Duplicate Publication and 'Paper Inflation' in the Fractals Literature

By

Dr. Ronald N. Kostoff, Office of Naval Research,
875 North Randolph Street, Arlington, VA 22217
Phone: 703-696-4198; Fax: 703-696-3098
Internet: kostofr@onr.navy.mil

Mr. Dustin Johnson, Office of Naval Research
875 North Randolph Street, Arlington, VA 22217

Dr. J. Antonio Del Río, Centro de Investigación en Energía
UNAM, Temixco, Mor. México

Dr. Louis A. Bloomfield, University of Virginia
Charlottesville, VA

Dr. Michael F. Shlesinger, Office of Naval Research
875 North Randolph Street, Arlington, VA 22217

Mr. Guido Malpohl, University of Karlsruhe
Postfach 6980, 76128 Karlsruhe, Germany

(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)

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

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Duplicate Publication and "Paper Inflation" in the Fractals Literature	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research Dr. Ronald N. Kostoff 875 North Randolph Street Arlington, VA 22217**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**31**	

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.

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 (Braumoeller, 2001; Monostori, 2002; Cook, 2002; Hoad, 2003; Gilbert, 2003; Pecorari, 2003; Chen, 2004; Bao, 2004), duplicate/ redundant publication (Doherty, 1996; Jefferson, 1998;

Schein, 2001; Bailey, 2002; Von Elm, 2004; Mojon-Azzi, 2004; Gwilym, 2004), text/ document clustering (Maderlechner, 1997; Atlam, 2003; Dobrynin, 2004; Shin, 2004; Li, 2004; Bansal, 2004), and information retrieval (Salton, 1991; Hui, 2004; Leuski, 2004; Muresan, 2004; Chang, 2004). 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 to detect 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 [Kostoff et al, 2004], 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 Fractals 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.

APPROACH

2.1. 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 [Kostoff et al, 1997].

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 [Kostoff, 2000]. 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).

2.2. 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 Appendix (2).

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 Appendix (3).

Data Compression (Entropy) Clustering:

The compression algorithm approach [Benedetto et al, 2002] 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 Appendix (4).

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 (See Appendix (4)). 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 (5)). 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.

2.3. 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 range from 0 (least similar) to 100 (most similar), to the integer scale used in our manual evaluations, which ranged 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 creates 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 results 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

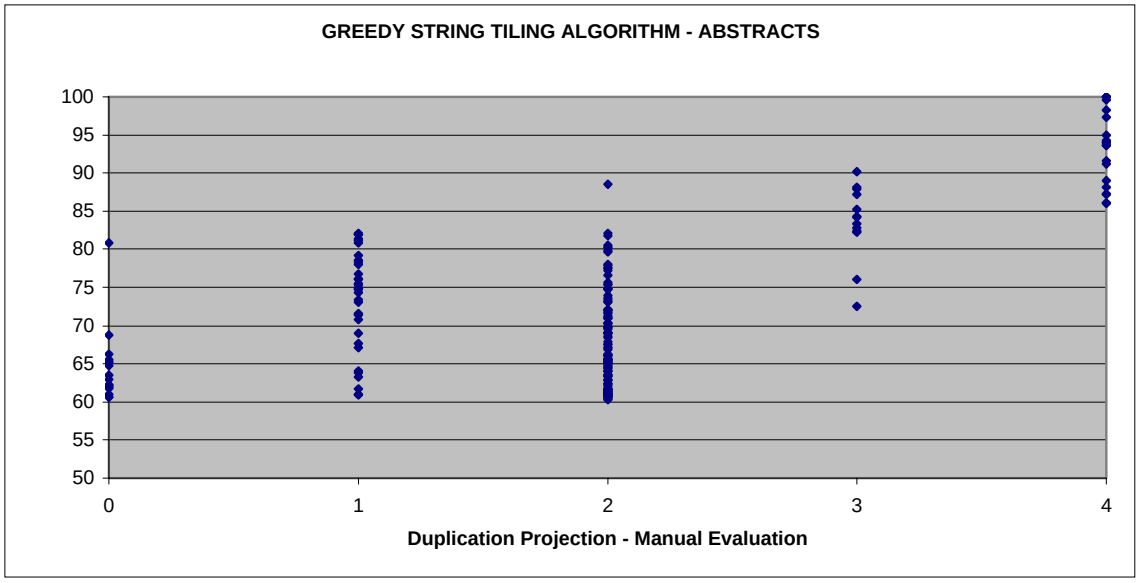
GST Approach

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 and were plotted against them as an initial qualitative test for correlation. In each plot shown, including similar plots in subsequent sections, the manual duplication projection is plotted on the x-axis with the algorithmically produced similarity index on the y-axis.

