Macromolecule Mass Spectrometry: Citation Mining of User Documents

Ronald N. Kostoff and Clifford D. Bedford
Office of Naval Research, Arlington, Virginia, USA

J. Antonio del Río and Héctor D. Cortes
Centro de Investigación en Energía, Universidad Nacional de Mexico, Temixco, México

George Karypis
University of Minnesota, Minneapolis, Minnesota, USA

Identifying research users, applications, and impact is important for research performers, managers, evaluators, and sponsors. Identification of the user audience and the research impact is complex and time consuming due to the many indirect pathways through which fundamental research can impact applications. This paper identified the literature pathways through which two highly-cited papers of 2002 Chemistry Nobel Laureates Fenn and Tanaka impacted research, technology development, and applications. Citation Mining, an integration of citation bibliometrics and text mining, was applied to the >1600 first generation Science Citation Index (SCI) citing papers to Fenn's 1989 Science paper on Electrospray Ionization for Mass Spectrometry, 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. 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. (J Am Soc Mass Spectrom 2004, 15, 281–287) © 2004 American Society for Mass Spectrometry

Over the past decade, electrospray ionization and laser desorption mass spectrometry have become the preferred methods for large molecule (especially biological) mass measurements. The present Background section describes the growth of the electrospray ionization and laser desorption mass spectrometry literatures, and relates the growth of these literatures to the original papers by Nobel co-recipients John B. Fenn and Koichi Tanaka, and to the papers of other principal contributors as well. The Background section then proceeds to describe the information technology approaches used in this analysis (text mining, bibliometrics, citation mining).

Growth of the Macromolecular Mass Spectrometry Literature

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 [1–7], and Tanaka concentrated on soft laser desorption [8–10].

The impact of these researchers on their respective disciplines can be viewed from a literature perspective. Figure 1 shows the growth in the SCI electrospray ionization mass spectrometry (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. The dashed curves are based on papers retrieved by a query applied to all text fields (Title, Abstract, Keywords), while the solid curves are based on a query applied to the Title field only. Before 1991, Abstracts were not available for SCI papers.

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 [1–7]. From the "bottom-up" perspective, references [1–7] received a total of 151 citations between 1984 and 1990, of which 143 were from external groups. The top twenty of these 143 citing papers received over

Published online January 15, 2004

Address reprint requests to Dr. R. N. Kostoff, Office of Naval Research, 800 North Quincy St., Arlington, VA, 22217, USA. E-mail: kostofr@onr.navy.mil

1044-0305/04/$30.00
doi:10.1016/j.jasms.2003.11.010

Received June 16, 2003
Revised November 8, 2003
Accepted November 16, 2003

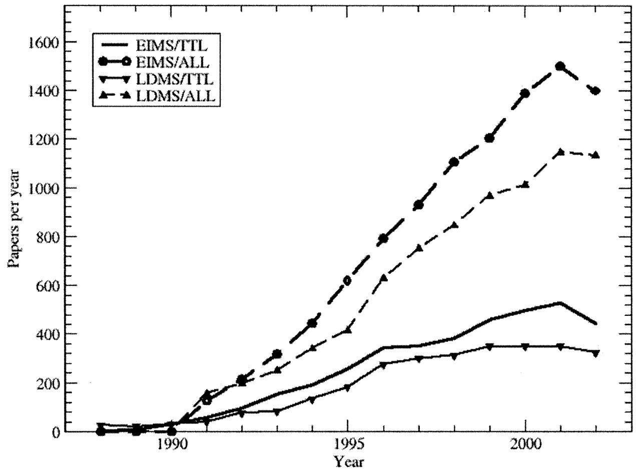

Figure 1. Growth in electrospray and laser desorption literatures (papers per year versus time).

150 citations apiece, with an aggregate second-generation citation total (for these top twenty 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 fifty cited papers of the 145 retrieved ranged in citations from 983 to 33. Tanaka's 1988 paper [8] was referenced in fifteen, one or more of R. C. Beavis' papers (e.g., [11–13]) were referenced in 37, and one or more of M. Karas' papers (e.g., [14, 15]) were referenced in 38 of these top fifty cited papers. Many of these Karas papers were published jointly with F. Hillenkamp. Reference [14] in particular has received over 1450 citations to date. From the "bottom-up" perspective, reference [8] received a total of 69 citations between 1988 and 1992, of which all were from external groups. The top fourteen of these 69 citing papers received over 100 citations apiece, with an aggregate second-generation citation total (for these top fourteen alone) of 3140 citations.

