

Fig. 4.13 High eigenvector centrality nodes are concentrated in Cluster #1 blast over-pressure. Zooming into #1 at the next level reveals high betweenness centrality nodes such as Mallonee1996 and Burns1993

structure of Cluster #1, but also reveals additional details. For instance, the new network reveals that it does not find any references published in 1996, 2000, and 2002. Instead, it contains publications in 1997–1999, 2001, and 2003. Cluster #1 is further divided into several clusters, including Level-1 clusters such as #0 major incident, #1 sublethal blast overpressure, #12 biological terrorism, and #6 pediatric trauma.

Two prominent nodes have strong betweenness centrality scores: MALLONEE1996 and BURNS1993. The betweenness centrality of a node measures the extent to which the node is in the middle of two or more dense areas. Suppose node v is connecting two sub-networks A and B. If the only way to reach one of the sub-networks from the other one is to go through node v, then the betweenness centrality of the node will reach the maximum possible level. The more alternative paths there are to bypass the node, the lower its betweenness centrality value will be.

The two nodes with strong betweenness centrality scores nicely illustrate the meaning of betweenness centrality in the visualization. MALLONEE1996 plays a central role in connecting at least three Level-1 clusters in three different colors. Removing MALLONEE1996 from the network will effectively disconnect these clusters because MALLONEE1996 is the only common node they share. Furthermore, MALLONEE1996 connects a 2003 cluster—#6 pediatric trauma (brown)—with clusters formed a few years ago (in blue and green years between 1997 and 1999), suggesting that in 2003 researchers revisited issues that had been addressed in 1997–1999. Such visits and revisits to the same research topics may explain the concentration of high eigenvector centrality nodes.

Figure 4.14 shows three displays of different metrics, namely, betweenness centrality, PageRank, and eigenvector centrality. The distributions of these metrics are different because they are designed to highlight different properties.

Betweenness centrality is effective in identifying critical information for understanding interrelationship between two or more clusters. Eigenvector centrality generalizes degree centrality by incorporating the importance of the

Fig. 4.14 The size of a node represents its betweenness centrality (left), PageRank (middle), and eigenvector centrality (right)

neighbors. Eigenvector centrality implemented in CiteSpace follows Zafarani et al. (2014).

Hirsch Index

Extrinsic factors are more common in the literature because of their relatively longer history. The most widely known examples include the Hirsch-index (Hirsch 2005), or the h-index, and the journal impact factor. Both of them have been extensively used and both have been subject to a wide variety of criticisms and modifications.

The *h*-index was introduced as an indicator of the productivity of a scientist in terms of all his/her *N* publications $\{a_i\}$ and corresponding citations $\{c(a_i)\}$, where $i = 1, 2, \ldots, h, \ldots, N$. For simplicity, assume the publications are sorted by their citations in descending order. The magic number *h* is the largest number of top cited *h* publications that have at least *h* citations, $c(a_i) \geq h$, for the scientist.

$$h = \max_i (i | c(a_i) \geq h), \text{ where } c(a_i) \geq c(a_j) \text{ if } i < j$$

Since the coverage of one's publications varies from one source to another, one's h-index varies depends on whether the calculation is based on Google Scholar, the Web of Science, Scopus, or anything else.

For instance, as of August 8, 2017, Loet Leydesdorff, an active and productive researcher in scientometrics and several other fields, has a total of 36,474 citations for his hundreds of publications that we can find on Google Scholar. His h-index on Google Scholar is 86. By definition, among his numerous publications, 86 of them have at least 86 citations. In fact, many of his publications have much higher citations. In particular, two of his joint papers with Etzkowitz on a triple helix model of university-industry-government relations have been cited 6357 and 3367 times, way above the h-index of 86. The h-index is very simple in that it tags the

productivity and the citations of a scientist with a single number. Broadly speaking, the higher the h-index, the more likely the scientist has made influential contributions to research.

On the other hand, the simplicity of the h-index also means that it does not represent some of the important aspects of a scientist's productivity or citations. Considering the complexity of citation distributions in reality, it is unlikely that any simplistic indicator can provide a comprehensive coverage of the underlying phenomenon that is so complex and dynamic. Scientists with the same h-index can still differ significantly before and after their the hth most cited publication. Scientists may have significantly different research profiles and yet still have the same h-index. For example, one researcher may have published exactly h papers and each of them has received h citations, which would make his/her h-index to be h. Another researcher may have published much more than h articles, say 10 times of h, but has a small number k of exceedingly influential and highly cited papers, $k \ll h$. As in with Loet Leydesdorff's case, his highest single paper citation is 6357, which is about 74 times of his h-index of 86. Although from the skewed distributions of citations we know that the former scenario is less likely to occur, the diversity within the class of scientists with the same h-index tends to be too large to be reliable for any evaluative purposes. After all, the h-index is biased towards researchers who have a sustained productivity as well as a long-lasting scholarly impact.

The g-Index

Many factors that influence citations may be used to normalize indicators such as the h-index. The academic age t of a researcher can be defined as the number of years since the first publication of a peer reviewed article, Hirsch proposed a normalized h-index m, which is the ratio of h to t. The stability of the m-index has been questioned, especially when the scientist is in his/her earlier career.

The h-index does not preserve any citation information about articles that are in the group of articles above the h citation mark, nor does it tell us anything about the size of the group below h. The h-index divides the publications of a scientist into two groups. One contains articles that have at least h citations, whereas the other contains articles that have fewer citations. Leo Egghe (2006) introduced the g-index as an enhanced modification of the h-index by taking into account the citations of the highly cited group. Similarly as in the h-index, the g-index divides the entire set of articles published articles into two groups using a single number g such that the top g highly cited articles as a whole have at least g^2 citations.

$$\sum_{i=1}^{g} c(a_i) \geq g^2$$

Alternatively, the g-index can be expressed in terms of the average citation of the g top cited articles' citations.

$$g \leq \frac{\sum_{i=1}^{g} c(a_i)}{g}$$

There are numerous ways one can normalize citation-based indicators such as the h-index and the g-index. For example, *Publish or Perish* normalizes the h-index by dividing the original citation counts by the number of co-authors first and then calculates the h-index on the author-normalized citation counts. Given the skewed citation distribution, instead of using the average of the g top cited articles' citations, one may consider using the median of the g citations or define an indicator G using a cumulative density function.

Other Measures

A key criterion of an indicator of scholarly impact of a scientific article should reflect how many researchers it has reached and how many people's thinking and behaviors have been changed. Thus, the number of citations an article has received or the number of citations a journal has received is commonly used measures. At the global level, Fig. 4.15 shows a dual-map overlay visualization of a set of publications on Terrorism research. There two maps in the visualization, hence it is called a dual-map visualization (Chen and Leydesdorff 2014). The map on the left is a set of citing journals organized according to their citing patterns, the map on the right is a set of cited journals positioned according to how similar they are cited. The curves represent citations from a citing journal on the left to a cited journal on the right.

Figure 4.16 shows some of the salient referential connections between clusters of citing and cited journals. For example, articles in this dataset frequently appeared in journals relevant to psychology, education, and health. These articles frequently cited references in similar types of journals. There are 17,276 such instances, which is equivalent to a z-score of 8.423. The strong pathway is visualized as a thick line. Some of the most cited journals are shown in Fig. 4.17.

Figure 4.18 depicts the distributions of citations by year of publication in Scientometrics (2010–2014). As expected, these distributions are strongly skewed towards the lower end of the citation scale. Most articles have zero or few citations, although highly cited articles do exist.

