Duplicate Publication and 'Paper Inflation' in the Fractals Literature

Ronald N. Kostoff,[α] Dustin Johnson,[α] J. Antonio Del Rio,[β] Louis A. Bloomfield,[γ] Michael F. Shlesinger,[α] Guido Malpohl[δ] and Hector D. Cortes[β]

[α]Office of Naval Research, Arlington, VA, USA; [β]Centro de Investigación en Energía UNAM, Temixco, Mor. México; [γ]University of Virginia, Charlottesville, VA, USA; [δ]University of Karlsruhe, 76128 Karlsruhe, Germany

Keywords: Text Mining; Redundant Publications; Text Matching; Paper Inflation; Document Plagiarism; Concept Matching; Fractals; Greedy String Tiling; CopyFind; Data Compression.

ABSTRACT: *The similarity of documents in a large database of published Fractals articles was examined for redundancy. Three different text matching techniques were used on published Abstracts to identify redundancy candidates, and predictions were verified by reading full text versions of the redundancy candidate articles. A small fraction of the total articles in the database was judged to be redundant. This was viewed as a lower limit, because it excluded cases where the concepts remained the same, but the text was altered substantially.*

Far more pervasive than redundant publications were publications that did not violate the letter of redundancy but rather violated the spirit of redundancy. There appeared to be widespread publication maximization strategies. Studies that resulted in one comprehensive paper decades ago now result in multiple papers that focus on one major problem, but are differentiated by parameter ranges, or other stratifying variables. This 'paper inflation' is due in large part to the increasing use of metrics (publications, patents, citations, etc) to evaluate research performance, and the researchers' motivation to maximize the metrics.

The views in this paper are solely those of the authors, and do not necessarily represent the views of the Department of the Navy or any of its components, UNAM, University of Virginia or University of Karlsruhe.

Address for correspondence: Dr. Ronald N. Kostoff, Office of Naval Research, 875 North Randolph Street, Arlington, VA 22217, USA; email: kostofr@onr.navy.mil.

Paper received, 2 September 2005: revised, 24 November 2005: accepted, 12 January 2006.

1353-3452: 2006 Opragen Publications, POB 54, Guildford GU1 2YF, UK. http://www.opragen.co.uk

1. BACKGROUND

Concept matching in textual documents has become important in myriad contexts and applications. Identifying document clusters, whether for discovery of new knowledge, ease of routing, estimation of effort levels, or improved information retrieval, is becoming increasingly valuable as the volume of documentation in electronic format explodes. Plagiarism of documents has become a more serious problem with the wider availability of Web documents and the increased difficulty of heritage traceability. Increasing emphasis on simplistic metrics in the evaluation of research effort encourages researchers to maximize publication bibliometrics, including publishing similar concepts in multiple forums.

A number of studies have been performed on different aspects of concept matching in text, in order to address some of the applications described above. These include plagiarism,[1-8] duplicate/redundant publication,[9-15] text/document clustering,[16-21] and information retrieval.[22-26] These studies have shown that, in general, identifying similar documents through concept matching is quite difficult. A concept can be expressed in many word formats and combinations. The tools that are commonly used for detecting similar documents work on matching the concept expressions, or words/ phrases, and can be viewed as text matching. Much software for text matching is commercially available, prototypically available, and under development as well. Text matching is straight-forward, to some degree almost mechanistic. Most real-world applications intrinsically require concept matching, but in most cases have to settle for text matching.

In the course of a text mining study on the discipline of Fractals,[27] the first author noticed a few journal articles that appeared to be replications, or near replications. This phenomenon had been observed in other discipline text mining studies as well. Since text mining involves quantitative analysis of word/ phrase occurrences, and since one underlying assumption is that the documents from which these words/ phrases are extracted are relatively unique (i.e., more or less independent), then replicate publication of essentially the same article in multiple journals would skew the quantitative results.

It was desired to estimate the degree of duplicate publishing for an expanded version of the Fractals database. The first author assembled a team of experts in the fields of text similarity, text mining, and Fractals, and initiated a study of duplication in this database.

2. OVERVIEW

There were two de facto objectives for this study. The first objective was to examine different text matching techniques for their capabilities in identifying potentially duplicate documents. The second objective was to estimate the levels of different types of redundant documents in a Fractals database.