References [1–8] have been cited highly. In particular, references [1–7] have received ~590, 210, 670, 210, 370, 1630, and 890 citations, respectively, by November 2002, and reference [8] has received 410 citations. The citing community can be viewed as a sub-set of the total user community. Identifying the characteristics of the citing community would provide one perspective on the diversity of *impact* that these papers have had or, more accurately, on the diversity of *citings* that these papers have had.

Text Mining

Science and technology (S&T) text mining [16–19] is a computational linguistics-based process for extracting useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from frequently occurring technical phrases. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. Text mining can be used for:

- Enhancing information retrieval and increasing awareness of the global technical literature [20–22].
- Potential discovery and innovation based on merging common linkages between very disparate literatures [23–26].
- Uncovering unexpected asymmetries from the technical literature [27, 28].
- Estimating global levels of effort in S&T sub-disciplines [29–31].
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their Impact Factors [32].
- Tracking myriad research impacts across time and applications areas [33, 34].

A typical text mining study of the published literature develops a query for comprehensive information retrieval, processes the database using computational linguistics and bibliometrics, and integrates the processed information.

Bibliometrics

Evaluative bibliometrics [35–37] uses counts of publications, patents, citations, and other potentially informative items to develop science and technology performance indicators. Its validity is based on the premises that, (1) counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents or papers, (2) the number of times those patents or papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers, and (3) the citations from papers to papers, from patents to patents, and from patents to papers provide indicators of intellectual linkages between the organizations which are producing the patents and papers, and knowledge linkage between their subject areas [38]. Evaluative bibliometrics can be used to:

- Identify the infrastructure (authors, journals, institutions) of a technical domain.
- Identify experts for innovation-enhancing technical workshops and review panels.
- Develop site visitation strategies for assessment of prolific organizations globally.
- Identify impacts (literature citations) of individuals, research units, organizations, and countries.

Citation Mining

Citation Mining [34, 39] is a technique developed for the purpose of characterizing the aggregate citing papers of a research unit. A research unit can consist of one paper, selected papers from an author, or selected papers from a group or technical discipline. In Citation Mining, text mining and bibliometrics analyses are performed on the aggregate citing papers. The bibliometrics component yields the infrastructure information (e.g., prolific authors, journals, institutions, countries, most cited authors, papers, journals, etc.), and the computational linguistics component yields the pervasive technical thrusts and the relationships among the thrusts. A temporal component documents the dissemination of information to the research and user community as a function of time.

The Science Citation Index (SCI) is a database that links papers (P1) in journals indexed by the SCI to other SCI papers (P2) that cite the original papers P1, and contains references (P3) in the original papers P1 as well. While the SCI accesses many of the premier research journals, it does not access all technical journals published. In the present study, the SCI is used to identify the citing papers to Fenn's and Tanaka's original papers. Thus, only those in journals accessed by the SCI will be identified.

This paper describes the application of Citation Mining to the subset of the most highly cited papers of Fenn [6] and Tanaka [8] referenced above, using the SCI as the source for citing papers. Because temporal dissemination and impacts of the initial cited papers is also a key feature of Citation Mining, it was desired to limit the analysis to one paper from each researcher, in order to have a sharp starting point in time.

Results

The results from the publications bibliometric analyses are presented first, followed by the citations bibliometrics analysis. Results from the computational linguistics analyses are shown last. The SCI bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, Keywords, and references. Due to space limitations, only journal bibliometrics are presented here. Reference [40] contains the details of the complete study results.

Publication Bibliometrics

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred, since these papers are published in the (typically) high caliber journals accessed by the SCI. There were 1628 papers that cited Fenn's 1989 paper, and 410 papers that cited Tanaka's 1988 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 computational linguistics are performed on those papers with Abstracts.

Journal Frequency Results

For both the Fenn and Tanaka citing papers, the most prolific journals focus on mass spectrometry, chemistry, and biology. Three journals stand out as the first tier for containing the most citing papers: Analytical Chemistry, Journal of the American Society for Mass Spectrometry, Rapid Communications in Mass Spectrometry. Twelve journals are in common between the two authors. The non-common Fenn citing journals tend to focus on biology and biochemistry (Analytical Biochemistry, Biochemistry, Protein Science, European Journal of Biochemistry), while those of Tanaka focus on the technique and instrumentation (Review of Scientific Instruments, Organic Mass Spectrometry, European Mass Spectrometry). This observation supports the later document clustering finding of the greater emphasis on bio-molecules in the Fenn citing papers relative to the Tanaka citing papers.