For the Abstracts case (FIG. (1)), the numbers of records with duplication scores of 0, 3, and 4 were approximately the same; larger numbers had a 1, and the largest were in the 2 category. Because the same set of manual scores was used in all comparisons, this trend is apparent in all subsequent plots as well.

In the case of GST, the 3 and 4 scored records are concentrated in the high similarity index region, with the 4 scored records occupying the higher segment, and the 3 scored records occupying the lower segment. The 1 and 2 scored records are distributed more or less uniformly over the mid-similarity index region, with the 2 scored records actually appearing to have slightly higher density at the lower end, and the 1 scored records appearing to have slightly higher density at the higher end. The observation from this chart is that 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. However, the 0 scored records overlap with the lower end of the 1 and 2 scored records, suggesting that indices of about 80 or below, and especially 70 or below, can reflect multiple duplication projections.

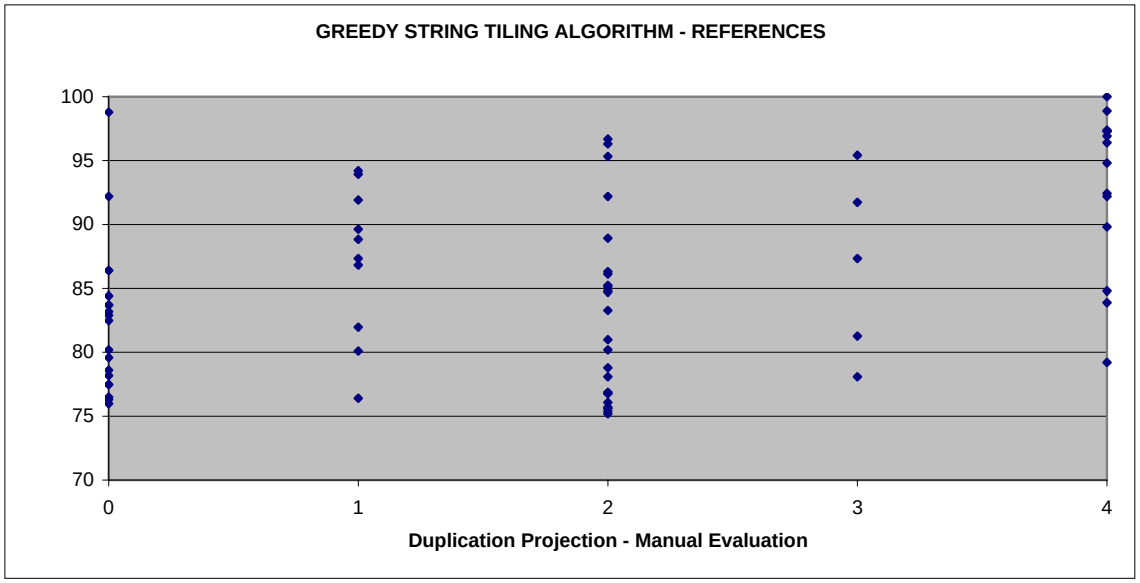
FIGURE (1) – ABSTRACT-BASED GST SIMILARITY INDEX VS. DUPLICATION PROJECTION



For the references case (FIG. (2)), the correlations are much weaker. For example, there is overlap among all categories in the 80-100 similarity index region, although the density of the 4 scored records is higher at the upper similarity index end and the density of the 0 and 2 scored projects is higher at the lower end. 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.