Figure 4.19 depicts the average number of references per paper in Terrorism (1982–2017). The thin solid line in green shows the average number of references per paper of the article type with citations. The dash-and-dot line in green shows the average number of references from articles without citations. Both lines are steadily increasing over time and the solid green line has about 15 references more on average. The thick solid line in blue represents the average number of references

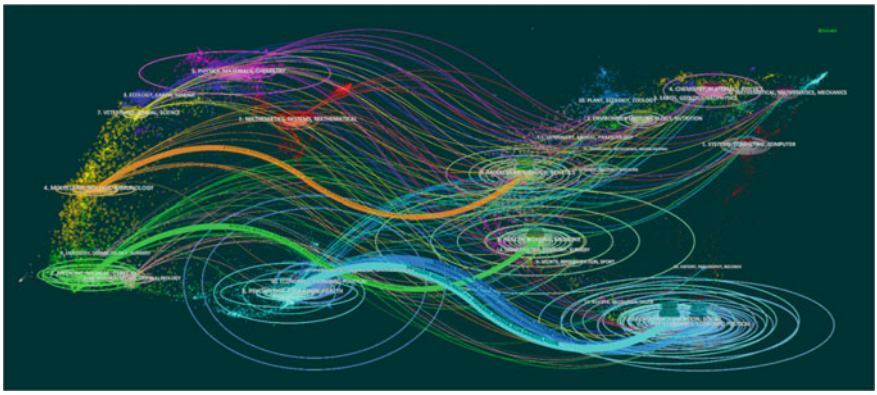

Fig. 4.15 A dual-map overlay visualization of the terrorism2017 dataset (N = 14,656 articles and reviews)

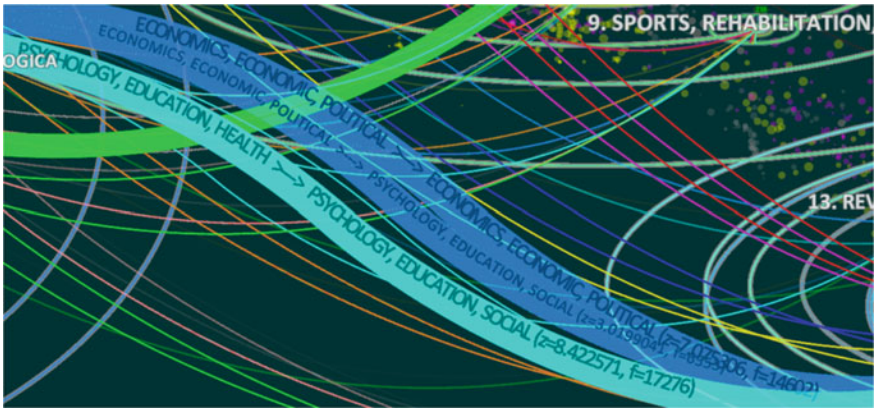

Fig. 4.16 The main field-level citation paths include Psychology|Education|Health to Psychology|Education|Social (z = 8.423, f = 17,276), Economics|Economic|Political → Economics|Economic|Political (z = 7.075, f = 14,602)

from review articles with citations, whereas the dashed line in blue represents the average of references from review articles with no citations. Reviews with citations have cited more references than reviews with no citations. We cannot draw conclusions on any possible causal relations between references and citations. Although some journalists indeed attempted to make more shocking headlines by claiming such relations, we believe one has to examine the nature of citations to avoid picking up the wrong end of the stick. We refer to the number of references and many similar types of indicators as extrinsic factors as opposed to intrinsic ones when one aims to explain the scholarly impact (Chen 2012; Onodera and Yoshikane 2015).

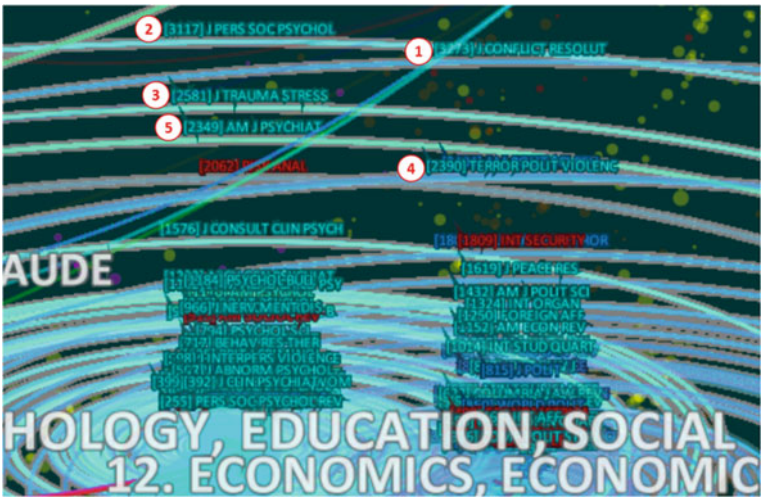

Fig. 4.17 Some of the most cited journals: 1. *Journal of Conflict Resolution*, 2. *Journal of Personality and Social Psychology*, 3. *Journal of Traumatic Stress*, 4. *Terrorism and Political Violence*, and 5. *The American Journal of Psychiatry*

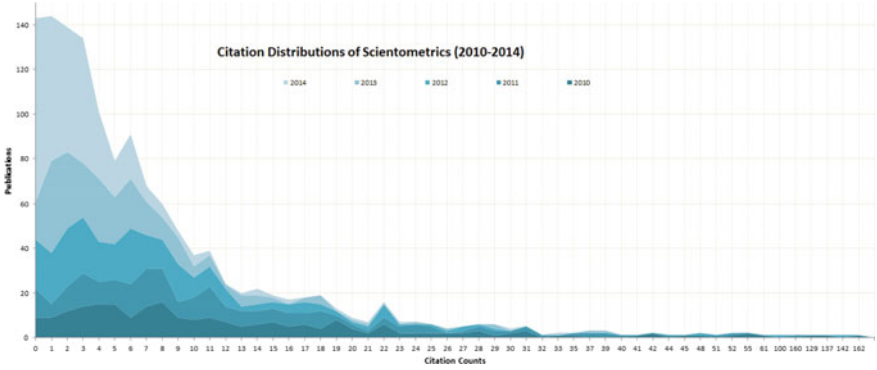

Fig. 4.18 Distributions of citations by year of publication in *Scientometrics* (2010–2014)

Normalization of Metrics

We all know that, to be fair, we should avoid comparing apples with oranges. Similarly, one should only pick on someone of his own size; otherwise, he would be considered either a bully or a coward. In weightlifting, athletes are grouped by their body mass. There are eight male divisions and eight female divisions. Men's weight classes include the 56 kg (123 lb) class, 62 kg (137 lb) class, and the highest 105 kg and over class. Athletes compete with others in the same class. In contrast, swimmers with longer arms have definite advantages over other

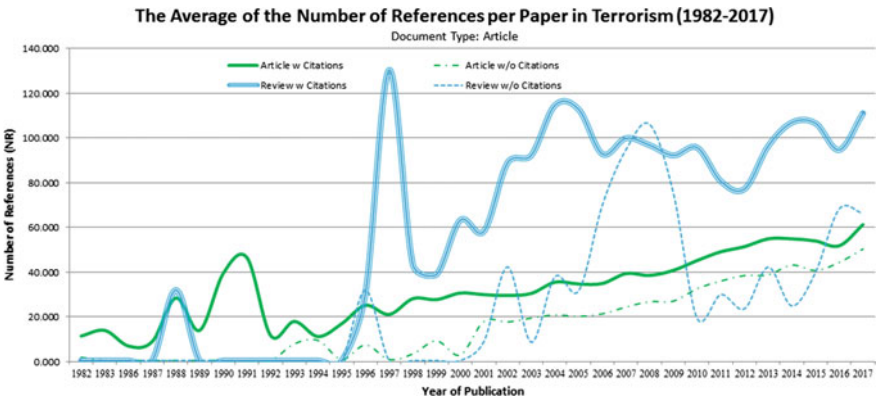

Fig. 4.19 Distributions of the average number of references per paper in Terrorism (1982–2017)

swimmers, but they compete regardless their height. Furthermore, there are four different styles in Olympic swimming: breaststroke, butterfly, backstroke, and free style. Different styles differ in their speed. One would expect it will take a swimmer a longer time to complete 100 m in breaststroke than in butterfly. On the other hand, we wouldn't be surprised if an Olympian swimmer's breaststroke is faster than a high schooler's freestyle. Given all these variabilities, we may still demand answers to questions that may sound like comparing apples with oranges after all. Who is the most powerful weightlifting athlete? Which swimmer's world record is the most remarkable?

