JSCIRES REVIEW ARTICLE

Text mining for science and technology: A review – Part II-citation and discovery

Ronald N. Kostoff*

School of Public Policy, Georgia Institute of Technology, 13500 Tallyrand Way, Gainesville, VA 20155, USA

ABSTRACT

This second part of a two-part review addressed three complementary components of text mining: Citation scientometrics, seminal literature reviews (SLR), and literature-related discovery and innovation (LRDI). All three have at their core the development of very comprehensive and precise queries for retrieving the data of interest. For any literature of interest, the citation scientometrics approach analyzes in detail the papers that cite the literature of interest (citation mining), and/or the papers that are cited by the literature of interest. The SLR uses the highly-cited references in a retrieved literature of interest to map out the intellectual heritage of that literature. The LRDI integrates (a) discovery generation from disparate literatures with (b) the wealth of knowledge contained in the prior art to (c) potentially solve technical problems that appear intractable. The review highlights each of the approaches drawing from studies undertaken by author and his research group.

Keywords: Citation analysis, citation mining, citation normalization, literature-related discovery and innovation, literature review, scientometrics

OVERVIEW

This is the second part of a two-part review that addresses three major text mining sub-divisions: Characterization; seminal literature review (SLR); literature-related discovery and innovation (LRDI). Part I, published in the inaugural issue of The Journal of Scientometric Research (Kostoff 2012a),[1] focuses on characterization, mainly its non-citation components. Part II, published in this second issue of the Journal of Scientometric Research, focuses on the citation component of characterization, the citation-based SLR, and the citation-enabled LRDI.

Characterization is the assignment of metrics to the technical literature of interest to identify patterns that will increase understanding of the topical matter. Of interest in the present paper are metrics related to the citation network associated with one or more selected articles. While isolated metrics may have specific uses, the key challenges are to identify "signatures", or combinations of metrics that provide unique insights into technology literatures or to countries' technology portfolios.

SLR presents the intellectual heritage of technical literature, mainly by identifying the most highly cited documents in that literature, and LRDI generates discovery and innovation by linking disparate literatures to produce value-added concepts. More detailed definitions of SLR and LRDI can be found in Part 1 of this review (Kostoff 2012a).[1]

***Address for correspondence:**
E-mail: rkostoff@gmail.com

Access this article online

Quick Response Code:

Website: www.jscires.org

DOI: 10.4103/2320-0057.115862

ANALYSIS

Citation Scientometrics

Case study 1: Sandpile vibration dynamics

Most citation studies focus on counts of citations. The goal of this study was to examine the characteristics of documents that cite one or more selected documents, and

identify objectively some of the impacts of basic research on the R and D community. Use of the total citation mining process could help answer questions such as:

- What types of people and organizations are citing the research outputs; is this the desired target audience?
- What development categories are citing the research outputs?
- What technical disciplines are citing the research outputs?
- What are the relationships between the citing technical disciplines and the cited technical disciplines?

In addition to scientometric analysis of the citing papers, text mining of the cited and citing papers was performed to supplement the information derived from the semi-structured field bibliometric analyses. Technical thrusts of the cited and citing papers, and the relationships among those thrusts, were identified and related to the purely bibliometric quantities. Text mining illuminated the thematic relationships that existed between the cited paper literature and the citing paper literature, and provided insights of knowledge diffusion to intra-discipline research, advanced intra-discipline development, and extra-discipline research and development. The addition of text mining to citation bibliometrics could make feasible the large-scale multi-generation citation studies that are necessary to display the full impacts of research.

One of the highlights of the study was a time plot of the evolution of citing paper characteristics. A highly cited 1992 Science article on sandpile vibration dynamics (Jaeger and Nagel 1992)[2] was selected for the base paper, and the approximately 300 citing papers at the time of this study were evaluated for (a) level of development and (b) alignment of the main themes of the cited and citing papers. In Figure 1 (Kostoff *et al*., 2001),[3] which represents the evolution of citing paper characteristics, the abscissa represents time. The ordinate, in the second column from the left, is a two-character tensor quantity. The first number represents the level of development characterized by the citing paper (1 = basic research; 2 = applied research; 3 = advanced development/applications), and the second number represents the degree of alignment between the main themes of the citing and cited papers (1 = strong alignment; 2 = partial alignment; 3 = little alignment). Each matrix element represents the number of citing papers in each of the nine categories. The metric values were obtained through visual inspection and human judgment.

There are three interesting features on the figure. First, the tail of total annual citation counts is very long and shows little sign of abating. This is one characteristic feature of a seminal paper.

Second, the fraction of extra-discipline basic research citing papers to total citing papers ranges from about 15-25% annually, with no latency period evident. This lag-free extra-disciplinary diffusion may have been due to the combination of intrinsic broad-based applicability of the subject matter and publication of the paper in a high-circulation science journal with very broad-based readership.

Third, a 4-year latency period exists prior to the emergence of the higher development category citing papers. This correlates with the results from the bibliometrics component. From the present study, it is not possible to differentiate the reasons for this important result.

		1992	1993	1994	1995	1996	1997	1998	1999	2000
TECH DEV	33									
TECH DEV	32					1				
TECH DEV	31									
APPL RES	23							1	1	1
APPL RES	22					1			3	
APPL RES	21						1	1		
BAS RES	13	1	2	2	2	2	3		1	
BAS RES	12		2	3	6	4	10	8	10	1
BAS RES	11	3	23	28	27	43	43	30	33	4

TIME

CODE: MATRIX ELEMENT IS NUMBER OF PAPERS

Figure 1: Development category and cited paper theme alignment of citing papers

The latency could have been due to the inability of the technology community to immediately recognize the potential applications of the science. Or, it could have been due to the information remaining in the basic research journals, and not reaching the applications community. Or, the time that an application needs to be developed in this discipline is of the order of 4 years. Thus, the basic science publication feature that may have contributed heavily to extra-discipline citations may also have limited higher development category citations for the latency period.