Citation Statistics

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics [36], *much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers* [41, 42]. Only author citation frequency results are presented here.

Author Citation Frequency Results

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 [1–5, 7], in addition to the citation of the Science article [6]. The next highly cited group, R. D. Smith and J. A. Loo, worked on different mass spectrometry techniques, including electrospray ionization (e.g., [43–45]).

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 [46]. Second, in 1988, Karas and Hillenkamp reported a MALDI approach applied to proteins [47] 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 (e.g., [48]). 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 laser desorption mass spectrometry (e.g., [11–13]), many of the papers that cite Tanaka tend to multiply cite Beavis/Chait.

There are five names in common between the two lists of most highly cited authors in the Fenn and Tanaka citing papers (Fenn, Smith, Karas, Beavis, Hillenkamp). All five have made broad contributions to mass spectrometry.

Of the 21 most cited authors in the Fenn citing papers, fourteen are from universities, three are from research institutions, and four are from industry. Of the 21 most cited authors in the Tanaka citing papers, sixteen are from universities, one is from a research institute, and four 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 papers [34, 39], 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 [49] served as the research unit. It was found that the citing papers had a substantially higher basic research fraction in aggregate than the Fenn or Tanaka citing papers, there was a four-year lag time before any applied citing papers emerged, and the Jaeger and 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).

These two characteristics were evaluated in the present paper. The detailed approach and results are presented in reference [40].

In aggregate, 80% of the Tanaka citing papers were concentrated in basic research, compared to 62% of the Fenn citing papers. Seventeen percent of the Tanaka citing papers were concentrated in the most non-aligned category, compared to 11% of the Fenn citing papers. Twenty-one percent 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.

Computational Linguistics (Taxonomy Generation)

Three statistically-based clustering methods, factor matrix, multi-link aggregation, and partitional document clustering, were used to develop taxonomies. They each offered a modestly different perspective on taxonomy category structure. Only partitional document clustering is summarized here. The detailed results of all three methods are contained in reference [40].

Partitional Document Clustering

Document clustering is the grouping of similar documents into thematic categories. Different approaches exist [50–59]. The approach presented here is based on a partitional clustering algorithm [60, 61] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements. The CLUTO algorithm then aggregates the clusters in a hierarchical taxonomy.

Fenn Citing Papers Document Clustering Taxonomy

Overall, the main category, Level 1, contains 1431 records, with a broad focus of bio-molecular applications and the ionization-charge components of the mass detection and analysis process. Level 2 contains the first major categorical split of two categories: Applications and Ionization Process. There are 532 records in Applications, focused on large bio-molecules. Additionally, there are 899 records in Ionization Process, focusing on the charging process and charge state, as well as the sample solution prior to ionization. Level 3 contains the next categorical split of four categories: Bio-molecule Structure, MALDI Protein Mapping, Ionization, and Sample Preparation.

The Applications category of Level 2 subdivides into Bio-molecule Structure and MALDI Protein Mapping. There are 349 records in Bio-molecule Structure, focused on proteins, peptides, binding states, and amino acid sequencing. There are 183 records in MALDI Protein Mapping, focused on the use of MALDI for protein mapping. Sampling of these records shows the main focus to be MALDI, with Fenn/ESI appearing mainly as a reference. Appearance of MALDI papers in the Fenn citing papers implies that either ESI is being cited as a MALDI alternative for Protein Mapping or that ESI is being cited historically as a demonstration that large bio-molecule mass measurements were possible.

The most cited soft laser desorption researchers in the Fenn citing papers are Karas/Hillenkamp. Tanaka does not appear in the top twenty 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 in the Background section (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). 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. As shown in the Background section, this was true over a decade ago, and as shown in this paragraph, it remains true today.

The Ionization Process category of Level 2 subdivides into Ionization and Sample Preparation. There are 398 records in Ionization, focused on characteristics of the charged state. There are 501 records in Sample Preparation, focused on the process and components preparatory to ionization.

Tanaka Citing Papers Document Clustering Taxonomy

Overall in Level 1, the total database contains 344 records, with a broad focus of MALDI, bio-molecular, and non-biomolecular applications. Level 2 contains the first major categorical split: Applications and Analytical Process. There are 131 records in Applications, focused on large bio-molecules, oligomers, and polymers. Additionally, there are 213 records in Analytical Process, focusing on charging process and sample preparation.