Inevitably, scientists often find themselves in similar situations—others would like to compare their performance as a scientist with other scientists' performance, for example, for recruiting, tenure and promotion, and prestigious awards. Strictly speaking, every scientist is unique in numerous and fundamental ways such that comparing scientists based on quantitative measures alone may be even more ridiculous than comparing apples with oranges. In reality, the attraction of quantitative assessments is so strong that we will have to deal with a wide variety of issues along this line of inquiry and practice.

Research indicators, or academic indicators, are numeric figures that can give us a sense of something that maybe otherwise intangible. For example, a researcher's resume routinely includes the number of journal articles publishes, the number of presentations made at international conferences, the total amount of research grants secured, and the number of prestigious awards received. More recent years, researchers include additional indicators such as the number of citations to their publications in the Web of Science, the number of citations on Google Scholar, or relatively more mysterious h-index.

In addition to the evaluation of individual researchers' performance, their productivity and their scholarly impact, groups of researchers, institutions, and nations as well as journals and disciplines are subject to various evaluative assessments in a growing number of countries. It is important to understand the basics of commonly

used indicators of research productivity and scholarly impact, especially their strengths and weaknesses.

Distributions of Citation Counts

The simplest indicator of a scientist's productivity is perhaps the number of articles he/she has published. Suppose one has published 400 articles and the other 200. Then the former clearly has a higher productivity. However, here is the first twist, what if we learn that the 400 publications together have received fewer citations than the 200 publications? In terms of the utility, who is more effective? Should we modify our assessment of the productivity based on the new information? Even if they have received the same number of citations N, the citation per paper rate (CPP) for the former is lower than the latter (N/400 < N/200). The efficiency of the latter is twice of that of the former.

It has been long realized that different disciplines of science may have drastically different citation rates. For example, mathematics is well known for its low citation rate, whereas biomedicine has the reputation of a high citation rate. Thus, being cited by 5 times may be not a big deal for a biomedical scientist, but it probably means a lot more to a mathematician. The differences between mathematics and biomedicine are probably much more profound than that between apples and oranges!

The age of a publication is also a known factor that may significantly influence the amount of citations. The diffusion of information takes time. The longer a publication has been exposed to the scientific community, the more likely it will be noticed and subsequently cited.

Normalization is a term that has been overloaded with multiple meanings. In our context, the term normalization refers to a transformation process that aims to eliminate or reduce the biases due to the heterogeneities between disciplines and between different durations of disclosure. The central idea of normalization is simple: how does the performance of our scientist compare with a *typical* scientist if everything else remains to be equal? As it turns out, in most of the cases it may not as straightforward as we wish to find our *typical* guy.

The distribution of citations is skewed. It means that the average number of citations does not evenly divide the distribution. Rather, one side of the mean may have a lot more instances than the other side. It would be nice and neat if citations are normally distributed. Then we can measure how far away an observed value from the average—the central tendency theory. We would be able to compare our observed value with the average. We would be able to look up the probability of observing a given value and we would be able to see how hard an achievement it might be.

A reference set is the term used by some researchers to refer to the baseline group to be taken into account. Once the performance of the reference set has been taken into consideration, their bias can be minimized or eliminated. In the early

years, the average of citations in a reference set was used in initial attempts to normalize citations. However, it would work nicely only if citations follow a normal distribution.

The skewness of citations, or the skewness of science, is discussed in detail by Seglen (1992). First, the article age contributes to the skewness. The citedness of scientific articles changes with their age. Citations usually peak in the third year after the publication, then citations will decline steadily over time. The decline is considered to do with the obsolescence of the content. Seglen concluded that neither productivity nor citedness can adequately serve as general indicators of scientific quality and that the skewness shown in these indicators are probably in common in other indicators or potential indicators of scientific quality. After all, the evidence is more than sufficient that a small number of scientists contributed a lion share of the major advances of science.

Citation counts are a measure of utility rather than a direct measure of scientific quality. Citations measure the degree of attention from the scientific community. In this sense, citations measure the degree of perturbations to the complex system of scientific beliefs held by scientists as a whole. Direct measurements of scientific quality should characterize the core of scientific advances in terms of the novelty and the potential of transformative change.

Cross-field normalization of citation counts is primarily motivated by the inevitable fact that scientists from different fields of study are subject to quantitative evaluation from time to time. The general idea is to identify the scientific field in which a scientist should be evaluated so that the performance of the scientific field can be used to serve as a baseline reference. Slightly different terminologies have been used to refer to the baseline, including a reference standard or a reference set of publications.

Ideally, if there is a readily available classification system of scientific publication, then it is probably a good idea to consider utilizing the existing classification system. The most widely used such systems is the Subject Categories from the Web of Science. Each article indexed in the Web of Science is assigned with one or more subject category terms, for example, astronomy and astrophysics, artificial intelligence, and psychology. The research of a computer scientist specialized in artificial intelligence should be assessed in this particular context. Similarly, the research of a psychologist should be evaluated with peer researchers in the same subject category of psychology.

In an influential study published in PNAS, Radicchi et al. (2008) focused on the normalization of the citation performance of single publications. Given an article a in a particular field of research F, they considered the average number c_{mean} of citations received by all N articles $b_1, \ldots, b_N$ in the field F published in the same year y, $F(y = \text{year}(a))$, as a normalized citation indicator c_{f} with reference to the particular field. Note that our notations may differ from those in Radicchi et al.'s original paper.

$$c_f(a) = \frac{c(a)}{c_{mean}} = \frac{c(a)}{\frac{\sum_{i=1}^{N} c_i(b_i)}{N}}, \text{ where } b_i \in F(y = year(a))$$

Radicchi et al. utilized the Subject Categories in the Web of Science as the definition of a scientific field. They found the chance of having a particular value of c_f is the same across distinct fields determined by the subject categories for articles published in the same year. More specifically, they found that the rescaled probability distribution $c_{mean}P(c, c_{mean})$ of the relative indicator c_f follows a lognormal distribution with a variance σ^2 of 1.3. If a random variable X has a lognormal distribution, then it means that ln(X) is normally distributed with μ as the mean and σ as the standard deviation. More specifically, X = exp(μ + σZ), where ln (X) = μ + σZ is normally distributed.

What is remarkable about the finding is that a wide variety of subject categories such as allergy, astronomy & astrophysics, biology, mathematics, and tropical medicine appear to have the same property.

A lognormal distribution is defined by the following probability density function (PDF):

$$PDF(x) = \frac{1}{x} \bullet \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

where $ln(x)$ follows a normal distribution, μ is the mean, and σ is the standard deviation. In Radicchi et al.'s study, the equation $\sigma^2 = -2\mu$ reduces the number of fitting parameters to 1. A lognormal distribution with the same mean and the same standard deviation as the one in Radicchi et al.'s paper is shown in Fig. 4.20.

The results obtained by Radicchi et al. is very strong because it suggests that the rescaled lognormal distribution is independent of particular fields of study. On the other hand, when Radicchi et al. experimented with the universal characteristics of citation distributions across scientific fields, their study left out some common and

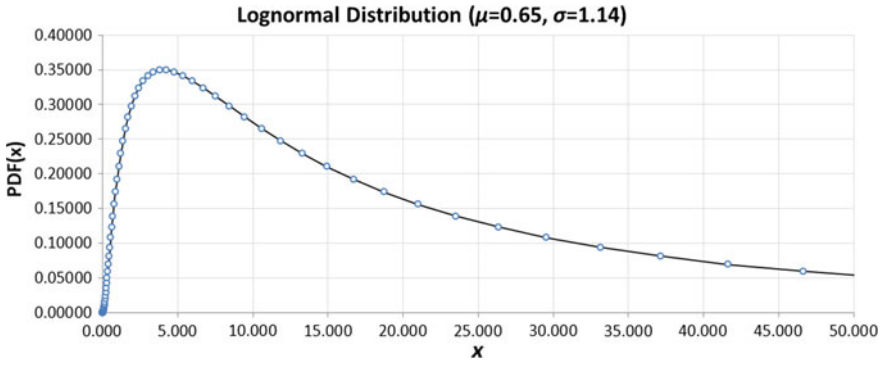

Fig. 4.20 A lognormal distribution

potentially significant categories, notably the multidisciplinary sciences category, which includes the most prestigious journals such as Science, Nature, and PNAS. Furthermore, their calculations exclude uncited articles.