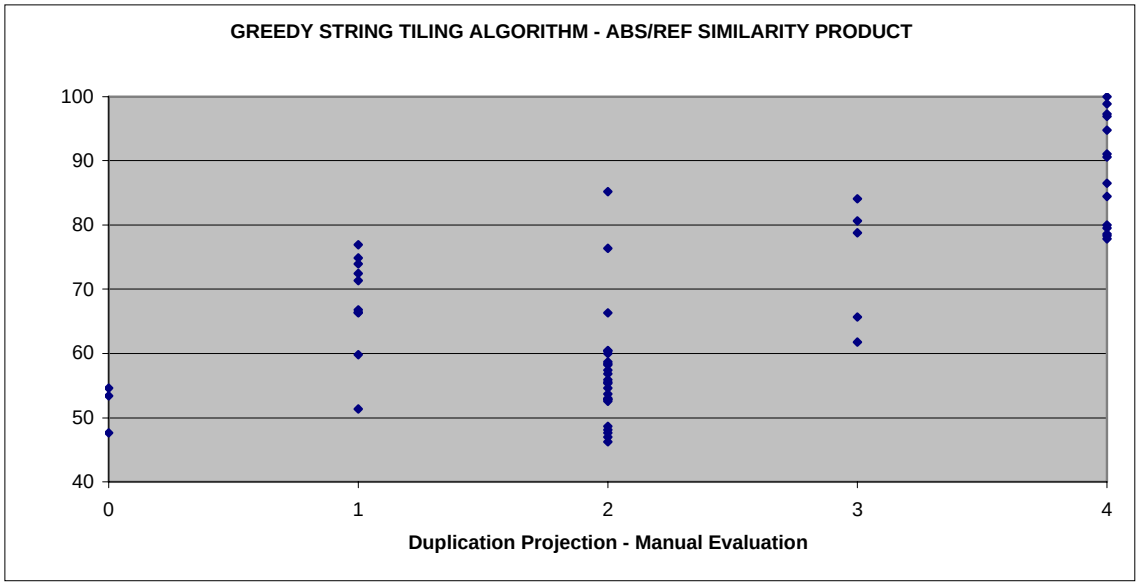
FIGURE (2) – REFERENCES-BASED GST SIMILARITY INDEX VS. DUPLICATION PROJECTION



Furthermore, and probably most importantly for explaining the somewhat counter-intuitive density inversion of the 1 and 2 scored records in the Abstracts plot, 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. These papers tended to receive either 0 or 1 scores, depending on the strength of the differences perceived. The papers with the 2 score may have had more of a difference in the Abstract text and the references than some of the 0 or 1 scored papers, but were judged to have been changed superficially, not intrinsically.

When the similarity scores for Abstracts and references were combined, the trends were sharpened somewhat. In a hybrid plot (FIG. (3)), the product of the Abstract and references similarity index is used as the y-axis metric. This plot shows minimal overlap between the 3 and 4 scored regions, and very modest overlap between the 3 scored records and the bulk of the 2 scored records. The same density inversion persists between the 1 and 2 scored records, for the same reasons stated above. Almost all of the 0 scored records have been eliminated.

FIGURE (3) – GST SIMILARITY INDEX PRODUCT VS. DUPLICATION PROJECTION



The conclusion to be drawn from these 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.

To determine the suitability of the GST algorithm to the current problem, Table (1) was generated according to the previously described procedure (SEC. 3.3). This table and the resultant mapping scheme (Table (2)) are shown below:

TABLE (1) – BAND SIZE OPTIMIZATION FOR GST

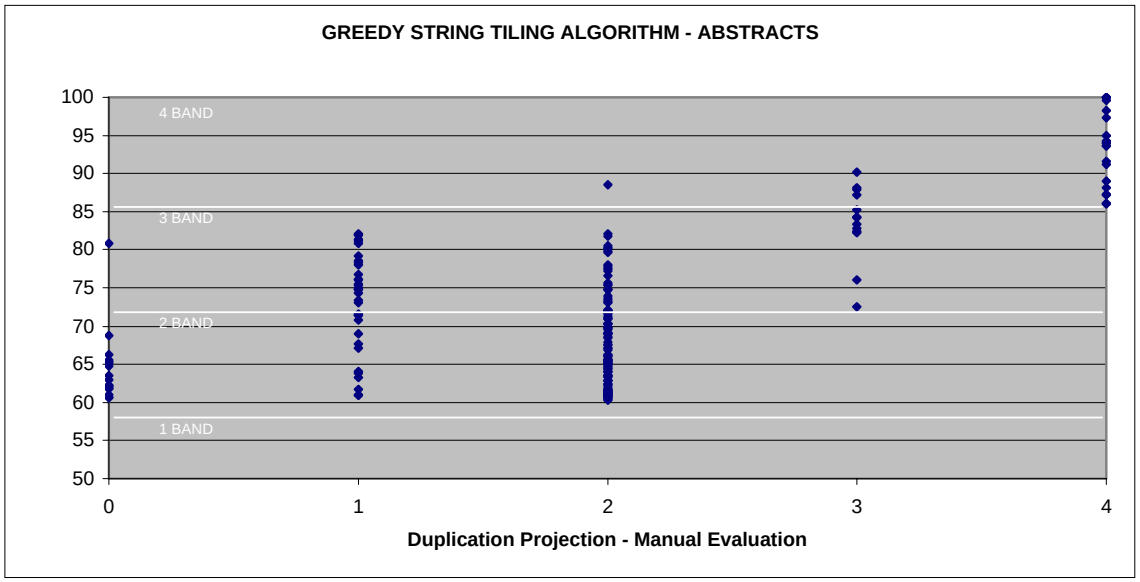
Band Size	Average Deviation
10	0.3178
11	0.3044
12	0.2844
13	0.2756
14	0.2756
15	0.3067
16	0.3400
17	0.3711
18	0.4333

TABLE (2) – MAPPING SCHEME FOR GST RESULTS

Similarity Index	New Mapping
86-100	4
72-85.99	3
58-71.99	2
44-57.99	1
0-43.99	0

Band sizes of 13 and 14 resulted in the smallest average deviation from the manual results, which was 0.2756 (14 was arbitrarily chosen in this case, as indicated by the highlighted entry in Table (1)). Figure (4) shows how the GST output is categorized according to these bands:

FIGURE (4) – GST SIMILARITY INDEX PRODUCT VS. DUPLICATION PROJECTION WITH BAND BREAKDOWN



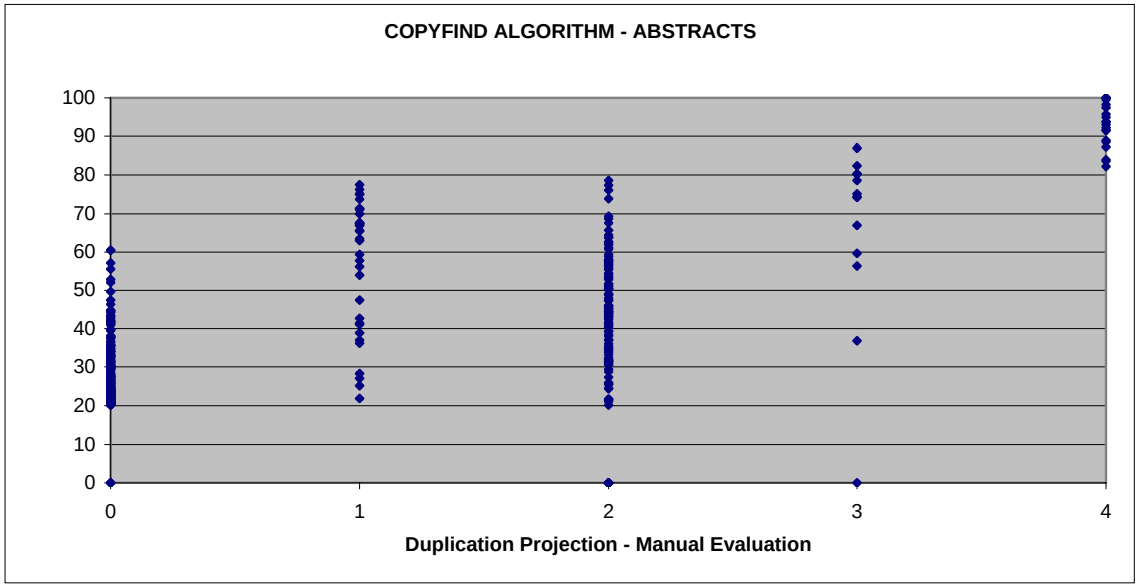
Copyfind Approach

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. Again, the 3 and 4 score records are concentrated in the high similarity index region with the 4 scores occupying the higher segment and the 3 scores occupying the lower segment (Figure (4)). The 1 and 2 score records are again distributed uniformly over the mid-similarity index region and exhibit a similar density inversion to the plots shown in the GST analysis. The 0 score records overlap with a more significant part of the 1 and 2 score records than seen in the GST, but the 4 score records overlap slightly less than the lower scored records and are more concentrated at the high end of the spectrum.

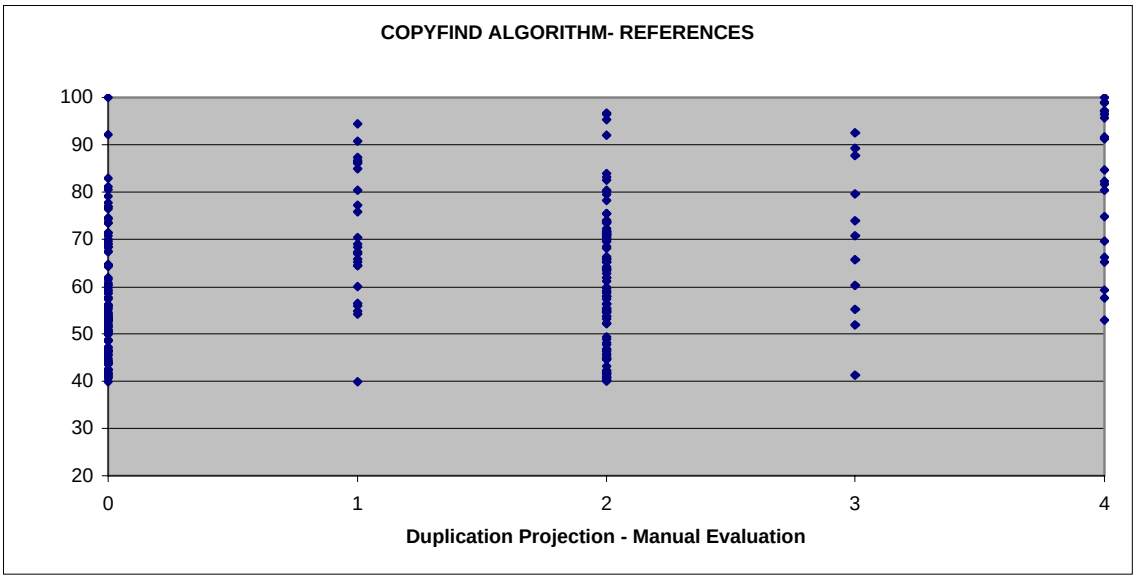
It may also be noted that the range of indices produced by the Copyfind algorithm is much wider than that produced by the GST algorithm. Where the majority of indices produced by GST lie between 60 and 100, the Copyfind algorithm ranges from 20 to 100. For this reason, the band-based normalization used in the determination of suitability metrics was crucial.