Case study 2: Soft desorption ionization methods for mass spectrometry

The 2002 Nobel Prize in chemistry was shared by John B. Fenn, Koichi Tanaka, and Kurt Wuthrich for their work in developing methods to enable the identification and structural analysis of biological macromolecules. In particular, Fenn and Tanaka focused on soft desorption ionization methods. Fenn concentrated on electrospray ionization, and Tanaka concentrated on soft laser desorption. Following the awarding of the Nobel Prize, there was substantial controversy in the mass spectrometry community over whether Tanaka should have received a share of the prize rather than two German researchers Karas and Killenkamp. The author became aware of this controversy only after the following scientometric study was completed, and the scientometric study presents an objective perspective of why such a controversy might have occurred.

This study (Kostoff *et al*., 2004)[4] identified the literature pathways through which two highly-cited papers of Fenn and Tanaka (on which the prize was based) impacted research, technology development, and applications. Citation mining was applied to the >1600 first generation science citation index (SCI) citing papers to Fenn *et al*. 1989 science paper on electrospray ionization for mass spectrometry (EIMS) (Fenn *et al*., 1989),[5] and to the >400 first generation SCI citing papers to Tanaka's 1988 rapid communications in mass spectrometry paper on laser ionization time-of-flight mass spectrometry (Tanaka *et al*., 1988).[6] Bibliometrics was performed on the citing papers to profile the user characteristics. Text mining was performed on the citing papers to identify the technical areas impacted by the research and the relationships among these technical areas. Some very unexpected findings show the potential value of these types of analyses as precursors for major awards of this type.

The impact of these researchers on their respective disciplines can be viewed from a literature perspective. Analysis of the growth in the SCI EIMS literature (retrieved by the query electrospray and (mass or ion * or spectrometry)) and the growth in the laser desorption mass spectrometry (LDMS) literature (retrieved by the query laser and desorption and ion * or mass spectrometry) from 1988 to mid-2002 showed the following.

In the years that EIMS growth accelerated initially (1988-1990), essentially all the papers retrieved from the database cited one or more of Fenn's papers dating from 1984. From the "bottom-up" perspective, these references received a total of 151 citations between 1984 and 1990, of which 143 were from external groups. The top 20 of these 143 citing papers received over 150 citations apiece, with an aggregate second-generation citation total (for these top 20 alone) of 5400 citations.

In the years that LDMS growth accelerated initially (1990-1992), 145 papers were retrieved from the title search only. The top 50 cited papers of the 145 retrieved ranged in citations from 983 to 33. Tanaka's 1988[6] paper was referenced in 15, one or more of R. C. Beavis' papers were referenced in 37, and one or more of M. Karas' papers were referenced in 38 of these top 50 cited papers. Many of these Karas papers were published jointly with F. Hillenkamp, including one that received over 1450 citations to date (2003). From the "bottom-up" perspective, Tanaka's 1988[6] paper received a total of 69 citations between 1988 and 1992, of which all were from external groups. The top 14 of these 69 citing papers received over 100 citations apiece, with an aggregate second-generation citation total (for these top 14 alone) of 3140 citations.

RESULTS

Citation Bibliometrics

There were 1628 papers that cited Fenn's *et al*. 1989[5] paper, and 410 papers that cited Tanaka's 1988[6] paper. Because the SCI did not start to publish abstracts until 1991 and since not all citing papers have abstracts, only 1433 Fenn and 344 Tanaka citing papers containing abstracts were used. The bibliometrics analyses are performed on the total number of citing papers, whereas the text mining/computational linguistics analyses are performed on those papers with abstracts.

In the Fenn citing papers, Fenn is cited almost twice as much as the next ranked author. This is due to the citation of Fenn's other related papers between 1984 and 1989, in addition to the citation of the Science article. The

next highly cited group, RD Smith and JA Loo, worked on different mass spectrometry techniques, including electrospray ionization.

In the Tanaka citing papers, Tanaka ranks third in number of first-author citations. M. Karas of Frankfurt ranks first (along with F. Hillenkamp of Muenster, who co-authored many of these papers with Karas). This is due to three factors. First, in 1985, Karas, in conjunction with Hillenkamp, showed that a "strongly absorbing matrix at a fixed laser wavelength" could be used to vaporize small molecules without chemical degradation. Second, in 1988, Karas and Hillenkamp reported a Matrix-assisted laser desorption/ionization (MALDI) approach applied to proteins shortly after Tanaka's paper was published. Thus, the papers that cite Tanaka's paper also tend to cite the groundwork papers of Karas/Hillenkamp as well as their large molecule mass determination papers. Third, Karas and Hillenkamp were in the top tier of Tanaka citing authors, as well as prolific in their own right relative to Tanaka, and had more opportunity to cite their own foundational work in the papers in which they also cited Tanaka. Additionally, due to a series of highly-cited papers by R.C. Beavis (along with his co-author B. Chait) in the early 1990s on LDMS, many of the papers that cite Tanaka tend to multiply cite Beavis/Chait.

Of the 21 most cited authors in the Fenn citing papers, 14 are from universities, 3 are from research institutions, and 4 are from industry. Of the 21 most cited authors in the Tanaka citing papers, 16 are from universities, 1 is from a research institute, and 4 are from industry. This relatively high fraction (~20%) of cited papers from industry suggests relatively applied citing papers. The validity of this implication is confirmed in the sections on temporal citing patterns and document clustering.