It is now generally agreed that citation counts from different fields should not be directly compared with each other. To a lesser degree, it is also realized that one should be very careful when comparing citations of articles published in different years as well as publications of different types such as original research, review papers, editorials, and letters. In fact, scientometricians have studied a large number of factors that may influence how many citations research articles may get, when they are likely to peak, and how soon they may begin to decay. In a 2012 article on predictive effects of structural variations in a network of cited references, we distinguish factors as intrinsic and extrinsic. Intrinsic factors reflect the semantic and structural characteristics of the underlying scientific activity, whereas extrinsic factors do not have direct connections.

Examples of intrinsic indicators include structural variation metrics such as the ones we developed in our study of predictive effects of structural variations on citations. to measure the transformative potential of an article based on whether and to what extent it introduces novel and potentially groundbreaking links. The modularity change rate, for example, measures the degree to which a newly published article alters the structure of the network of scientific knowledge in terms of the change of modularity scores. Each newly published article brings us a set of references it cites. This set of references casts new lights on the existing network of scientific knowledge, which may be organized with cited references as nodes and co-citation relations as connecting links. The newly casted sub-network may introduce unprecedented links as well as reinforce existing ones with reference to the baseline network. The modularity of a network measures the degree to which the network is modularized. In other word, a network with a high modularity is organized in terms of a number of rather self-contained sub-networks. Interconnections between these sub-networks are minimal. In contrast, a network with a low modularity involves a considerable number of interwoven sub-networks.

Influential Factors on Citations

Researchers have identified some of the major sources of the skewness of science. Onodera and Yoshikane (2015), for example, published a study that systematically investigated several factors affecting citation rates. Ludo Waltman (2016) reviewed the literature on citation impact indicators, including a section on issues concerning normalizing citation-based indicators. In Table 4.13, we group some of the most commonly seen factors of citations in several broadly defined categories. Citation counts may be influenced by various factors about the authors of an article, including the productivity of the author, the academic age of the author, citations the author has received so far, and how the author connects with others in the academic network of collaborators.

Table 4.13 Factors that may influence citations of a scientific publication

Category	Factors on citation counts
Author	Productivity Reputation, citedness Gender Discipline Institution, Country Academic age: the number of years since the publication of the first peer reviewed article, the number of years since the first Ph.D. degree Academic network: eigenvector centrality, betweenness centrality
Article	Citations to date Altmetrics: Downloads, Views, Tweets Accessibility: Open access Visibility: Journal Impact Factor Co-authors: the number of co-authors, their diversity in author attributes Document type: original research, review, letter, etc. Extrinsic properties: the number of pages, the number of figures and equations Exposure: Duration since its publication date Language
References	The number of cited references The diversity of the references in terms of journals and disciplines The novelty of co-cited references
Discipline	The scientific field or fields to which the article belongs
Quality	Significance of research questions Rigor of methodology Clarity of presentation

Quantitative measures such as the citations and the number of cited references are relatively easy to handle. Factors that are of quality in nature are much more challenging to define (Hicks et al. 2015, Zhu et al. 2015). For example, identifying the scientific field that an article belongs too requires a substantial level of domain knowledge even with existing taxonomies of a domain. The significance of research questions requires a good understanding of a subject area, sometimes, more than one. Developing indicators of quality is an ongoing and challenging research in its own right (Ding et al. 2014, Wang et al. 2013).

Improvements of Impact Factors

The Journal Impact Factor is probably one of the most widely used and misused indicators of scholarly impact. In its original form, given a journal J, its impact factor $IF(J)$ is defined as the ratio of the citations to the citable items published in the previous two years c_{-1} and c_{-2} over the total number of citable items s_{-1} and s_{-2} within the same time frame.

$$IF(J) = \frac{c_{-1} + c_{-2}}{s_{-1} + s_{-2}}$$

The calculation of the impact factor over a two-year time span can be easily extended to a 5-year span or an arbitrary *k*-year impact factor.

$$IF_k(J) = \frac{\sum_{i=-k}^{-1} c_i}{\sum_{i=-k}^{-1} s_i}$$

Loet Leydesdorff is among the first to argue that the calculation should be done in a different order (e.g., Leydesdorff 2012; Leydesdorff et al. 2011). Instead of summing up the citations and citable items separately first and then taking the ratio, a more reasonable calculation should take the average citation per citable items in each year first and then calculate the average over the number of years.

$$NIF_k(J) = \frac{\sum_{i=-k}^{-1} \frac{c_i}{s_i}}{k}$$

The new impact factor (*NIF*) then becomes a k-year moving average of the annual citation rate. The original impact factor is the ratio of two averages, whereas the *NIF* is the average of citation ratios. Which one is more appropriate? What difference does it make? These questions are in fact part of a more profound debate in cross-field normalization.

Earlier citation normalization such as the Crown Indicator are calculated as the ratio of the mean of observed citation rates (OCR) over the mean of expected citation rates (ECR), which resemble to the way the original IF is calculated.

$$\frac{Mean(OCR)}{Mean(ECR)} = \frac{\frac{\sum_{i=1}^{n_{obs}} c_i}{n_{obs}}}{\frac{\sum_{j=1}^{n_{exp}} C_j}{n_{exp}}}$$

where $\{c_i\}$ are observed citations and $\{C_j\}$ are expected citations computed from a reference set such as all the publications from a field, i.e. biology or mathematics. In contrast, the more recently recommended citation normalization is the mean of the ratio of OCR to ECR:

$$Mean\left(\frac{OCR}{ECR}\right) = \frac{\sum_{j=1}^{n} \frac{c_i}{C_j}}{n}$$

The Mean(OCR)/Mean(ECR) is a division of two means. Using the Mean(OCR/ECR) has an advantage over the former—it comes with a standard deviation, which is additional information that is not available from the division of two means. Researchers have recognized the advantages of replacing the rate of averages with the average of rates.

The choice of using the mean of observed citation counts or the mean of the expected citation counts has also been subject to criticisms on the ground that the mean is no longer representative in a skewed distribution, which citation distributions typically fall into this category. Instead, the median would be a better choice. An ideal indicator should reflect the shape of the distribution and it should provide a metric that is independent from fields of study, the age of the article, and other major factors.

One of the most appealing indicators proposed in recent years is perhaps the approach that ranks articles on a percentile scale. It is proposed by Leydesdorff et al. (2011). The rank of an article is defined as the percentage of papers in the reference set that have citations fewer than the citation of the paper. The percentile is then rounded as an integer as the rank. Most cited 1% papers on the top of Mount Kilimanjaro should belong to the 99 percentile class. Given an article a, the probability that it belongs to the 99 percentile class is way below one in a million, considering that the size of the Web of Science as the reference set is about 50 million, depending on particular subscriptions.

The rank of articles in the *k*th percentile class can be expressed as the cumulated relative frequencies p(r) weighted by their corresponding rank r:

$$R(k) = 1 \cdot p(1) + 2 \cdot p(2) + \ldots + k \cdot p(k) = \sum_{r=1}^{k} r \cdot p(r)$$

where f_r is the number of articles in the rth bin (there are 100 bins; one for each percentile class), $p_r = f_r/n_r$ and $n_r = \sum_i f_i$. The maximum weight is 100, which appears in R(100). The minimum weight is 1. Leydesdorff et al. gave an example of R(6) = 1*0.5 + 2*0.25 + 3*0.15 + 4*0.05 + 5*0.04 + 6*0.01 = 1.91.

Note that the range of R is not [0, 1]. One will need additional information to tell whether 1.91 is large or small. A further improvement can scale the range to the unit interval [0, 1] so that it is instantly clear about the position of 0.89 on a scale of [0, 1].

$$R(k) = \frac{\sum_{r=1}^{k} r \cdot p(r)}{\sum_{r=1}^{100} r \cdot p(r)}$$

More generally, in addition to work with percentiles, one can extend it to an arbitrary number of classes, for example, with 1000 bins or 100,000 bins, especially when dealing with a large number of articles at the disciplinary level. The more finer sliced bins we use, the more accurate the indicator tracks the underlying distribution. This line of reasoning leads to an ideal indicator I in an integral form, which suggests that when necessary, one can use finer grained bins to improve the accuracy of the indicator with reference to the underlying distribution. Here the p(x) is the probability density function.

$$I(c) = \int_{-\infty}^{c} p(x)\mathrm{d}x$$

The value of the cumulative density function at an arbitrary level of citation count c is between 0 and 1. It reaches the maximum of 1 when the probabilities of all sorts of scenarios are accounted for. In this way, the metric of quality is both intuitive and field-independent. Questions concerning quantifying scientists' performance can be answered in terms of the cumulative probability of observing a particular level of performance. If the performance of a mathematician has the cumulative probability of 0.90, then we know that this is a better performance of a molecular scientist with a cumulative probability of 0.80 in terms of scholarly publications.