To achieve these objectives, the following conceptual approach was used. First, the database was generated. Second, three text matching approaches were applied to paper

Abstracts to quantify similarity of these Abstracts. Third, the full-text versions of the potentially most similar documents were obtained, and manually compared by experts for a final judgment of similarity. The next section describes these steps.

3. APPROACH

3.1 Database Generation

A key step in the Fractals literature analysis is the generation of the database to be used for processing. For the present study, the SCI database (including both the Science Citation Index and the Social Science Citation Index) was used. The approach used for query development was the first author's iterative relevance feedback concept of Simulated Nucleation.[28]

Science Citation Index/ Social Science Citation Index (SCI) [SCI, 2002]

The retrieved database used for analysis consists of selected journal records (including the fields of authors, titles, journals, author addresses, author keywords, Abstract narratives, and references cited for each paper) obtained by searching the Web version of the SCI for Fractals articles. At the time the final data was extracted for the present paper (Fall 2002), the version of the SCI used accessed about 5600 journals (mainly in physical, engineering, and life sciences basic research) from the Science Citation Index, and over 1700 journals from the Social Science Citation Index.

The SCI database selected represents a fraction of the available Fractals (mainly research) literature, that in turn represents a fraction of the Fractals S&T actually performed globally.[29] It does not include the large body of classified literature, or company proprietary technology literature. It does not include technical reports or books or patents on Fractals. It covers a finite slice of time (2000-2002). The database used represents the bulk of the peer-reviewed high quality Fractals research literature, and is a representative sample of all Fractals research in recent times.

To extract the relevant articles from the SCI, the Title, Keyword, and Abstract fields were searched using a query of terms relevant to Fractals. The resultant Abstracts were culled to those relevant to Fractals. The final efficient query, consisting of the highest marginal utility terms, is shown in Appendix (1), p. 554.

3.2 Text-Similarity Algorithms

The most thorough way of identifying all duplications and plagiarisms would be a manual comparison of all full text versions of the database records. This process would capture even those duplications and plagiarisms where the language was completely changed but the concept remained the same. However, for an 8352 record database such as Fractals, this procedure would involve tens of millions of full text manual comparisons. Limiting the scope to duplications would still require manual comparison of all full text versions of records that are linked by at least one common author

(single-link clustering). Again, the large number of full text manual comparisons that would be required is not feasible given limited resources. It was decided that the best approximation to identifying all duplicates was to manually evaluate the full text of the subset of records having the highest probability of being duplicates. This probability was determined by computer-based comparison of the Abstracts of all articles in the database.

To determine the likelihood of duplication, three distinct computational approaches were examined. Each approach employs a similar operational structure: comparing each record in the database with every other record and generating similarity metrics based on these comparisons. Abstracts, as well as references, were used as input to account for trends within author groups (e.g. repetitive self-citing, shared reference collections, etc.) in addition to conceptual similarities. The text-similarity algorithms characterizing each approach can be described as follows:

Greedy String Tiling (GST):

GST clustering forms groups of documents based on the cumulative sum of shared strings of words. Each group is termed a cluster. The number of records in each cluster, and the highest frequency technical keywords in each cluster, are two outputs central to this analysis. This process is described in more detail in Kostoff et al, 2005.[30]

Copyfind Algorithm:

The Copyfind algorithm examines a collection of documents, extracting the text portions of those documents and searching for matching words in phrases of a specified minimum length. When two files are found that share enough words in those phrases, Copyfind generates HTML report files. These reports contain the document text with the matching phrases underlined. The application of this process to the present study is described in more detail in Kostoff et al, 2005.[30]

Data Compression (Entropy) Clustering:

The compression algorithm approach[31] assumes that the entropy of a string can be measured when this string is zipped (compressed). The main idea is that when one compresses two strings sequentially, the compression rate will increase if the second string is similar to the first one, and then the zipped string will have less disorder (entropy) than the previous two strings. The application of this process to the present study is described in more detail in Kostoff et al, 2005.[30]

Though each of these algorithms is intrinsically different, it was hoped that they would produce complementary results exemplifying common trends. Both GST and Copyfind assign a similarity index between 0 and 100 to each pair of articles, with 0 indicating no similarity and 100 indicating an exact match. The Data Compression Clustering technique, which was originally modified to produce metrics that paralleled those used in our manual evaluation, was also normalized to produce a comparably scaled similarity index. Likely candidates for paper reuse were identified based on a

threshold function using the highest index assigned by any one of the algorithms as input. To determine the accuracy and appropriateness of each computational approach, candidate Abstracts were then compared by manual evaluation and were given a similarity ranking based on specified criteria (Appendix (2), p. 554). As manual evaluation is often time and resource intensive, the threshold function was set according to this constraint. Approximately 450 of the 8352 articles were selected for an initial manual review, which focused on Abstracts rather than full text articles.