Temporal Citing Patterns

In the original citation mining paper (Kostoff *et al.*, 2001),[3] two characteristics of the citing papers were evaluated as a function of time. These were: (1) the level of development of the work reported in the citing paper (basic research, applied research, technology development) and (2) the alignment between the technical thrusts of the citing paper and the cited paper (strongly aligned, partially aligned, not aligned). The Jaeger and Nagel fundamental physics paper on dynamic granular systems (Jaeger and Nagel 1992)[2] served as the research unit. It was found in Kostoff *et al.* (2001)[3] that the Jaeger/Nagel citing papers had a substantially higher basic research fraction in aggregate than the Fenn or Tanaka citing papers, there was a 4-year lag time before any applied citing papers emerged, and the Jaeger/Nagel citing papers reached a wider variety of more extreme non-aligned categories than the Fenn or Tanaka citing papers (e.g., earthquakes, avalanches, traffic congestion, war games, flow immunosensors, shock waves, nanolubrication, thin film ordering).

In aggregate, 80% of the Tanaka citing papers were concentrated in basic research, compared to 62% of the Fenn citing papers. 17% of the Tanaka citing papers were concentrated in the most non-aligned category, compared to 11% of the Fenn citing papers. 21% of the Fenn citing papers were concentrated in the applied research most-aligned category, compared to 13% of the Tanaka citing papers. These three findings emphasize the greater concentration of the Fenn citing papers in applications. The temporal evolution showed that about a decade was required before the applied technology citing papers became evident.

Fenn Citing Papers Document Clustering Taxonomy

The most cited soft laser desorption researchers in the Fenn citing papers are Karas/Hillenkamp. Tanaka does not appear in the top 20 list. To test whether this result applies beyond the Fenn citing papers, in a more recent context, a database of 300 papers was generated from the SCI. The query used was the same as previously (laser and desorption and [ion * or mass spectrometry]), and the records were the most recent prior to October 2002 (so as not to be influenced by the Nobel awards made at that time). After the elimination of (few) self-citations, the citation results were as follows: Karas-70 citations; Hillenkamp-25 citations; Tanaka-18 citations; Beavis-12 citations. Of the 70 Karas citations, 79% were pre-1989 (1985-1988). These results mirror those using MALDI as the query term. Remembering that the SCI provides the first author in citation print-outs, and most of the early soft laser desorption papers of Karas and Hillenkamp were joint, it appears that the most referenced early works on soft laser desorption/MALDI are those of Karas/Hillenkamp.

CONCLUSIONS

Citation mining produced very different patterns for Fenn and Tanaka from the bibliometrics component of the analysis. Fenn clearly stimulated the development and

growth of EIMS, as the magnitude and timing of his citations showed.

It was unclear from the bibliometrics that Tanaka stimulated the development and growth of soft laser desorption ionization mass spectrometry/MALDI more than Karas and Hillenkamp. Both the early citations (from papers published in 1990-1992) and more recent citations (from papers published immediately pre-October 2002) show a more voluminous association of Karas'/Hillenkamp's early papers with soft laser desorption ionization mass spectrometry/MALDI than Tanaka's. This issue is further exacerbated when comparing the factor matrix taxonomies of Fenn's and Tanaka's citing paper databases. There are more factors focused on applications in Fenn's citing papers, whereas there are more factors focused on mass spectrometer components in Tanaka's citing papers. A more in-depth analysis would be required to address the implications of these pattern differences, including the examination of many of the full text papers that cite Tanaka's and Karas'/Hillenkamp's works. Such an analysis was beyond the scope of the present study, but the bibliometrics has served as an agent to flag the anomaly.

After the study was completed and published, the author learned that there was controversy on the selection of Tanaka over Karas/Hillenkamp for the Nobel Prize (e.g., Anon 2002).[7] The citation metrics/results presented above flagged the basis for this controversy.

Citation Analysis – Comparison of Three Neuropsychology Journals

A scientometric analysis of articles published in the journal Cortex (a neuropsychology journal) was performed (Kostoff *et al.*, 2005).[8] One highlight was the comparison of citation performance of Cortex with two similar competitive journals: Brain and Neuropsychologia. The following experiment was run. All articles published in Cortex, Neuropsychologia and Brain in the years 1998-1999 were retrieved from the SCI. There were 110 Cortex articles, 278 Neuropsychologia articles, and 341 Brain articles. Then, the 10 most cited articles from each of the retrieval were extracted, as well as the 10 least cited articles and various characteristics compared. While the standard citation metrics were used in the analysis (e.g., numbers of authors, references, citations, etc.,) perhaps the most interesting result came from the use of a non-standard metric. All the articles were inspected visually, and a taxonomy was constructed that was inclusive of each article. The taxonomy had four categories: Clinical behavior studies; surgical interventions; non-invasive diagnostic tests; invasive diagnostic tests.

The results showed there was a distinct shift in type of study (according to the taxonomy category) in proceeding from cortex to neuropsychologia to brain. Clinical behavioral studies, many of them essentially case studies, predominated the most cited cortex papers. There were only two papers characterized as diagnostic-non-invasive (e.g., positron emission tomography, magnetic resonance imaging, etc.). Neuropsychologia had more of a balance between behavioral and diagnostic-non-invasive in its ten most cited papers. Brain showed a heavy emphasis on diagnostic-non-invasive (7/10), two papers on surgical procedures, and one on diagnostic-invasive. Based on reading abstracts from each of these journals, the types as represented in the top 10 most cited articles roughly approximate the types of papers published overall. Thus, as citations increase in absolute amounts, the study type transitions from the clinically oriented behavioral focus to the correlates with more objective measurements. Also, as the study type transitions from the clinically oriented behavioral focus ("soft" technology) to the more objective measurements ("hard" technology), the most cited papers tend to become more recent.