Furthermore, given an article with citations of c, the cumulative density function will return a value between 0 and 1. The value can be considered as a rarity measure. The rarer a citation frequency, the harder it is to achieve and thus the more excellent it is.

In summary, cross-field normalization of citation-based indicators of scholarly impact has produced many indicators. However, researchers continue to refine the normalization procedures to reduce various biases that may be originated from the delineation of disciplinary boundaries or from the way to estimate expected levels of citation with reference to year of publication as well as relevant fields. Researchers have identified a large number of potential factors (See Table 4.13). We need to further develop our understanding of the magnitudes of the effects of these factors and how they interact at multiple levels of granularity. Most normalizations focus on a very small number of factors. It remains to be found to what extent existing normalizations preserve the order of articles in terms of their relative positions in their own crowd. Normalization should transform the values of apples and oranges into numbers within [0, 1].

Science Mapping

In this section, we will illustrate some of the important concepts with a collection of 17,731 papers on science mapping. A systematic review of science mapping published in 2017 is based on this dataset (Chen 2017). 17,721 of the 17,731 records are successfully loaded into a database. The following examples are based on the 17,721 records (Fig. 4.21). The dataset contains 14,794 articles (83.48%), 1861 proceeding papers, 1034 review, and a few items of other types such as book review, editorial, and book chapter. A copy of the dataset is downloadable from the ResearchGate project of the book.

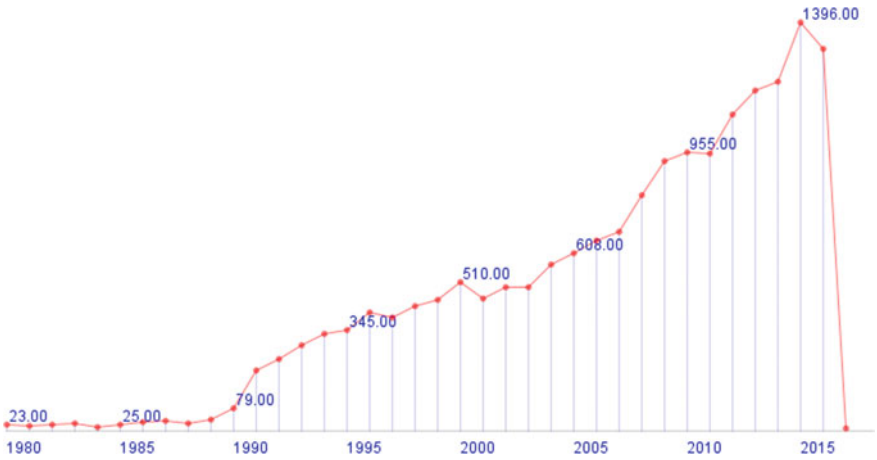

Fig. 4.21 The number of records in the dataset of Science Mapping (1980–2017)

Exploring the Science Mapping Dataset with CiteSpace's Database

We first loaded the dataset to a MySQL database on the localhost through an interface provided by CiteSpace. We demonstrated the example with MySQL queries such that interested readers can practice with their own datasets.

Table 4.14 illustrates the information stored in the Articles table of the wos database regarding the 2006 publication on CiteSpace. Each record from the Web of Science has a unique ID such as WOS:000234932600008. Similarly, a record from Scopus can be converted to the same format. The Scopus ID contains the DOI of the article, which appears to make the rest of the long string redundant. The information from the two sources has some discrepancies, which are highlighted in the table. For instance, the author name is Chen, Cm in the Web of Science, but Chen, C in Scopus. The journal title is abbreviated slightly different. More interestingly, citation counts differ substantially: 331 in the Web of Science and 503 in Scopus. A quick inspection of citing articles' sources reveals that many of the Scopus records are from conferences such as ISSI 2007 (8 papers), 2009 (8), 2013 (4), and 2015(3). These conferences, to our best knowledge, are not included in the Web of Science. This discrepancy in citation counts underlines practical issues one should consider for mixing citation records from distinct sources.

Table 4.15 lists the index terms assigned to the article. The author of the article did not provide any keywords. The index terms are algorithmically assigned as so-called KeywordPlus in the Web of Science. The keywords assigned by Scopus such as knowledge domain visualization and scientific literature are more accurate than the keywords under the Web of Science. The nearest term from the Web of

Table 4.14 Information stored in the Articles table of the wos database

Field	Example from the Web of Science	Example from Scopus (Format Converted)
id	433075	258910
uid	WOS:000234932600008	Scopus:2-s2.0-33644531603&doi = 10.1002%2fasi.20317
project	sciencemapping17731	scopus651
author	Chen, Cm	Chen, C
title	CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature	CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature
abstract	This article describes the latest development ……	This article describes the latest development ……
source	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY	Journal of the American Society for Information Science and Technology
j9	J AM SOC INF SCI TEC	J AM SOC INF SCI TECHNOL
volume	57	57
issue	3	3
nr	61	61
bp	359	359
ep	377	377
page	359–377	359–377
dt	Article	Article
doi	10.1002/asi.20317	10.1002/asi.20317
year	2006	2006
citations	331	503

Science is domain visualization. Other keywords on the list are more related to the case studies included in the article than CiteSpace as a tool as the focus of the paper.

The type sc in the last two rows of the table stands for Subject Category (SC). Two subject categories are assigned to the paper, namely Computer Science and Information Science & Library Science. It is not surprising that many articles published in the journal involve these two subject categories. In the Science Mapping dataset, Computer Science is the second largest subject category, whereas Information Science & Library Science is the third largest one.

Major Subject Categories in Science Mapping

In this section, we will explore several aspects of the Science Mapping with reference to the need for cross-field normalization and cross-time normalization. Corresponding MySQL queries are included for interested readers to replicate the results if they wish.

Table 4.15 The same article is indexed differently in different sources

Web of Science			Scopus		
id	Keyword	Type	id	Keyword	Type
2901483	Triassic mass extinction	id	1223053	Knowledge domain visualization	id
2901484	Domain visualization	id	1223054	Scientific literature	id
2901485	Terrorist attack	id	1223055	Algorithm	id
2901486	Science	id	1223056	Computer programming language	id
2901487	Paradigm	id	1223057	Information retrieval	id
2901488	Knowledge	id	1223058	Information science	id
2901489	Network	id	1223059	Research	id
2901490	City	id	1223060	Natural sciences computing	id
2901491	September-11	id			
2901492	Technology	id			
2901493	Computer Science	sc			
2901494	Information Science & Library Science	sc			

The number of records distributed per year in the dataset is shown in Fig. 4.21. The volume steadily increases. In 2015 alone, there are 1396 publications in the dataset. In 2000, the number of 510. The plot is generated in CiteSpace with the following MySQL query.

```
SELECT year, count(year)
FROM articles
WHERE project='sciencemapping17731'
GROUP BY year
ORDER BY year
```

Science Mapping is a field of interdisciplinary research. The dataset involves 149 distinct Web of Science Subject Categories. The Subject Categories of each record are stored in the keywords table. The following MySQL query finds the number of distinct subject categories. Each subject category is considered as a field of study. Researchers commonly identify the fields of study in terms of the Subject Category classification system.

```
SELECT count(distinct(keyword))
FROM keywords
WHERE project='sciencemapping17731'
AND type='sc';
```

The top 10 largest subject categories in the science mapping dataset are shown in Table 4.16. The largest subject category is Engineering, which has 4387 publications (24.8% of the entire dataset). The second largest one, Computer Science, has