Following this review, another threshold was used in the same manner to identify a smaller subset of article pairs whose full text versions were to be obtained and examined manually. This was deemed necessary to establish how effectively each Abstract represented its full text article's actual content.

3.3 Algorithm Suitability Metrics

To determine how well-suited each algorithm is to the identification of redundant publication, it became necessary to devise a system of metrics. The term "well-suited" is used here instead of "accuracy" because each of the techniques demonstrates varying strengths in different applications and thus a generalized statement of quality would be unwarranted.

First, a system was developed for mapping most of the algorithmically-produced indices, which ranged from 0 (least similar) to 100 (most similar), to the integer scale used in our manual evaluations, which ranges from 0 (least similar) to 4 (most similar). For Data Compression Clustering this was not necessary as the original output was already in this form. For the remaining algorithms, the output was grouped into a series of "bands," each containing a range of indices to be mapped to a specified score between 0 and 4. Because each method produced very different distributions of indices, the size of the bands corresponding to each algorithm was uniquely determined by the output distribution of that algorithm. Additionally, since a generally higher concentration of indices was observed in the upper spectrum, the five bands were arranged beginning at the high similarity end and proceeding toward the low similarity end. This created a larger band ranging from 0 to the lowest value contained in the next highest band (a value determined by the band size), but ultimately resulted in a better fit to all observed data. From this point forward, "band size" will be used to refer to the size of the four upper bands.

An optimization technique was used to simultaneously determine the band size and the suitability of each algorithm. Since one objective was to identify the computer-based method whose results most closely resembled those produced from manual comparisons, the manual ratings were used as a standard for comparison (benchmark). Absolute deviations between the remapped algorithmic indices and corresponding manual ratings were calculated. The average value of these deviations was then computed to determine the overall closeness of each automated process to the established standard, producing a precise measure of suitability. Band sizes for individual mappings were chosen to maximize this suitability metric for each algorithm's output separately. This straightforward metric was chosen – rather than more rigorous statistical measurements, which would have contributed little to our purpose – to keep the results relevant and widely accessible.

4. RESULTS

4.1 Abstracts Analysis

GST Approach

The Greedy String Tiling algorithm was applied, with one-word resolution (i.e., word strings of unit length were included in the text comparisons), to the Abstract field and the References field of the full Fractals database. Results were compared to the manually produced duplication projection scores. Very high similarity indices appeared to correlate with high duplication projections, and lower similarity scores correlated moderately well with duplication projections that could not be ruled out from reading the Abstracts alone. For the References case, the correlations were much weaker.

Manual evaluation of the Abstracts showed that many of the records having strong textual similarity and shared References were modest variants of the same problem. The authors appeared to have done one substantive study, and then subdivided the written product among two or more papers. Since these different papers were actually parts of one large paper, there was little need to change References or the Abstract text. When the similarity scores for Abstracts and References were combined, the trends were sharpened somewhat

The conclusion to be drawn from results charts is that, with the GST approach, the highest similarity indices are probably a good predictor of duplications, particularly when the similarity indices from Abstracts and References are combined. The low range combined similarity indices probably reflect minimal or no duplication. The mid-range similarity indices, where the 1 and 2 scored duplication records mainly exist, provide inconclusive projections based on manual Abstract evaluation alone.

Copyfind Approach

The Copyfind algorithm was applied with strings of 6 words in a row as the minimal phrase match and a moderate tolerance of imperfections between phrases it identified as matching. It was also applied to both the Abstract and References fields of the full Fractals database and the results were plotted against the manually produced scores to test for correlation.

Similar trends to those in the GST approach were observed in the cases of both Abstracts and References. When the similarity scores for Abstracts and References were combined, the trends were sharpened only very slightly.

Data Compression (Entropy) Approach

The entropic algorithm based on data compression was applied to the Abstract field of the full Fractals database. In this approach, very high similarity indices correlated very well with high duplication projections and low similarity indices correlated very well with low duplication projections. For the most part, the overall correlation of the algorithmically produced similarity indices to the manual duplication projections was quite high.