Citation Analysis – Journal Lancet

The purpose of this study (Kostoff 2007)[9] was to identify differences between highly cited and poorly cited medical articles, and the reasons for these differences. Characteristics of highly and poorly cited research articles (with abstracts) published in The Lancet over a 3-year period were examined. A database of Lancet papers published in a narrow time window (for time normalization) and accessed through the SCI was generated, and the detailed attributes (characteristics) of most and least cited papers were identified. Specifically, all documents classified by the SCI as articles and published in Lancet from 1997 to 1999 were examined initially.

The key component of this study was the identification of the broad range of metrics required for a comprehensive analysis. These characteristics included numerical (numbers of authors, references, citations, abstract words, journal pages), organizational (first author country, institution type, institution name), and medical (medical condition, study approach, study type, sample size, study outcome). Compared to the least cited articles, the most cited have

three to five times the median number of authors per article, 50-600% greater median number of references per article, 110-490 times the median number of citations per article, 2.5 to almost 7 times the median number of abstract words per article, and 2.5-3.5 times the median number of pages per article.

The most cited articles' medical themes emphasize breast cancer, diabetes, coronary circulation, and human immunodeficiency virus immune system problems, focusing on large-scale clinical trials of drugs. The least cited articles' themes essentially do not address the above medical issues, especially from a clinical trials perspective, cover a much broader range of topics, and have much more emphasis on social and reproductive health issues. Finally, for sample sizes of clinical trials specifically, those of the most cited articles ranged from a median of about 1500-2500, whereas those of the least cited articles ranged from 30 to 40.

Medicine has many facets, including adequacy and affordability of health care, disease and injury prevention, public health education, lab research and clinical trials, theory and experiment, individual and global health issues, and epidemiology. Out of all these possibilities that could be of substantial interest to the medical research community, the Lancet readership community has chosen to emphasize high citations to large-scale clinical drug trials on breast cancer, diabetes, coronary circulation, and immune system problems, reported by many authors in long well-referenced papers, for the time period chosen.

Citation Normalization – Study of Journal Oncogene

One method for assessing the quality of research outputs across different technical disciplines is comparing citations received by the research output documents. However, cross-discipline citation comparison studies require discipline normalization, in order to eliminate discipline differences in cultural citation practices and discipline differences in numbers of active researchers available to cite. The "definition" and number of documents used to represent a discipline becomes critical. This study (Kostoff and Martinez 2005)[10] attempted to determine whether the citation characteristics (average, median) of a discipline's domain stabilized as the domain's size was decreased. The purpose of the study was to examine citations of published papers in a given domain, allow the domain to get smaller, and ascertain whether iso-citation regions of documents become relatively size-independent (the region-average citations would remain approximately constant as the region size changes). The approach started with a collection of documents from a technical "discipline", performed document clustering that grouped the documents by similarity, allowed the groupings to get smaller, and thereby allowed the constituent documents of each group to become more similar in technical content. If the average group member citation value changed with size, this would raise questions as to whether any of the groups could be used as a denominator for clustering, and would raise more serious questions about whether credible normalization is possible.

A sample of papers (classified as research articles only, not review articles, by the Institute for Scientific Information) published in the journal Oncogene in 1999 was clustered hierarchically, and the citation averages and medians were computed for each cluster at different cluster hierarchical levels. The citation characteristics became increasingly stratified as the clusters were reduced in size, raising serious questions about the credibility of a selected denominator for normalization studies.

In summary, to compare the quality/impact of different research papers as represented by citations, the papers should be as similar thematically and typically (research article, review article, etc.) as possible. Publication dates, journals, and other factors should be normalized, where possible. For the Oncogene test case, segregation according to thematic similarity resulted in changing group citation averages. This suggests that a meaningful "discipline" citation average may not exist, and the mainstream large-scale mass production semi-automated citation analysis comparisons may provide questionable results. It further suggests that meaningful cross-discipline citation comparisons require the manually intensive approach of identifying those few research papers most closely related to the paper of interest, and normalizing on those papers (Kostoff 2002).[11] Finally, it confirms what many research evaluators recognize instinctively: There are really relatively few very thematically similar technical articles in any discipline, and any metrics used to evaluate research should be based on this reality.

SLR

Overview

The citation-assisted background (CAB) concept (Kostoff and Shlesinger 2005)[12] identifies the highly cited background documents for a research area using citation analysis. CAB rests on the assumption that a document

viewed as a significant building block for a specific research area will typically have been referenced positively by a substantial number of people who are active researchers in that specific area.

Implementation of the CAB concept then requires the following steps:

a. The research area of interest must be defined clearly
b. The documents that define the area of interest must be identified and retrieved
c. The references used most frequently in these documents must be identified and selected
d. These critical references must be analyzed, and integrated in a cohesive narrative manner to form a comprehensive background section or separate literature survey.