Level 3 contains the next categorical split of 4 categories: Bio-molecules, Non-bio-molecules, Sample Preparation, and Mass Resolution. The Applications category of Level 2 subdivides evenly into Bio-molecules and Non-bio-molecules. There are 66 records in Bio-molecules, focused on proteins, peptides, and amino acid sequencing. There are 65 records in Non-bio-molecules, focused on oligomers and polymers. *This Non-bio-molecules category does not appear in the Fenn citing papers, at least as a dominant theme.*

The Analytical Process category of Level 2 subdivides into Sample Preparation and Mass Resolution. There are 95 records in Sample Preparation, focused on the steps leading to ionization, especially on preparation of the matrix. There are 118 records in Mass Resolution, focused on the control of mass spectrometer fields and energies necessary to increase the precision of mass determination.

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 electrospray ionization mass spectrometry, 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 exascerbated 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.

The text mining identified the major technical thrusts of both the Fenn and Tanaka citing databases. The document clustering identified both the main technical thrusts and the number of papers devoted to each thrust. If an abbreviated text mining methodology is desired to identify major technical thrusts and approximate levels of effort devoted to each thrust, the document clustering methodology could provide a reasonable first approximation.

The main differences in the higher taxonomy levels appeared to be twofold. First, the Tanaka citing paper applications are evenly split between bio-molecules and oligomers/polymers, whereas the Fenn citing papers appear to focus predominately on bio-molecules. This reflects the ability of the MALDI approach to address both bio-molecules and a wide range of polymers, whereas electrospray requires soluble analytes that are readily ionizable. This restricts the classes of polymers that can be analyzed by ESI. Second, there is a MALDI component in the Fenn citing papers, but not an ESI component in the Tanaka citing papers. This reflects the practical situation that MALDI can be viewed as an alternative to ESI for bio-molecules, but ESI is much less an alternative to MALDI for polymers, for the analyte solubility reason shown above.

Disclaimer

The views in this paper are solely those of the authors, and do not necessarily represent the views of the United States Department of the Navy or any of its components, the Universidad Nacional Autonoma de Mexico, or the University of Minnesota.