Overall Algorithm Comparisons for Abstracts

A summary of the suitability metrics associated with each algorithm is given in Table (1):

TABLE (1) – OVERALL SUITABILITY METRICS

APPROACH	AVERAGE DIFFERENCE
GST	0.2756
COPYFIND	0.4867
ENTROPY	0.1689

The data compression (entropy) approach applied to Abstracts produced results that most closely mirrored those produced by manual evaluation of Abstracts. The next best approach was Greedy String Tiling, whose average difference was 0.4867, 63% higher than the data compression approach (Note here that a higher average difference corresponds to a lower degree of overall suitability). The Copyfind algorithm was third best, with an average difference of 0.4867 (188% higher than the data compression approach). Because the GST approach is only moderately less suitable according to the given metrics, it will probably still give accurate results for most practical purposes.

4.2 Full Text Analysis

Using the Abstract-based manual duplication projections, the 136 articles that appeared to be the most likely candidates for duplication were chosen for full-text manual evaluation. Of these 136 articles, 119, or 87.5%, were successfully obtained and reviewed. Criteria similar to those used in manual Abstract evaluation were used in the full-text evaluation (see Appendix (3), p. 554). The rankings used in the full text analysis were also integers ranging from 0 to 4.

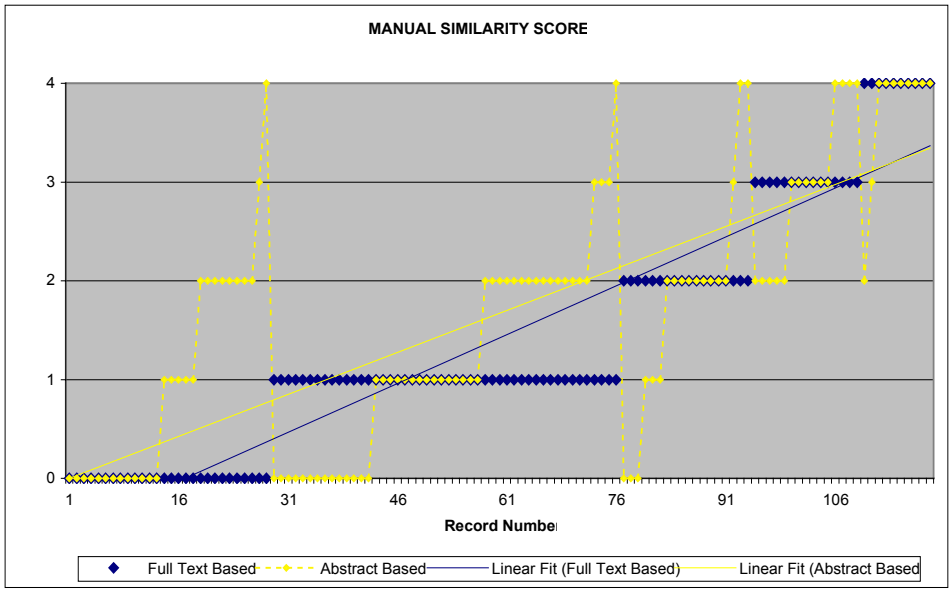
Results from this analysis showed a moderate correlation between Abstract scores and full text scores, which can be seen in Figure (1) overleaf. The plot uses the record number as the x-axis metric (an arbitrary index in this case) and the manual score as the y-axis. The data are sorted first according to full text scores and then according to Abstract scores. This allows one to observe the number of article pairs given the same score by both methods as well as the number of article pairs whose similarity was over- or under-estimated by evaluation of the Abstract alone.

For approximately 37% of the article pairs, the Abstract score was higher than the full-text score (false positive). In these cases, it appeared that the author(s) had become comfortable enough with the wording of a previously used Abstract to reuse it in a comparable but technically dissimilar study. For 26% of the article pairs, the full-text score was higher than the Abstract score (false negative). These represent the more problematic cases where articles are essentially reused, but Abstracts are changed more significantly to avoid detection by journals. Unfortunately, to discover how many of these cases actually exist in the literature would require an exhaustive comparison of

every full-text article pair, which is beyond the scope of this study. For the remaining 37% of article pairs, the manual Abstract and full text scores were the same.