These required steps are achieved in the following manner:

a. The research topic of interest is defined clearly by the researchers who are documenting their study results. For example, consider the topic of severe acute respiratory syndrome (SARS-the pandemic of 2002-2003). In a 2010-2011 text mining study of SARS (Kostoff and Morse 2011),[13] the topical area was defined to include SARS research, clinical issues, and epidemiology-related issues.
b. The topical definition is sharpened further by the development of a literature retrieval query, which in the SARS case consisted of only 20 terms because of the relatively sharp focus of the SARS literature
c. The query is entered into a database search engine, and documents relevant to the topic are retrieved. In the SARS text mining study mentioned above, 2874 documents were retrieved from the Web version of the SCI/SSCI for the years 2003-early 2008. The SCI/SSCI was used because it is the only major research database to contain references in a readily extractable format
d. These documents are combined to create a separate database, and all the references contained in these documents are extracted. Identical references are combined, number of occurrences of each reference is tabulated, and a table of references and their occurrence frequencies is constructed. In the SARS text mining study, ~45,000 useful separate references were extracted and tabulated (Kostoff 2010a).[14]

Two frequencies were computed for each reference in the SARS study: The number of times each reference was cited by the 2874 records in the retrieved database only, reflecting the importance of a given reference to the specific discipline of SARS; the total number of citations the reference received from all sources, reflecting the importance of a given reference to all the fields of science that cited the reference. This latter number is obtained from the citation field or citation window in the SCI. In CAB, the first frequency is used initially since it is topic-specific. Using the first discipline-specific frequency number obviates the need to normalize citation frequencies for different disciplines (as a consequence of different levels of activity in different disciplines), as would be the case if total citation frequencies were used to determine the ordering of the references. Then, the 2874 core literature records were sorted by total citations from all sources, and any highly cited documents that were not identified using the first discipline-specific frequency number are captured in this step.

Caveats

First, listing and selection of the most highly cited references are dependent on the comprehensiveness and balance of the total records retrieved. Any imbalances (from skewed databases or incorrect queries) can influence the weightings of particular references, and result in some references exceeding the selection threshold where not warranted, and others falling below the threshold where not warranted.

Second, it is important that the query used for record retrieval be extensive as was shown for the SARS application. The query needs to be checked for precision and recall, which becomes complicated when assumptions of binary relevance and binary retrieval are relaxed. There are myriad issues to be considered when evaluating queries and their impact on precision and recall. The author's experiences with the handful of studies done so far with CAB have shown that modest query changes may substitute some papers at the citation selection threshold, but the truly important papers have citations of such magnitude that they are invulnerable to modest query changes. For this reason, the cut-off threshold for citations has been, and should be, set slightly lower, to compensate for query uncertainties.

Third, there may be situations where at least minimal citation representation is desired from each of the major technical thrust areas in the documents retrieved. In this case, the retrieved documents could be clustered into the major technical thrust areas, and the CAB process could be performed additionally on the documents for each cluster. The additional references identified with the cluster-level CAB process, albeit with lower citations than from the

aggregated non-clustered CAB process, would then be added to the list obtained with the aggregated CAB process. The author has not found this cluster-level CAB process necessary for the above purpose for any of the disciplines studied with CAB so far. However, in the SARS study, the author performed document clustering on the retrieval in order to structure the narrative, and it proved to be an invaluable aid to presenting the results. The highly cited papers were assigned to the biomedical categories generated by the clustering process, and the contribution of each paper to the theme of its respective category showed clearly its role in establishing the intellectual heritage of the SARS literature.

Fourth, there may be errors in citation counts because of references errors, and the subsequent fragmenting of a reference's occurrence frequency metric into smaller metric values. Care needs to be taken in insuring that a given reference is not divided into multiple large fragments, which are not subsequently combined. In all the SLR studies performed to date, considerable effort has been devoted to insure that the different representations of the same reference were aggregated into one, with the frequencies adjusted accordingly.

Fifth, the CAB approach is most accurate for recent references, and its accuracy drops as the references recede into the distant past. This derives from the tendency of authors to reference more recent documents and, given the restricted real estate in journals, not reference the original documents. To get better representation, and more accurate citation numbers, for early historical documents, the more recent references need to be retrieved, collected into a database, and have their references analyzed in a similar manner (essentially examining generations of citations).

Sixth, high citation frequencies are not unique to important documents only; different types of references can have high citation frequencies. Documents that contain critical research advances, and were readily accessible in the open literature tend to be cited highly, and represent the foundation of the CAB approach.

Application of CAB so far shows that this type of document is predominant in the highly cited references list. Books or review articles also appear on the highly cited references list. These documents do not usually represent new advances, but rather are summaries of the state of the art (and its background) at the time the document was written. These types of documents are still quite useful as background material. Finally, documents that receive large numbers of citations highly critical of the document could be included in the list of highly cited documents. In the studies performed so far, the author has not identified such papers in the detailed development of the background.

Additionally, one of the application studies being completed concerns high speed compressible flow, a discipline in which the author worked decades ago. Using the CAB approach, the author found that all the key historical documents with which he was familiar were identified, and all the historical documents identified appeared to be important. Thus, for that data point at least, the weaknesses identified above (imbalances, undervaluing early historical references, unwanted highly cited documents) did not materialize. To ensure that any critical documents were not missed because of imbalance problems, the threshold was set a little bit lower to be more inclusive.

The converse problem to multiple types of highly cited references, some of which may not be the important documents desired, is influential references that do not have substantial citation frequencies. If the authors of these references did not publish them in widely and readily accessible forums, or if they do not contain appropriate verbiage for optimal query accessibility, then they might not have received large numbers of citations. Additionally, journal or book space tends to be limited, with limited space for references. In this zero-sum game for space, research authors tend to cite relatively recent records at the expense of the earlier historical records. Inclusion of the references that were not widely available when published is more problematical and tends to rely on the background developers' personal knowledge of these documents, and their influence.

Identification of very old or very new seminal references

Extremely recent but influential references have not had the time to accumulate sufficient citations to be listed above the selection threshold on the citation frequency table. Methods of including these influential records located at the wings of the temporal distribution will now be described in the following implementation section.