References

1. Yamashita, M.; Fenn, J. B. Electrospray Ion-Source—Another Variation on the Free-Jet Theme. *J. Phys. Chem.* **1984,** *88(20),* 4451–4459.
2. Yamashita, M.; Fenn, J. B. Negative-Ion Production with the Electrospray Ion-Source. *J. Phys. Chem.* **1984,** *88(20),* 4671–4675.
3. Whitehouse, C. M.; Dreyer, R. N.; Yamashita, M.; Fenn, J. B. Electrospray Interface for Liquid Chromatographs and Mass Spectrometers. *Anal. Chem.* **1985,** *57(3),* 675–679.
4. Wong, S. F.; Meng, C. K.; Fenn, J. B. Multiple Charging in Electrospray Ionization of Poly(Ethylene Glycols). *J. Phys. Chem.* **1988,** *92(2),* 546–550.
5. Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting Mass-Spectra of Multiply Charged Ions. *Anal. Chem.* **1989,** *61(15),* 1702–1708.
6. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. *Science* **1989,** *246(4926),* 64–71.
7. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization—Principles and Practice. *Mass Spectrom. Rev.* **1990,** *9(1),* 37–70.
8. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y. Protein and Polymer Analysis up to M/Zx 100,000 by Laser Ionization Time-of-Flight Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **1988,** *2,* 151–153.
9. Tanaka, K.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. *Proceedings of the Second Japan-China Joint Symposium on Mass Spectrometry*. Editors Matsuda, H. and Xiao-tian L. Osaka, Japan, September 15-18, 1987, 185-188.
10. Yoshida, T.; Tanaka, K.; Ido, Y.; Akita, S.; Yoshida, Y. *Mass Spectrosc*. (Japan). 1988, 36, 59.
11. Beavis, R. C.; Chait, B. T. High-Accuracy Molecular Mass Determination of Proteins Using Matrix-Assisted Laser Desorption Mass-Spectrometry. *Anal. Chem.* **1990,** *62(17),* 1836–1840.
12. Beavis, R. C.; Chait, B. T. Cinnamic Acid Derivatives as Matrices for Ultraviolet Laser Desorption Mass Spectrometry of Proteins. *Rapid Commun. Mass Spectrom.* **1989,** *3(12),* 432–435.
13. Beavis, R. C.; Chait, B. T. Factors Affecting the Ultraviolet Laser Desorption of Proteins. *Rapid Commun. Mass Spectrom.* **1989,** *3(7),* 233–237.
14. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60(20),* 2299–2301.
15. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. Matrix-Assisted Ultraviolet-Laser Desorption of Nonvolatile Compounds. *Int. J. Mass Spectrom. Ion Processes* **1987,** *78,* 53–68.
16. Kostoff, R. N. Text Mining for Global Technology Watch. In *Encyclopedia of Library and Information Science, Vol. 4,* Second Edition Drake, M., Ed.; Marcel Dekker, Inc: New York, NY, 2003; pp 2789–2799.
17. Hearst, M. A. Untangling Text Data Mining. Proceedings: ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. 1999. University of Maryland, June 20-26.
18. Zhu, D. H.; Porter, A. L. Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting. *Technological Forecasting and Social Change* **2002,** *69(5),* 495–506.
19. Losiewicz, P.; Oard, D.; Kostoff, R. N. Textual Data Mining to Support Science and Technology Management. *J. Int. Info. Syst.* **2000,** *15,* 99–119.
20. Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. Database Tomography for Information Retrieval. *J. Info. Sci.* **1997,** *23(4),* 301–311.
21. Greengrass, E. Information Retrieval: An Overview. National Security Agency. 1997. TR-R52-02-96, 28 February.
22. TREC (Text Retrieval Conference), Home Page, http://trec.nist.gov/.
23. Swanson, D. R. Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge. *Perspect. Biol. Med.* **1986,** *30(1),* 7–18.
24. Swanson, D. R.; Smalheiser, N. R. An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. *Artif. Intel.* **1997,** *91(2),* 183–203.
25. Kostoff, R. N. Stimulating Innovation. *International Handbook of Innovation;* Shavinina, L. V., Ed.; Elsevier Social and Behavioral Sciences: Oxford, UK, 2003, pp 388–400.
26. Gordon, M. D.; Dumais, S. Using Latent Semantic Indexing for Literature Based Discovery. *J. Am. Soc. Info. Sci.* **1998,** *49(8),* 674–685.
27. Goldman, J. A.; Chu, W. W.; Parker, D. S.; Goldman, R. M. Term Domain Distribution Analysis: A Data Mining Tool for Text Databases. *Methods Info. Med.* **1999,** *38,* 96–101.
28. Kostoff, R. N. Bilateral Asymmetry Prediction. *Med. Hypotheses* **2003,** *61(2),* 265–266.
29. Kostoff, R. N.; Green, K. A.; Toothman, D. R.; Humenik, J. A. Database Tomography Applied to an Aircraft Science and Technology Investment Strategy. *J. Aircraft.* **2000,** *37(4),* 727–730.
30. Kostoff, R. N.; Shlesinger, M.; Malpohl, G. Fractals Roadmaps Using Bibliometrics and Database Tomography. *Fractals* **2004,** *12,* 1.
31. Viator, J. A.; Pestorius, F. M. Investigating Trends in Acoustics Research from 1970–1999. *J. Acoust. Soc. Am.* **2001,** *109(5),* 1779–1783.
32. Kostoff, R. N.; Shlesinger, M.; Tshiteya, R. Nonlinear Dynamics Roadmaps Using Bibliometrics and Database Tomography. *Int. J. Bifurcat. Chaos* **2004**.
33. Davidse, R. J.; Van Raan, A. F. J. Out of Particles: Impact of CERN, DESY, and SLAC Research to Fields Other than Physics. *Scientometrics* **1997,** *40(2),* 171–193.
34. Kostoff, R. N.; Del Rio, J. A.; García, E. O.; Ramírez, A. M.; Humenik, J. A. Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. *J. Am. Soc. Info. Sci. Technol.* **2001,** *52(13),* 1148–1156.
35. Narin, F. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity* (monograph); NSF C-637. National Science Foundation: 1976; Contract NSF C-627. NTIS Accession No. PB252339/AS.
36. Garfield, E. History of Citation Indexes for Chemistry—A Brief Review. *JCICS* **1985,** *25(3),* 170–174.
37. Schubert, A.; Glanzel, W.; Braun, T. Subject Field Characteristic Citation Scores and Scales for Assessing Research Performance. *Scientometrics* **1987,** *12(5/6),* 267–291.
38. Narin, F.; Olivastro, D.; Stevens, K. A. Bibliometrics Theory, Practice, and Problems. *Eval. Rev.* **1994,** *18(1),* 65–76.
39. Del Río, J. A.; Kostoff, R. N.; García, E. O.; Ramírez, A. M.; Humenik, J. A. Phenomenological Approach to Profile Impact of Scientific Research. *Adv. Complex Syst.* **2002,** *5,* 19–42 Also available at http://arxiv.org/physics/0112047.