FIGURE (4) – ABSTRACT-BASED COPYFIND SIMILARITY INDEX VS. DUPLICATION PROJECTION



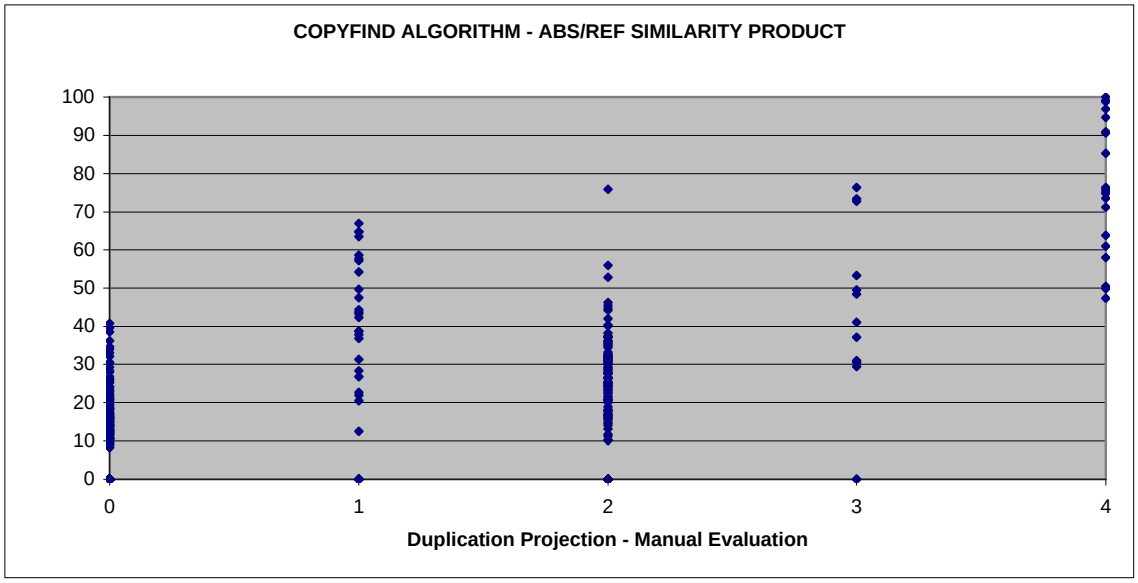
As with the GST approach, the references case (FIG. (5)) shows a much weaker correlation. In this case, there is overlap among all categories in the 50-100 similarity index region. Again, the density of the 4 scored records is higher at the upper similarity index end and the density of the 0 and 2 scored records is higher at the lower end, though the lower end in this case covers a broader range than with GST.

FIGURE (5) – REFERENCES-BASED COPYFIND SIMILARITY INDEX VS. DUPLICATION PROJECTION



When the similarity scores for Abstracts and references were combined, the trends were sharpened only very slightly. Another hybrid plot (FIG. (6)) was generated using the product of the Abstract and references similarity index as the y-axis metric. This plot shows a reduced and lowered range of 0 scores, with the majority of scores situated between similarity indices of 10 and 40, compared to approximate ranges of 20-60 and 40-80 for Abstracts and references respectively. However, the severity of overlap between the 3 and 4 scored records lies somewhere between the Abstract-only and references-only results, with a better correlation observed with the Abstracts. A density inversion between 1 and 2 scored records persists for the same reasons mentioned earlier.

FIGURE (6) – GST SIMILARITY INDEX PRODUCT VS. DUPLICATION PROJECTION



The suitability of the Copyfind algorithm was determined by generating Table (3), analogous to that used for the GST algorithm. Table (3), and the resultant mapping scheme (Table (4)), are shown below:

TABLE (3) – BAND SIZE OPTIMIZATION FOR COPYFIND

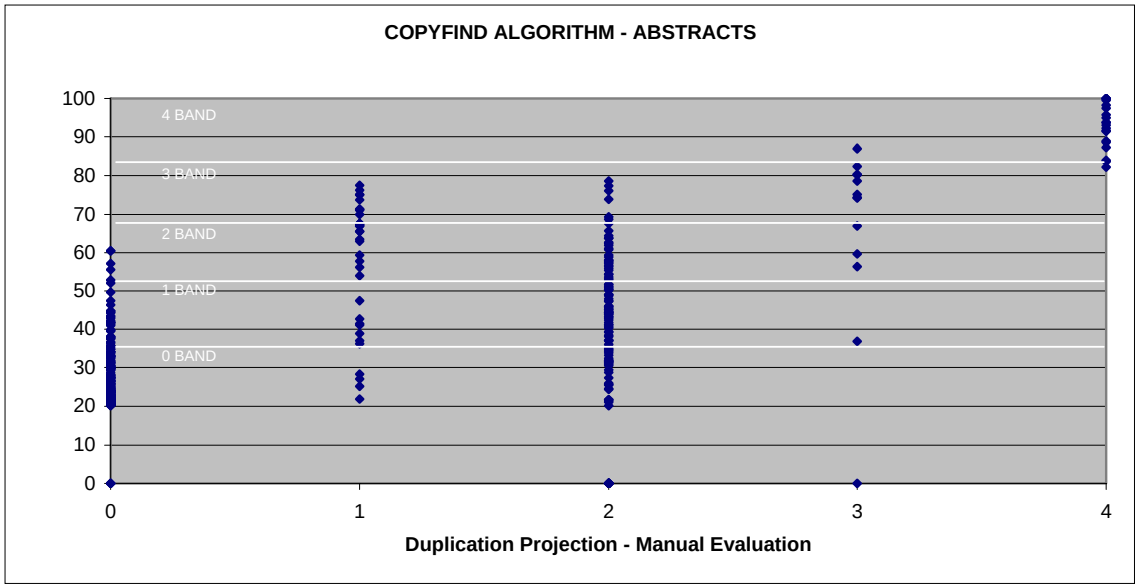
Band Size	Average Deviation
12	0.5511
13	0.5178
14	0.5000
15	0.4911
16	0.4867
17	0.5067
18	0.5444
19	0.6133
20	0.9467