To identify the total candidate references for the background section, a table containing all the references from the retrieved records, is constructed. In the SARS case, this table contained approximately 45,000 references. A threshold frequency for selection can be determined by arbitrary inspection (e.g., a background section consisting

of 150 key references is arbitrarily selected). The author has found a dynamic selection process more useful. In this dynamic process, references are selected, analyzed, and grouped based on their order in the citation frequency table until the resulting background is judged sufficiently complete by the background developers.

To ensure that the influential documents at the wings of the temporal distribution (very old and very new) are included, the following total process is used. The reference frequency table is ordered by inverse frequency initially, as above, and a high value of the selection frequency threshold is selected initially. Then, the table is re-ordered chronologically. The early historical documents with citation frequencies substantially larger than those of their contemporaries are selected, as are the extremely recent documents with citation frequencies substantially larger than those of their contemporaries. By contemporaries, it is meant documents published in the same time frame, not limited to the same year. Then, the dynamic selection process defined above is applied to the early historical references, the intermediate time references (those falling under the high frequency threshold), and the extremely recent references.

Topical CAB studies performed

CAB topical areas studied include non-linear dynamics (Kostoff and Shlesinger 2005),[12] nanotechnology (Kostoff *et al*., 2006; Kostoff *et al*., 2009; Kostoff *et al*., 2011);[15-17] anthrax (Kostoff *et al*., 2007a);[18] SARS (Kostoff 2010a);[14] high speed compressible flow (unpublished).

LRDI

Journal special issue on LRDI technique and medical/ technical studies

LRDI (formerly LRD-literature-related discovery) integrates (a) discovery generation from disparate literatures with (b) the wealth of knowledge contained in the prior art to (c) potentially reverse chronic and infectious diseases and/or (d) potentially solve technical problems that appear intractable. A detailed review of the LRDI literature has been published (Kostoff *et al*., 2008a),[19] and an updated review of the LRDI technique and findings has also been published (Kostoff 2012b).[20] First generation LRDI efforts culminated in a special issue of technological forecasting and social change in 2008, which included eight papers from the author's research group (Kostoff 2008b-c; Kostoff *et al*., 2008d-i).[21-24] Four of the papers were on medical topics (Raynaud's Phenomenon (Kostoff *et al*., 2008e),[25] Cataracts (Kostoff 2008c),[22] Parkinson's disease (PD) (Kostoff *et al*., 2008f),[26] multiple sclerosis (MS) (Kostoff *et al*., 2008g),[27] and the fifth was on a technical non-medical topic (Alternative Water Desalinization Approaches) (Kostoff *et al*., 2008h).[28] All of the studies were of the open discovery system type, where one starts with a problem (e.g., the disease) and identifies solutions (e.g., preventatives, treatments).

The approach used to identify discovery in the medical LRDI papers was through a query that contained terms of disease characteristics to be eliminated. This query was intersected with classes of potential discovery (these classes were limited to non-drugs non-advanced technology; they were mainly foods and food extracts), and those papers in the retrieval that included the name of the disease under consideration were eliminated from further analysis for discovery (they were prior art). Typically, there were hundreds of papers in the prior art category. There were also hundreds of papers in the discovery candidate category. A sub-set of these papers was selected for validation of no prior art, and those that passed this validation process were published as potential discovery. Typically, half the papers that underwent validation were prior art. Thus, the total results included many papers of prior art and many papers of potential discovery. A comprehensive report was generated (Kostoff *et al*., 2007b)[29] that included much of this data on prior art, as well as some of the potential discovery data.

In the published papers, sample results from the prior art, directly-related potential discovery and indirectly-related potential discovery were presented. Treatment protocols were not presented. Thus, for each disease examined, the research product was identifying hundreds of papers describing potential preventatives/treatments, probably comprising over a hundred different "treatment" concepts. In order for these results to be implemented, they had to be culled down to perhaps the 10 or 20 most important. This is a difficult procedure because of unknown synergies, positive or negative that could arise from the astronomical number of different combinations of these potential "treatments". The only feasible way for the culling to occur would be a panel of experts using their best judgments.

Next generation of LRDI - SARS study

In 2011, an LRDI study on SARS was published (Kostoff 2011).[30] The main advance over the previous LRDI technique was the form of the query. A functional form was developed that basically expressed what outcome was

desired (e.g., enhance humoral immunity, restrict viral entry, etc.). Combined with proximity search capability this query proved to be a more effective filter for potential discovery than previous filters. It could also be applied to the full text, rather than limited to abstracts as before. Full text experiments showed one order of magnitude or more increase in retrievals compared to using abstracts.

One important finding of the SARS study came from the background literature review. Approximately 8000 people worldwide presented with SARS symptoms, of whom about 10% succumbed. This was not a random 10%. The people who succumbed had significant co-morbidities and weak immune system parameters. None of the drugs worked; the effective treatments were good hygiene, isolation, and quarantine. What kept the 90% alive was a strong immune system; therefore, strengthening the immune system became the target for the discovery study.

Bibliographic coupling to enhance potential discovery

In 2010, an LRDI study on the relationship between PD and Crohn's disease (CD) was published (Kostoff 2010b).[31] PD is a neurodegenerative disease while CD is an autoimmune disease; the question arose whether there could be any common features. This was the first of the author's LRDI studies that was the closed discovery system type, where one starts with two problems, or a problem and a solution, and searches for mechanisms/features that link them. The study combined two approaches for identifying common features in two literatures: Text-based and citation-based. The text-based approach was to identify records in each literature that contained common phrases. Now, this could be done at the full-text level, at the abstract level, at the title level, or at combinations of all three levels. For demonstration purposes, common phrases in titles were used. The citation-based approach was to identify records that had common references. For this component, bibliographic coupling was used.