Table 4.16 The number of articles distributed in subject categories

Publications	% of 17,721	Keyword
4387	24.7559	Engineering
3467	19.5644	Computer Science
2075	11.7093	Information Science & Library Science
1080	6.0945	Physics
1076	6.0719	Business & Economics
708	3.9953	Environmental Sciences & Ecology
623	3.5156	Telecommunications
605	3.4140	Optics
599	3.3802	Science & Technology—Other Topics
538	3.0359	Materials Science

3467 publications. The third one, Information Science & Library, has 2075 publications.

```
SELECT count(*), count(*)/17721, keyword
FROM keywords
WHERE project='sciencemapping17731'
AND type='sc'
GROUP BY keyword
ORDER BY count(*) DESC limit 10;
```

One may not anticipate to see Engineering appearing as the largest subject category in this dataset; after all, Science Mapping should be more closely related to computer science and information science. The following query lists the top 20 most frequent keywords assigned to Engineering papers in this dataset.

```
SELECT count(*), k2.keyword
FROM
    keywords AS k1,
    keywords AS k2
WHERE
    k1.project='sciencemapping17731' AND
    k2.project='sciencemapping17731'
AND k1.uid=k2.uid
AND k1.type='sc' AND k2.type!='sc'
AND k1.keyword='Engineering'
GROUP BY k2.keyword
ORDER BY count(*) DESC
LIMIT 20;
```

As shown in Table 4.17, Engineering papers are related to time-domain analysis, frequency-domain analysis, scattering, electromagnetic scattering, and information visualization. Although information visualization is semantically connected to

Table 4.17 Top 20 keywords associated with papers from the Engineering subject category

Engineering		Computer Science		Information Science	
Count (*)	Keyword	Count (*)	Keyword	Count (*)	Keyword
488	Time-domain analysis	635	**Citation analysis**	748	**Citation analysis**
268	Frequency-domain analysis	594	Information visualization	525	**Science**
258	Time domain analysis	481	**Science**	217	**Bibliometrics**
255	System	234	Visualization	196	**Journal**
253	Frequency domain analysis	223	**Network**	193	**Impact**
221	Model	198	**Bibliometrics**	170	**Network**
189	Design	191	System	166	**Indicator**
162	Simulation	189	Model	150	**Citation**
109	Algorithm	167	**Impact**	141	**Publication**
98	Scattering	157	**Information**	131	**Information-science**
91	Information visualization	149	**Journal**	118	**Co-word analysis**
89	Identification	147	**Indicator**	114	Scientometrics
86	Performance	146	Design	112	Library
73	Stability	141	**Citation**	109	**Information**
71	Vibration	117	**Pattern**	108	Impact factor
70	Equation	116	**Co-word analysis**	104	**Pattern**
67	Domain analysis	115	**Publication**	100	Index
65	Dynamics	108	Visual analytics	95	h-index
62	Electromagnetic scattering	99	Knowledge	93	Web
61	Wave	97	**Information-science**	90	Cocitation analysis

science mapping, the inclusion of papers on time-domain analysis and frequency-domain analysis appears to be a side effect of the set of queries used to retrieve the 17,731-record dataset from the Web of Science. In particular, domain analysis is one of the sub-topics in Science Mapping. Apparently, domain analysis is a term that is also used in Engineering for a completely different subject. When using CiteSpace, our advice to how to handle such unanticipated and potentially irrelevant topics is to proceed to the network analysis stage without attempting to eliminate the potentially irrelevant records. There are at least two good reasons for deferring any actions to eliminate any records prematurely:

The suspicious irrelevancy at this stage is based on our current knowledge. If we conclude the irrelevancy without further investigation, we may lose the opportunity to learn anything new from the process. After all, there may exist profound connections that we are simply not aware of.

The best time to eliminate irrelevant data is probably after we have a chance to inspect the resultant network model. It is much easier to identify an isolated sub-network in a visualization of the network than try to determine the relevancy from the dataset of such complexity.

The top 20 keywords for Computer Science and Information Science are quite consistent, including common ones such as citation analysis, science, bibliometrics, impact, network, and indicator. Common keywords are highlighted in the table. Unique keywords in Computer Science include information visualization, visualization, design, and visual analytics, whereas unique keywords in Information Science include scientometrics, h-index, and cocitation analysis.

Many publications are indexed with multiple subject categories. For example, there are 455 publications in common between Engineering and Computer Science, 1556 shared publications between Computer Science and Information Science. Interestingly, while Engineering, Physics, Telecommunications, Materials Science, and Environmental Science and Technology overlap one another, Information Science does not overlap with any of them within this dataset. To compute the number of overlapping records between two subject categories, one can use the following query by substituting K1.keyword and K2.keyword accordingly.

```
SELECT count(*)
FROM
    keywords AS K1,
    keywords AS K2
WHERE
    K1.project='sciencemapping17731' AND
    K2.project='sciencemapping17731'
AND K1.type='sc' AND K2.type='sc'
AND K1.uid=K2.uid AND
    K1.keyword='Information Science & Library Science'
AND
    K2.keyword='Computer Science';
```

The total number of papers in Information Science & Library Science is 2075, apart from 1556 papers that are jointly indexed as Computer Science papers, there are only 519 papers that do not share the Computer Science category. This is an indication of the role of computer science in Science Mapping.

Citation Distributions

Based on our earlier discussions, one would expect that citation rates are field-dependent as well as time-variant. One would also expect that the number of references cited by an article varies across distinct subject categories. Using the query below, we can find that the average of citations of the dataset is 16.79, the minimum citations is 0, and the maximum is 1547.

```
SELECT avg(citations), min(citations), max(citations)
FROM articles
WHERE project='sciencemapping17731';
```

Figure 4.22 shows a log-log plot of the frequencies of citations per paper in Science Mapping. Citation counts are log-transformed, so are the frequencies of citations. Since *log*(citations) is not defined for zero citations, a common practice is to add 1 to the citation count of each paper. As expected, papers with zero citations are most common, whereas highly cited papers are increasingly unusual.

The total number of references cited by the 17,731 articles is 672,899, of which 508,564 are distinct. On average, each publication in the dataset has 37.95 references. Publications in the dataset received a total of 297,529 citations across publications indexed in the Web of Science. On average, each paper has a citation count of 16.78.

In addition to the average over the entire dataset, to what extent does a particular subject category differ from the overall dataset? Using the following query, we can find the average number of citations and cited references specifically for a particular subject category.

```
SELECT
      avg(citations),
      avg(nr),
      keywords.year
FROM
      articles,
      keywords
WHERE
      articles.project='sciencemapping17731' AND
      keywords.project='sciencemapping17731' AND
      articles.uid=keywords.uid AND
      keywords.keyword='Computer Science'
GROUP BY keywords.keyword, keywords.year;
```

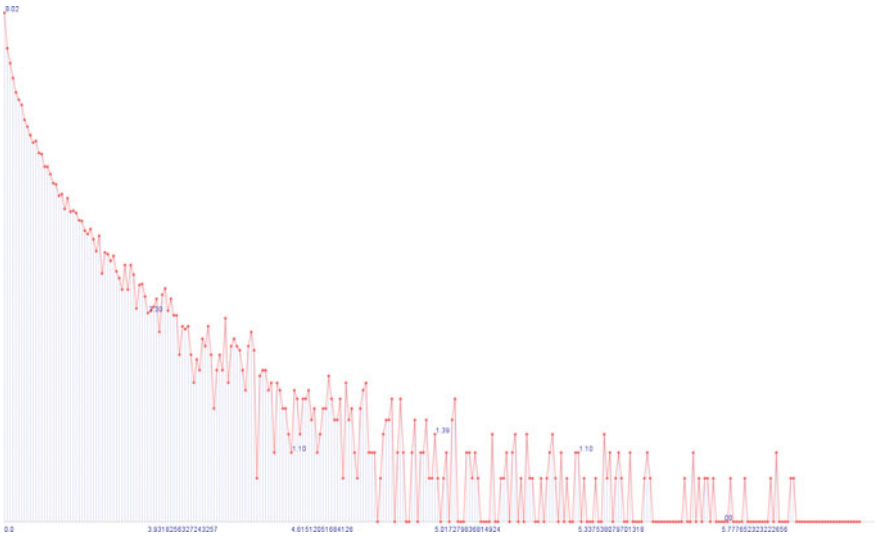

Fig. 4.22 A log-log plot of the frequencies of citations per paper in Science Mapping (1980–2017)

The average number of citations per paper and the average of references per paper of the largest four subject categories show that Engineering and Physics papers have an average of citations per paper about 12, whereas Computer Science, Information Science & Library Science papers have a citation count of 16. Furthermore, for papers in Information Science & Library Science, they have 42 references on average (Table 4.18).