Linear fit lines are included on the plot to indicate the overall trend similarity of the Abstract and full-text scores. Despite deviations between scores, a reasonable correlation can be observed. However, if our previous method is used to determine the suitability of Abstract-based comparisons to predicting the actual study similarity as indicated by full-text comparisons, the average difference is calculated to be 0.7863 – a value higher than those produced by any of the three approaches examined. This observation reinforces our previous conviction that, for a truly comprehensive analysis of the dual-use problem, full text versions of the entire literature of interest must be used.

FIGURE (1) – COMPARISON OF MANUAL ABSTRACT AND FULL TEXT SIMILARITY SCORES



5. DISCUSSION AND CONCLUSIONS

GST

For the Abstracts case, very high similarity indices appear to correlate with high duplication projections, and lower similarity scores correlate moderately well with duplication projections that can't be ruled out from reading the Abstracts alone.

For the References case, the correlations are much weaker. However, the References results are believed to reflect an important reality. If one does a literature search in the SCI using common references as a criterion for retrieving related records, one finds the following. For papers written by different authors, relatively few

references are shared, even though the topics can be quite similar. In the present study, for the high ranking matches that in many cases involve the same author groups, shared references are quite high. This is probably because the authors are familiar with a finite group of references and tend to refer to these, in addition to repetitive self-citing.

Manual evaluation of the Abstracts showed that many of the records having strong textual similarity and shared references were modest variants of the same problem. The authors appeared to have done one substantive study, and then subdivided the written product among two or more papers. Since these different papers were actually parts of one large paper, there was little need to change References or the Abstract text. When the similarity scores for Abstracts and References were combined, the trends were sharpened somewhat

The conclusion to be drawn is that, for the GST approach, the highest similarity indices are probably a good predictor of duplications, particularly when the similarity indices from Abstracts and References are combined. The low range combined similarity indices probably reflect minimal or no duplication. The mid-range similarity indices provide inconclusive projections based on manual Abstract evaluation alone.

COPYFIND

The Copyfind algorithm was applied with strings of 6 words in a row as the minimal phrase match and a moderate tolerance of imperfections between phrases it identified as matching. It was also applied to both the Abstract and References fields of the full Fractals database and the results were plotted against the manually produced scores to test for correlation. Similar trends to those in the GST approach were observed in the cases of both Abstracts and References.

DATA COMPRESSION

The data compression results indicate that very high similarity indices correlate very well with high duplication projections and low similarity indices correlate very well with low duplication projections. With the exception of the records receiving a score of two, which exhibit some degree of ambiguity, the overall correlation of the algorithmically produced similarity indices to the manual duplication projections is quite high.

GENERAL CONCLUSIONS

The prediction approaches examined in this paper (and the subsequent manual evaluation of full texts) have identified a small number of redundant Fractals publications in a much larger sample of publications in SCI-accessed journals. However, while the fraction of redundant publication found in this study is extremely small, that value should be viewed as a lower limit. Redundancy among 1) papers in SCI journals and 2) papers in journals not accessed by the SCI could not be evaluated by the present SCI-focused techniques. Papers whose Abstracts had substantial wording changes to new terminology (as opposed to wording re-arrangements) could

not be accessed by the present techniques as well. Either algorithms with thesaurus-access capabilities would have to be used to detect terminology changes for the same concept, or single-link clustering would have to be used for author names, and full text of all papers in each cluster would have to be evaluated manually, a massive undertaking.

Additionally, the computer-based similarity prediction algorithms based on Abstracts are only moderately successful in predicting redundant publications. Full text analysis is required for more than cursory evaluations.

Far more pervasive than redundant publications are publications that do not violate the letter of redundancy but rather violate the spirit of redundancy. There appear to be widespread publication maximization strategies. Studies that resulted in one comprehensive paper decades ago now result in multiple papers that focus on one major problem, but are differentiated by parameter ranges, or other stratifying variables.

Rather than addressing the major problems to be solved, across a relatively broad swath of topics, a number of researchers are focusing on a set of experimental or theoretical tools, and maximizing the number of papers they can generate by modestly varying a number of parameters. The trend among this group is tool-centric, rather than problem-centric.