What was the value of combining a text-based approach with a citation-based approach? Methods that use a text-based approach only, such as the excellent Arrowsmith software (Swanson and Smalheiser 1999),[32] tend to produce thousands or tens of thousands of these intermediate common phrases, depending on the size of the literatures linked. Since the evaluation of each phrase for potential discovery requires reading the records associated with the phrase, the problem quickly becomes infeasible without provision of additional filtering criteria. Arrowsmith has a number of built-in filtering options (Smalheiser 2012);[33] the author previously developed a filtering approach based on phrases identified through document clustering and factor analysis. There was the possibility that bibliographic coupling superimposed on text phrase matching could provide an even more effective filter.

The study details are contained in a comprehensive report (Kostoff 2010b).[31] In terms of the findings, there were three major themes that unified the PD and CD literatures: Genetics; Neuroimmunology; Cell Death. Some new concepts at the sub-set level of the main themes were identified. The synergy of matching phrases and shared references provided a strong prioritization to the selection of promising matching phrases as discovery mechanisms.

Second-generation improvement of LRDI technique - Vitreous restoration study

The author and a collaborator are presently completing a study on vitreous restoration (vitreous is the gel between the lens and the retina in the eye; its degradation can enhance the development of cataracts in the lens and more serious retinal diseases). This team is using an improved version of the functional query first shown in the SARS study, including proximity searching capability. The team is using a text-based query approach in concert with a citation-based query, which exploits the strengths of each approach and eliminates the weaknesses. Previous LRDI study restrictions limiting discovery to non-drug non-advanced technology concepts only have been removed, and all potential forms of treatment are being considered. Initial results show that general systemic and local problem-focused treatments are both required for optimal healing, but treatment effectiveness will be strongly related to the ability to identify and remove causes of disease. More effort is being placed on identifying the widest spectrum of potential causes for vitreous degradation, in order to insure that the potential treatments identified cover the widest spectrum of causes possible. Initial results also show that a number of potential causes have not been researched in the literature, and these un-researched potential causes have been identified as research gaps.

DISCUSSION AND CONCLUSIONS

This review has examined three superficially different techniques for knowledge discovery: Citation scientometrics; SLR; LRDI. However, all three use text-based queries and citation-based queries to extract information from databases over space and time. Unlike most standard practice, the variants of the techniques described in this review focuses on the retrieval aspect of citations rather

than the counting aspect. The analysis of these articles for knowledge discovery is key, not the numerics. Additionally, when constructed properly, the queries allow the retrieval of information from many disparate disciplines when necessary to help solve the problems of interest. Thus, cross-disciplinary retrieval is not treated as a target to be achieved, but rather as an unrestricted capability that is available when necessary.

Two examples of citation mining were presented: Sandpile vibration dynamics, and soft ionization mass spectrometry. One highlight of the sandpile literature was a temporal analysis of the citing papers of a heavily cited basic research sandpile dynamics paper. The fraction of extra-discipline basic research citing papers to total citing papers ranged from about 15-25% annually, with no latency period evident. However, a 4-year latency period existed prior to the emergence of the higher development category citing papers.

For the soft ionization mass spectrometry example, the most highly cited papers of each of the two co-recipients of the Nobel Prize in Chemistry in 2002 (Fenn and Tanaka) were analyzed scientometrically. Fenn clearly stimulated the development and growth of EIMS, as the magnitude and timing of his citations showed. It was unclear from the bibliometrics that Tanaka stimulated the development and growth of soft laser desorption ionization mass spectrometry more than Karas and Hillenkamp (two German researchers). In this example, both the numerics and the content of the citing papers proved to be important.

A comparison of three neuropsychology journals showed that as citations increased in absolute amounts, the study type transitioned from the clinically oriented behavioral focus to the correlates with more objective measurements. Also, as the study type transitioned from the clinically oriented behavioral focus ("soft" technology) to the more objective measurements ("hard" technology), the most cited papers tended to become more recent. In plain English, more research funds are available for the non-invasive high-tech approaches in this field, more researchers work on the high-tech areas and more journals publish the high-tech papers, and consequently the high-tech papers and the journals that publish more of these high-tech papers are the most highly cited.

A citation analysis of most cited Lancet articles shows that the Lancet readership community has chosen to emphasize high citations to large-scale clinical drug trials on breast cancer, diabetes, coronary circulation, and immune system problems, reported by many authors in long well-referenced papers, for the time period chosen. As in the neuropsychology example shown above, whether these well-funded high-tech highly-cited approaches are in the best long-term interest of patients remains to be demonstrated. A recent publication by the author questions the heavy emphasis by the medical community on high-tech treatments for chronic diseases without equal emphasis on cause determination and elimination (Kostoff 2012b).[20]

In a citation normalization study of the journal Oncogene, segregation according to thematic similarity resulted in changing group citation averages. This suggested that a meaningful "discipline" citation average may not exist, and the mainstream large-scale mass production semi-automated citation analysis comparisons may provide questionable results. More studies of this type on different disciplines are needed to determine the universality of this conclusion.

In a study combining bibliographic coupling with text matching to identify similarities between PD and CD, the synergy of matching phrases and shared references provided a strong prioritization to the selection of promising matching phrases as discovery mechanisms.

The CAB approach has been applied to a handful of disciplines. It was most effective when subject matter experts were involved in the selection of seminal documents and the subsequent analysis, and when clustering was applied to the retrieved seminal documents for the purpose of structuring the subsequent narrative.

LRDI has been applied to a handful of diseases (and upgraded in parallel). In its latest incarnation, all restrictions on types of potential treatments have been removed. In a recent summary of the evolution of LRDI (Kostoff 2012b),[20] one example was provided of the relationship between the findings of the 2008 LRDI study on preventatives and treatments for MS (Kostoff *et al.*, 2008g)[27] and a recent demonstration of the reversal of an advanced case of MS. There is nothing in the LRDI technique that precludes these types of positive impacts from applying to any chronic or infectious disease.