Figure 4.23 illustrates the differences between the four largest subject categories in the science mapping dataset in terms of the average number of references per paper and the average number of citations per paper. The curves in green plot the average number of citations per paper, whereas those in red represent the average number of references per paper. Overall, the red curves show an upward trend. It means that the average number of references per paper is increasing over the years regardless of subject categories. There are a few outliers of papers in Information Science & Library Science. In 1999, Wilson CS for example published a paper that cited 491 references. In 2004, Phillips LI cited 400 references in a single paper. More recently, Guimaraes cited 346 references in a paper and Waltman cited 342. These papers are review papers. Engineering and Physics papers in this dataset have a lower average number of cited references per paper, whereas Computer Science and Information Science have about 10–15 more references on average. The average is steadily increasing for both groups of subject categories. The growth rates are the same because the four lines are essentially parallel to one another.

The lines representing the average numbers of citations are more complex than their reference counterparts, although they diminish towards the present time because recently published papers are yet to receive their citations. Citations of Engineering papers are relatively smooth over the years. In contrast, citations of Information Science & Library Science fluctuated over time, but their citation average is higher than that of Engineering. The earliest outliers include a 1981 paper by Howard White with 406 citations and a 1989 paper by Macroberts. Other prominent papers include Callon, Small, and Chen from Information Science. Papers by Holten, Shneiderman, and Bostock respectively are from the subject category of Computer Science, more precisely, from information visualization and visual analytics.

Table 4.18 The average number of citations per paper and that of references per paper are both field-dependent

Subject category	Papers	Average (Citations)	Average (References)
Overall	17,721	16.7896	37.9718
Engineering	4387	12.3855	24.8849
Computer Science	3467	16.1554	39.5953
Information Science & Library Science	2075	16.4308	42.1667
Physics	1080	12.0820	27.7991

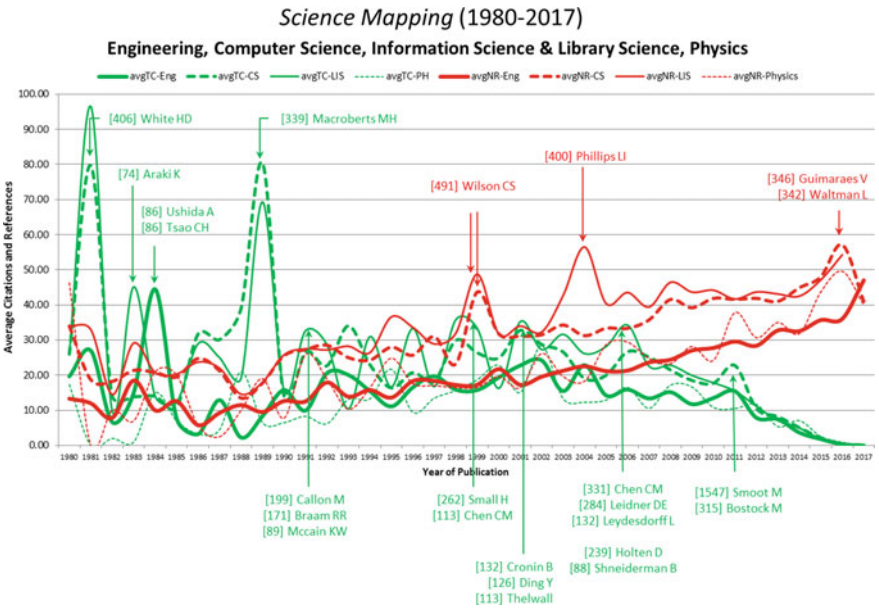

Fig. 4.23 Trends of the number of references and the number of citations of the four largest subject categories

The following query searches for most highly cited papers in a particular year from a specific subject category. Using it along with the plots such as the one shown in Fig. 4.23, one can identify landmark works in Science Mapping as well as general trends in terms of the average number of citations and the average number of references.

```
SELECT
    citations, author, doi, k.keyword, a.year
FROM
    articles AS a, keywords AS k
WHERE
    a.project='sciencemapping17731' AND
    k.project='sciencemapping17731' AND
    a.uid=k.uid AND
    (k.keyword='Engineering' OR k.keyword='Computer Science') AND
    (a.year=1981 OR a.year=1983 OR a.year=1984 OR a.year=1989)
ORDER BY citations DESC LIMIT 30;
```

Citation Normalization Over Time

Articles that are published earlier tend to have a higher number of citations on average. In order to remove or reduce the biases due to how long a paper is available to potential citers, the age of publication should be taken into account.

Figure 4.24 depicts the cumulative citation density function. For instance, given an article published in 2000, the probability that the article has no more than 5 citations (s5) is much lower than the probability that a 2015 article. In other words, a 2000 paper is more likely to get more than 5 citations than a 2015 paper. The formula in the figure suggests how one may estimate a probability in terms of relative frequencies. For example, c_0 is the number of items that have a zero citation, c_1 is the number of papers with citation counts of 1, and so on. Thus $i^* c_i$ is the sub-total of the citations corresponding to c_i. If there are $c_5 = 11$ papers with $i = 5$ citations each, these 11 papers collectively received 55 citations. If the entire set of publications is allocated to 100 evenly divided bins, then this method is very close to the percentile indicator proposed by Loet Leydesdorff and his coauthors. Furthermore, our indicator has two additional advantages:

The percentile-based indicator is an approximation to the cumulative citation density function in its integral form. Realizing its connection to the integral form, one can easily improve the approximation by using an arbitrarily large number of bins. In effect, we are taking the limit of the discrete sum of the citations over bins. With a sufficiently large number of data points and a sufficiently large number of bins, the estimate can be arbitrarily close to the integral value.

Our indicator is scaled to [0, 1], which makes it independent of its range and thus easy to understand and compare with other fields. Instead of wondering where a scientist with an indicator of 1.91 would be positioned on an irregular scale, The unit range of [0, 1] simplifies the interpretation and comparison.

Figure 4.25 depicts the probabilities of articles published in a particular year having c citations between 2000 and 2015 in Science Mapping. The citation probability distributions of articles published in the first 11 years (2000–2010) resemble to normal distributions with the highest probability is around p50, which is the middle of the [0, 100] scale and near-to-zero probabilities towards both ends. The probability curves of articles published in the recent five years (2011–2015) are increasingly higher towards the lower end of the citation scale. It appears that, in general, the peak of a citation distribution steadily shifts from left to right and the overall distribution is stabilized approximately after five years of publication. We suspect that the rate of the settlement is likely to be field-dependent.

In addition to the fluctuations of citation probability, we smooth the citation probabilities with 5-year average citation probabilities between 2000 and 2015. As shown in Fig. 4.26, the trends become more apparent—citation probability distributions are gradually shifted from low-citation probabilities to average—and higher-citation probabilities. The citation probability distribution of articles published in the recent five years has a substantial weight on the left, i.e. the probability of having few citations is relatively high. The two citation distributions of articles