REFERENCES

1. Braumoeller BF, Gaines BJ. (Dec 2001) Actions Do Speak Louder Than Words: Deterring Plagiarism with the Use of Plagiarism-Detection Software. *PS-Political Science & Politics* **34** (4): 835-839.
2. Monostori K, Finkel R, Zaslavsky A, Hodasz G, Pataki M. (2002) Comparison of Overlap Detection Techniques. Computational Science-ICCS 2002, Pt I, *Proceedings Lecture Notes In Computer Science* **2329**: 51-60.
3. Cook DE, Mellor L, Frost G, Creutzburg R. (2002) Knowledge Management and the Control of Duplication. Engineering and Deployment of Cooperative Information Systems, *Proceedings Lecture Notes in Computer Science* **2480**: 396-402.
4. Hoad TC, Zobel J. (Feb 1 2003) Methods for Identifying Versioned and Plagiarized Documents. *Journal of the American Society for Information Science and Technology* **54** (3): 203-215.
5. Gilbert FJ, Denison AR. (Jul 2003) Research Misconduct. *Clinical Radiology* **58** (7): 499-504.
6. Pecorari D. (Dec 2003) Good and Original: Plagiarism and Patchwriting in Academic Second-Language Writing. *Journal of Second Language Writing* **12** (4): 317-345.
7. Chen X, Francia B, Li M, Mckinnon B, Seker A. (Jul 2004) Shared Information and Program Plagiarism Detection. *IEEE Transactions on Information Theory* **50** (7): 1545-1551.
8. Bao JP, Shen JY, Liu XD, Liu HY, Zhang XD. (2004) Finding Plagiarism Based on Common Semantic Sequence Model. Advances in Web-Age Information Management: *Proceedings Lecture Notes in Computer Science* **3129**: 640-645.
9. Doherty M. (Nov 1996) Misconduct of Redundant Publication. *Annals of the Rheumatic Diseases* **55** (11): 783-785.
10. Jefferson T. (Apr 1998) Redundant Publication in Biomedical Sciences: Scientific Misconduct or Necessity? *Science and Engineering Ethics* **4** (2): 135-140.
11. Schein M, Paladugu R. (Jun 2001) Redundant Surgical Publications: Tip of the Iceberg? *Surgery* **129** (6): 655-661.

12. Bailey BJ. (Mar 2002) Duplicate Publication in the Field of Otolaryngology-Head and Neck Surgery. *Otolaryngology-Head and Neck Surgery* **126** (3): 211-216.
13. Von Elm E, Poglia G, Walder B, Tramer MR. (Feb 25 2004) Different Patterns of Duplicate Publication - An Analysis of Articles Used in Systematic Reviews. *Jama-Journal of the American Medical Association* **291** (8): 974-980.
14. Mojon-Azzi SM, Jiang XY, Wagner U, Mojon DS. (May 2004) Redundant Publications in Scientific Ophthalmologic Journals - the Tip of the Iceberg?. *Ophthalmology* **111** (5): 863-866.
15. Gwilym SE, Swan MC, Giele H. (Jul 2004) One in 13 'Original' Articles in the Journal of Bone and Joint Surgery are Duplicate or Fragmented Publications. *Journal of Bone and Joint Surgery-British Volume* **86b** (5): 743-745.
16. Maderlechner G, Suda P, Bruckner T. (Nov 1997) Classification of Documents by Form and Content. *Pattern Recognition Letters* **18** (11-13): 1225-1231.
17. Atlam ES, Fuketa M, Morita K, Aoe J. (Nov 2003) Documents Similarity Measurement Using Field Association Terms. *Information Processing & Management* **39** (6): 809-824.
18. Dobrynin V, Patterson D, Rooney N. (2004) Contextual Document Clustering. Advances in Information Retrieval, *Proceedings Lecture Notes in Computer Science* **2997**: 167-180.
19. Shin K, Han SY, Gelbukh A. (2004) Advanced Clustering Technique for Medical Data Using Semantic Information. Micai 2004: Advances in Artificial Intelligence *Lecture Notes in Computer Science* **2972**: 322-331.
20. Li WY, Ng WK, Lim EP. (2004) Spectral Analysis of Text Collection for Similarity-Based Clustering. Advances in Knowledge Discovery and Data Mining, *Proceedings Lecture Notes in Artificial Intelligence* **3056**: 389-393.
21. Bansal N, Blum A, Chawla S. (Jul-Sep 2004) Correlation Clustering. *Machine Learning* **56** (1-3): 89-113.
22. Salton G, Buckley C. (Aug 30 1991) Text Matching for Information-Retrieval. *Science* **253** (5023): 1012-1015.
23. Hui SC, Fong ACM. (2004) Document Retrieval from a Citation Database Using Conceptual Clustering and Co-Word Analysis. *Online Information Review* **28** (1): 22-32.
24. Leuski A, Allan J. (Jun 2004) Interactive Information Retrieval Using Clustering and Spatial Proximity. *User Modeling and User-Adapted Interaction* **14** (2-3): 259-288.
25. Muresan G, Harper DJ. (Aug 2004) Topic Modeling for Mediated Access to Very Large Document Collections. *Journal of the American Society for Information Science and Technology* **55** (10): 892-910.
26. Chang Y, Kim M, Ounis I. (2004) Construction of Query Concepts in a Document Space Based on Data Mining Techniques. Flexible Query Answering Systems, *Proceedings Lecture Notes in Artificial Intelligence* **3055**: 137-149.
27. Kostoff, RN, Shlesinger M, and Malpohl G. (March 2004) Fractals roadmaps using bibliometrics and database tomography. *Fractals.* **12**:1. 1-16.
28. Kostoff RN, Eberhart HJ., and Toothman DR. (1997) Database Tomography for information retrieval. *Journal of Information Science* **23**: 4.
29. Kostoff RN. (May 2000) The underpublishing of science and technology results. *The Scientist*. **14**:9. 6-6. 1.
30. Kostoff RN, Johnson D, Del Rio JA, Bloomfield LA, Shlesinger MF, and Malpohl G. (2005) Duplicate publication and 'paper inflation" in the fractals literature. DTIC Technical Report Number ADA440622 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA.
31. Benedetto D, Caglioti E, Loreto V. (Jan 28 2002) Language trees and zipping. *Physical Review Letters* **88** (4) 048702: 1-4.