TABLE (4) – MAPPING SCHEME FOR COPYFIND RESULTS

Similarity Index	New Mapping
84-100	4
68-83.99	3
52-67.99	2
36-51.99	1
0-35.99	0

A band size of 16 resulted in the smallest average deviation from the manual results, which was 0.4867. This slightly larger band, 16 compared to 14 used for the GST approach, accounts for the wider distribution of similarity indices produced by the Copyfind algorithm, as shown in the three plots above. Figure (7) shows how the Copyfind output is categorized according to these bands:

FIGURE (7) – GST SIMILARITY INDEX PRODUCT VS. DUPLICATION PROJECTION WITH BAND BREAKDOWN



Data Compression (Entropy) Approach

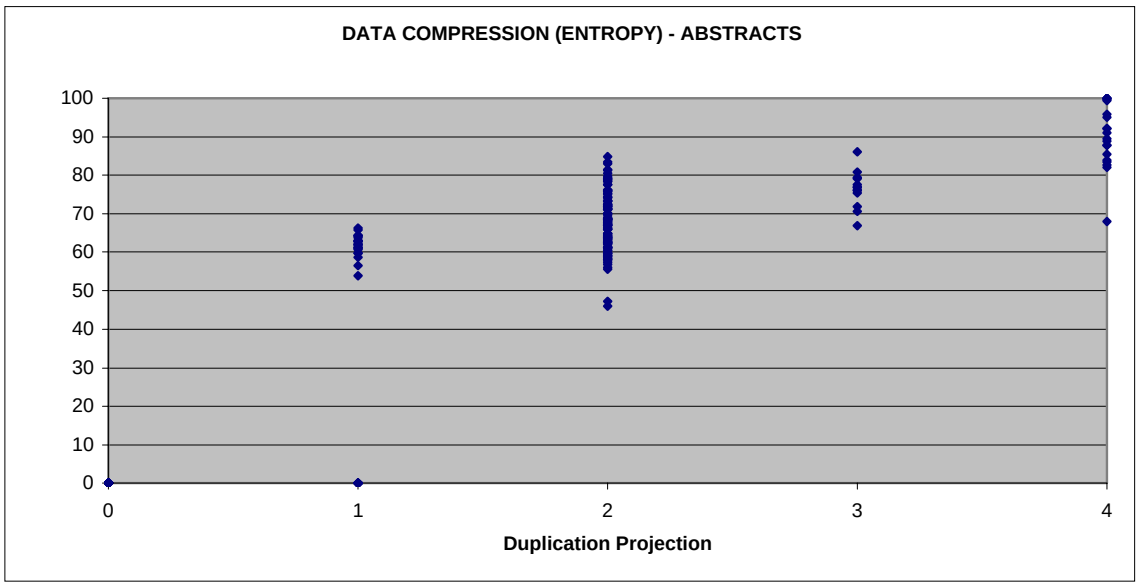
The entropic algorithm based on data compression was applied to the Abstract field of the full Fractals database. The non-integer variant of the output, which was originally generated in a form compliant with the manually produced scores (integer between 0 and 4), was used to produce a plot analogous to those used for the previous approaches (FIG. (8)). The formula used to generate this non-integer index is given in Appendix (4).