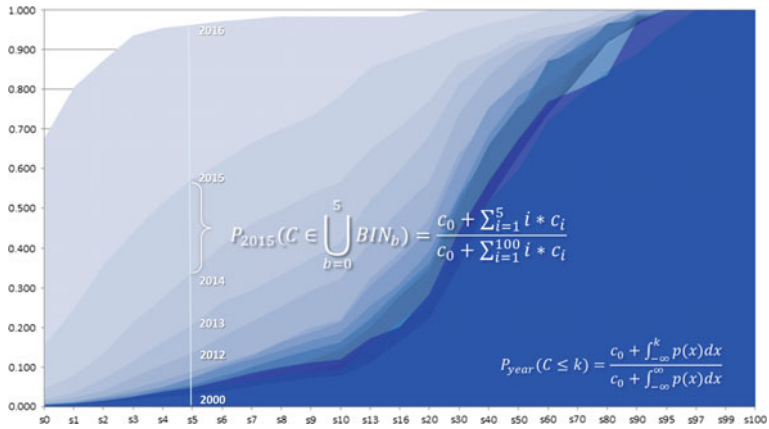

Fig. 4.24 Cumulative relative citations by year of publication

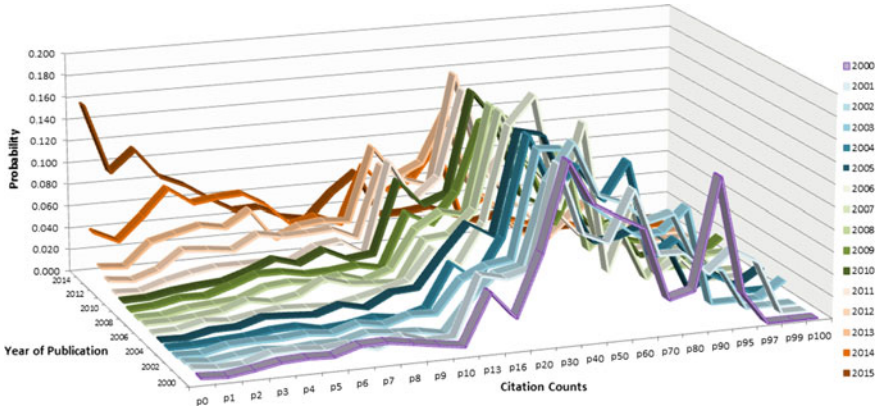

Fig. 4.25 Probabilities of articles published in each year having citations c in Science Mapping

published more than five years ago overlap considerably with one another, suggesting a relatively stable distribution. Normalizing citations over time is reasonably reliable for publications more than 5 years old. In contrast, citation probabilities fluctuated considerably more with articles published less than 5 years old. The key to citation normalization over time is to account these factors.

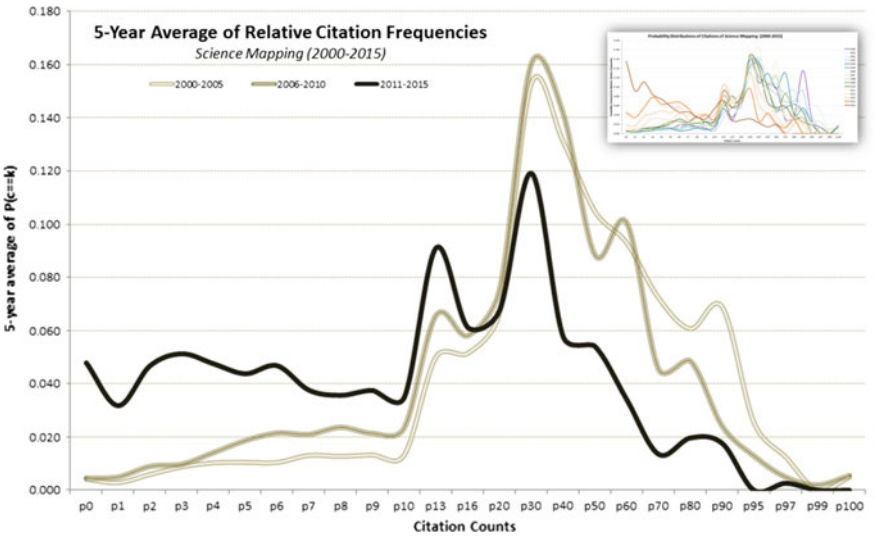

Fig. 4.26 5-year moving average of citation probabilities

Summary

Citation-based indicators should be normalized in terms of the fields of study involved, the year of publication. There are distinct advantages of utilizing standard cumulative citation probability functions as opposed to the development of indicators that may not share the universality in terms of their interpretability. More importantly, the wide variety of indicators should be taken into account collectively along with qualitative analyses of science to serve the purposes of research evaluation as well as learning the state of the art of scientific research.

References

Banerjee S, Pedersen T (2002) An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Gelbukh A (ed) CICLing 2002, LNCS 2276. Springer, Heidelberg, pp 136–145

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Sociol 2(1):113–120

Bookstein A, Klein ST, Raita T (1998) Clumping properties of content-bearing words. JASIS 49 (2):102–114

Chen C (2006) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J AM SOC INF SCI TEC 57(3):359–377. doi:10.1002/asi.20317

Chen C (2012) Predictive effects of structural variation on citation counts. J Am Soc Inform Sci Technol 63(3):431–449

Chen C (2014) The fitness of information: quantitative assessments of critical evidence. Wiley, Hoboken
Chen C (2017) Science mapping: a systematic review of the literature. J Data Inf Sci 2(2):1–40
Chen C, Leydesdorff L (2014) Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. J Assoc Inf Sci Technol 65(2): 334–351
Deerwester S, Dumais T, Landauer T, Furnas G, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6): 391–407
Ding Y, Rousseau R, Wolfram D (eds) (2014) Measuring scholarly impact: methods and practice. Springer, Heidelberg. doi:10.1007/978-3-319-10377-8
Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Comput Linguist 19(1):61–74
Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152. doi:10.1007/s11192-006-0144-7
Freeman LC (1977) A set of measuring centrality based on betweenness. Sociometry 40:35–41
Fries CC (1952) The structure of English. Harcourt Brace, New York
Hicks D, Wouters P, Waltman L, Rijcke Sd, Rafols I (2015) Bibliometrics: The Leiden Manifesto for research metrics. Nature 520(7548):429–431. doi:10.1038/520429a
Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569–16572. doi:10.1073/pnas.0507655102
Hirst G, St-Onge D (1998) Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum C (ed) WordNet: an electronic lexical database. The MIT Press, Cambridge, MA
Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of international conference research on computational linguistics (ROCLING X), Taiwan
Kleinberg J (2002) Bursty and hierarchical structure in streams. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, pp 91–101
Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. In: Fellbaum C (ed) WordNet: an electronic lexical database, Chapter 11. The MIT Press, Cambridge, MA
Leontief WW (1941) The structure of American economy, 1919–1929. Harvard University Press
Lesk M (1986) Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on systems documentation (SIGDOC'86). Toronto, Ontario, Canada, pp 24–26
Leydesdorff L (2012) Alternatives to the journal impact factor: I3 and the top-10% (or otp-25%) of the most-highly cited papers. Scientometrics 92:355–365
Leydesdorff L, Bornmann L, Mutz R, Opthof T (2011) Turning the tables on citation analysis one more time: principles for comparing sets of documents. J Am Soc Inform Sci Technol 62 (7):1370–1381. doi:10.1002/asi.21534
Lin D (1998) An information-theoretic definition of similarity. In: ICML '98 Proceedings of the fifteenth international conference on machine learning, pp 296–304, 24–27 July 1998
Miller GA, Charles WG (1991) Contextual correlates of semantic similarity. Lang Cogn Process 6 (1):1–28
Onodera N, Yoshikane F (2015) Factors affecting citation rates of research articles. J Assoc Inf Sci Technol 66(4):739–764
Piffer D (2012) Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. Thinking Skills Creativity 7(3):258–264. doi:10.1016/j.tsc.2012.04.009
Rada R, Bicknell E (1989) Ranking documents with a thesaurus. JASIS 40(5): 304–310
Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. Proc Natl Acad Sci 105(45):17268–17272

Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy
Seeley JR (1949) The net of reciprocal influence: a problem in treating sociometric data. Can J Psychol 3:234–240
Seglen PO (1992) The skewness of science. J Am Soc Inf Sci 43(9):628–638
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:3379–3423
Shwed U, Bearman PS (2010) The temporal structure of scientific consensusformation. American Sociological Review 75(6):817–840
Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352
Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342 (6154):127–132
Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics. Las Cruces, New Mexciao, pp 133–138, 27–30 June 1994
Zafarani R, Abbasi MA, Liu H (2014) Chapter 3, Network measures. In: Social media mining: an introduction. Cambridge University Press. http://dmml.asu.edu/smm/chapters/SMM-ch3.pdf
Zhu X, Turney P, Lemire D, Vellino A (2015) Measuring academic influence: not all citations are equal. J Assoc Inf Sci Technol 66(2):408–427