REFERENCES

1. Kostoff RN. Text mining for science and technology - A review: Part I - characterization/scientometrics. J Scientometr Res 2012;1:11-21..
2. Jaeger HM, Nagel SR. Physics of the granular state. Science 1992;256:1523-31.

3. Kostoff RN, Del Rio JA, García EO, Ramírez AM, Humenik JA. Citation mining: Integrating text mining and bibliometrics for research user profiling. J Am Soc Inf Sci Technol 2001;52:1148-56.
4. Kostoff RN, Bedford CD, del Río JA, Cortes HD, Karypis G. Macromolecule mass spectrometry: Citation mining of user documents. J Am Soc Mass Spectrom 2004;15:281-7.
5. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989;246:64-71.
6. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y. Protein and polymer analysis up to M/Zx 100,000 by laser ionisation time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 1988;2:151-3.
7. Anon. Nobel Prize Controversy. The Scientist. December 11, 2002.
8. Kostoff RN, Buchtel HA, Andrews J, Pfeil KM. The hidden structure of neuropsychology: Text mining of the journal Cortex: 1991-2001. Cortex 2005;41:103-15.
9. Kostoff RN. The difference between highly and poorly cited medical articles in the journal Lancet. Scientometrics 2007;72:513-20.
10. Kostoff RN, Martinez WL. Is citation normalization realistic? J Inf Sci 2005;31:57-61.
11. Kostoff RN. Citation analysis for research performer quality. Scientometrics 2002;53:49-71.
12. Kostoff RN, Shlesinger MF. CAB-citation-assisted background. Scientometrics 2005;62:199-212.
13. Kostoff RN, Morse SA. Structure and infrastructure of infectious agent research literature: SARS. Scientometrics 2011;86:195-209.
14. Kostoff RN. The highly cited SARS research literature. Crit Rev Microbiol 2010a;36:299-317.
15. Kostoff RN, Murday J, Lau C, Tolles W. The seminal literature of global nanotechnology research. J Nanopart Res 2006;8:193-213.
16. Kostoff RN, Koytcheff RG, Lau CG. Seminal nanotechnology literature: A review. J Nanosci Nanotechnol 2009;9:6239-70.
17. Kostoff RN, Koytcheff RG, Lau CG. Characteristics of the seminal nanotechnology literature. Encyclopedia of Nanoscience and Nanotechnology. Vol. 12: American Scientific Publishers; 2011. p. 271-300.
18. Kostoff RN, Morse SA, Oncu S. The seminal literature of anthrax research. Crit Rev Microbiol 2007a;33:171-81.
19. Kostoff RN, Block JA, Solka JA, Briggs MB, Rushenberg RL, Stump JA, *et al*. Literature-related discovery. Annu Rev Inf Sci Technol 2008a;43:243-85.
20. Kostoff RN. Literature-related discovery and innovation - update. Technol Forecast Soc Change 2012;79:789-800.
21. Kostoff RN. Literature-related discovery: Introduction and background. Technol Forecast Soc Change 2008b;75:165-85.
22. Kostoff RN. Literature-Related Discovery: Potential treatments for cataracts. Technol Forecast Soc Change 2008c;75:215-25.
23. Kostoff RN, Briggs MB, Solka JA, Rushenberg RL. Literature-related discovery: Methodology. Technol Forecast Soc Change 2008d;75:186-202.
24. Kostoff RN, Block JA, Solka JA, Briggs MB, Rushenberg RL, Stump JA, *et al*. Literature-related discovery: Lessons learned, and future research directions. Technol Forecast Soc Change 2008i;75:276-99.
25. Kostoff RN, Block JA, Stump JA, Johnson D. Literature-related discovery: Potential treatments for Raynaud's phenomenon. Technol Forecast Soc Change 2008e;75:203-14.
26. Kostoff RN, Briggs MB. Literature-related discovery: Potential treatments for parkinson's disease. Technol Forecast Soc Change 2008f;75:226-38.
27. Kostoff RN, Briggs MB, Lyons T. Literature-related discovery: Potential treatments for Multiple Sclerosis. Technol Forecast Soc Change 2008g;75:239-55.
28. Kostoff RN, Solka JA, Rushenberg RL, Wyatt JR. Literature-related discovery: Potential improvements in water purification. Technol Forecast Soc Change 2008h;75:256-75.
29. Kostoff RN, Block JA, Solka JA, Briggs MB, Rushenberg RL, Stump JA, *et al*. Literature-related discovery: A review. DTIC Technical Report Number ADA473643 Fort Belvoir, VA: Defense Technical Information Center; 2007b. Available from: http://www.dtic.mil/.
30. Kostoff RN. Literature-related discovery: Potential treatments and preventatives for SARS. Technol Forecast Soc Change 2011;78:1164-73.
31. Kostoff RN. Literature-related discovery: Common factors for Parkinson's Disease and Crohn's Disease. DTIC Technical Report Number ADA525269. 2010b. Defense Technical Information Center. Fort Belvoir, VA. Available from: http://www.dtic.mil/.
32. Swanson DR, Smalheiser NR. Implicit text linkages between Medline records: Using arrowsmith as an aid to scientific discovery. Libr Trends 1999;48:48-59.
33. Smalheiser NR. 2012. Available from: http://www.arrowsmith.psych.uic.edu/arrowsmith_uic/index.html.

How to cite this article: Kostoff RN. Text mining for science and technology: A review - Part II-citation and discovery. J Sci Res 2013;2:3-14.

Source of Support: Nil, **Conflict of Interest:** None declared