40. Kostoff, R. N.; Bedford, C.; Del Rio, J. A.; Cortes, H. D.; Karypis, G. Science and Technology Text Mining: Citation Mining of Macromolecular Mass Spectrometry. DTIC Technical Report. http:\\stinet.dtic.mil\.
41. Kostoff, R. N. The Use and Misuse of Citation Analysis in Research Evaluation. *Scientometrics* **1998,** *43(1),* 27–43.
42. MacRoberts, M.; MacRoberts, B. Problems of Citation Analysis. *Scientometrics* **1996,** *36(3),* 435–444.
43. Smith, R. D.; Loo, J. A.; Edmonds, C. G.; Barinaga, C. J.; Udseth, H. R. New Developments in Biochemical Mass Spectrometry-Electrospray Ionization. *Anal. Chem.* **1990,** *62(9),* 882–899.
44. Loo, J. A.; Edmonds, C. G.; Smith, R. D. Primary Sequence Information from Intact Proteins by Electrospray Ionization Tandem Mass Spectrometry. *Science* **1990,** *248(4952),* 201–204.
45. Loo, J. A.; Udseth, H. R.; Smith, R. D. Peptide and Protein Analysis by Electrospray Ionization Mass Spectrometry and Capillary Electrophoresis Mass Spectrometry. *Anal. Biochem.* **1989,** *179(2),* 404–412.
46. Karas, M.; Bachmann, D.; Hillenkamp, F. Influence of the Wavelength in High-Irradiance Ultraviolet-Laser Desorption Mass Spectrometry of Organic Molecules. *Anal. Chem.* **1985,** *57(14),* 2935–2939.
47. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60(20),* 2299–2301.
48. Karas, M.; Bahr, U.; Hillenkamp, F. UV Laser Matrix Desorption Ionization Mass Spectrometry of Proteins in the 100,000 Dalton Range. *Int. J. Mass Spectrom. Ion Processes* **1989,** *92,* 231–242.
49. Jaeger, H. M.; Nagel, S. R. Physics of the Granular State. *Science* **1992,** *256,* 1523–1531.
50. Cutting, D. R.; Karger, D. R.; Pedersen, J. O.; Tukey, J. W. Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections. *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval;* Copenhagen, Denmark, June, 1992, pp 318–329.
51. Guha, S.; Rastogi, R.; Shim, K. CURE: An Efficient Clustering Algorithm for Large Databases. *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data;* Seattle, Washington, June, 1998, pp 73–84.
52. Hearst, M. A. The Use Of Categories and Clusters in Information Access Interfaces. In *Natural Language Information Retrieval;* Strzalkowski, T., Ed.; Kluwer Academic Publishers, 1998.
53. Karypis, G.; Han, E.-H.; Kumar, V. Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *IEEE Comp. Special Issue on Data Analysis and Mining.* **1999,** *32(8),* 68–75.
54. Prechelt, L.; Malpohl, G.; Philippsen, M. Finding Plagiarisms Among a Set of Programs with JPlag. *J. Univ. Comput. Sci.* **2002,** *8(11),* 1016–1038.
55. Rasmussen, E. Clustering Algorithms. In *Information Retrieval Data Structures and Algorithms;* Frakes, W. B.; Baeza-Yates, R., Eds.; Prentice Hall: Upper Saddle River, NJ, 1992.
56. Steinbach, M.; Karypis, G.; Kumar, V. A Comparison of Document Clustering Techniques; Department of Computer Science and Engineering, University of Minnesota: 2000. Technical Report no. 00–034.
57. Willet, P. Recent Trends in Hierarchical Document Clustering: A Critical Review. *Info. Process. Management* **1988,** *24,* 577–597.
58. Wise, M. J. String Similarity via Greedy String Tiling and Running Karb-Rabin Matching; Dept. of CS, University of Sidney: ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps. 1992.
59. Zamir, O.; Etzioni, O. Web Document Clustering: A Feasibility Demonstration. *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval;* Zurich, Switzerland, August, 1998, pp 46–54.
60. Karypis, G. *CLUTO—A Clustering Toolkit*; http://www.cs.umn.edu/~cluto. 2002.
61. Zhao, Y.; Karypis, G. Criterion Functions For Document Clustering: Experiments and Analysis. Machine Learning. In press. 2003.