APPENDICES

APPENDIX (1) – FRACTALS QUERY

Fractal* or Self-similar* or Self-organized Criticality or Multifractal or Anomalous Diffusion or Scale Invariant or Hausdorff Dimension or Diffusion Limited Aggregation or Fractional Brownian Motion or Mandelbrot or Lacunarity or Cantor Set or Nonfractal or Monofractal not Fractalkine*

APPENDIX (2) – DEFINITION OF SIMILARITY LEVELS FOR ABSTRACT/REFERENCES COMPARISON

Level 4 – Only difference between papers is the journal in which they are published. The titles are either the same or very similar. The Abstracts and references are essentially the same, with large blocks of common text. The judgment would be, based on the record examined, but before the full text versions have been compared, that there is a very high probability that the papers are duplicates.

Level 3 – Substantial "wordsmithing" has been performed. Words may have been re-arranged in the title and Abstract, and one or two references may have been added or subtracted. There are modest sized blocks of common text, most technical words and phrases are in common, but in different order. The judgment would be, based on the record examined, but before the full text versions have been compared, that there is a high probability that the papers are duplicates.

Level 2 – Tenses have been changed as well as words re-arranged, and perhaps there are larger modifications in the references. The judgment would be, based on the record examined, but before the full text versions have been compared, that there is a medium probability that the papers are duplicates.

Level 1 – Extensive substitutions of synonyms have been made, but the fundamental concepts are unchanged. The judgment would be, based on the record examined, but before the full text versions have been compared, that there is a possibility that the papers are duplicates.

Level 0 – Seemingly dissimilar. The judgment would be, based on the record examined, but before the full text versions have been compared, that there is little to no possibility that the papers are duplicates.

APPENDIX (3) – DEFINITION OF SIMILARITY LEVELS FOR FULL TEXT COMPARISON

Level 4 - Essentially identical text and concept. Same title and Abstract; same references. Perhaps a couple of words changed.

Level 3 - Almost identical text and concept. Some shifting around of words. Perhaps title modified, but Abstract and references very similar. Objectives, approach, and results the same.

Level 2 - Similar text and almost identical concept. Concepts similar, but many words have been changed. Extensive use of synonyms. References quite similar. Objectives, approach, and results the same.

Level 1 - Similar text and complementary concepts. Much text in common, especially in 'boiler-plate' sections, Abstract, and references. Concepts in each paper are part of one larger concept. One parameter range may be studied in one paper; another parameter range studied in the second paper. Or, part one of a study may be in one paper, and part two in the other paper. Essentially, one large comprehensive paper has been divided into separate papers.

Level 0 - Different text and different concept. Two essentially different documents.