With the exception of the 2 scored records, very little overlap was observed using this approach. The 2 scored records overlap the 1 and 3 scored records indicating some degree of uncertainty between 1 and 2 scored records as well as between 2 and 3 scored records. However, the 4 scored records are concentrated almost exclusively higher than all other records. Likewise, the 3 scored records are concentrated above all 1 and 0 scored records, and 1 scored records are concentrated above most 0 scored records. Though it is difficult to see in the plot, many 0 scored records were observed, but were all given 0 similarity indices causing the data points to overlap in the bottom left corner of the plot.

The above observations 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 2 scored records, which exhibit some degree of ambiguity, the overall

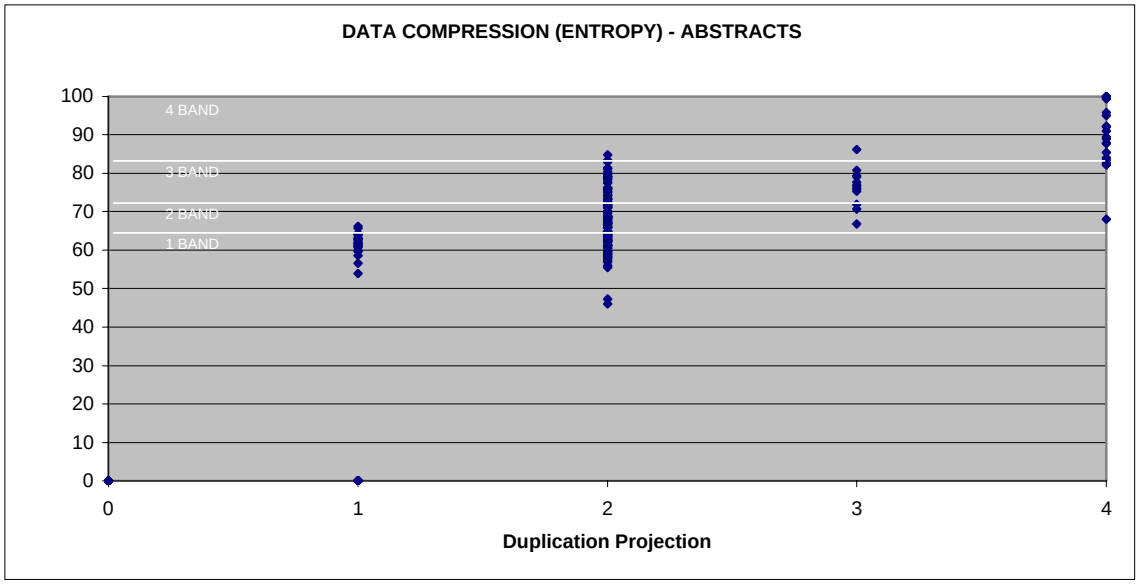
correlation of the algorithmically produced similarity indices to the manual duplication projections is quite high.

FIGURE (8) – ABSTRACT-BASED DATA COMPRESSION (ENTROPY) SIMILARITY INDEX VS. DUPLICATION PROJECTION



Because the original output of the data compression algorithm was on the same scale as the manually produced scores, there was no need for a remapping scheme. The suitability of the algorithm was measured in the same manner as with the other approaches, averaging the absolute deviations from the manual scores. This average deviation was calculated to be 0.1689 – a relatively low number, which reinforces our observations from the above plot. A plot of this data including the effective bands (i.e. the values generated by the algorithm prior to mapping to the uniform similarity index) in shown in Figure (9) below:

FIGURE (9) – ABSTRACT-BASED DATA COMPRESSION (ENTROPY) SIMILARITY INDEX VS. DUPLICATION PROJECTION WITH EFFECTIVE BAND BREAKDOWN



Overall Algorithm Comparisons for Abstracts

A summary of the suitability metrics associated with each algorithm is given in Table (5):

TABLE (5) – OVERALL SUITABILITY METRICS

Approach	Band Size	Avg. Diff.
GST	14	0.2756
Copyfind	16	0.4867
Entropy	N/A	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 (6)). The rankings used in the full text analysis were also integers ranging from 0 to 4.

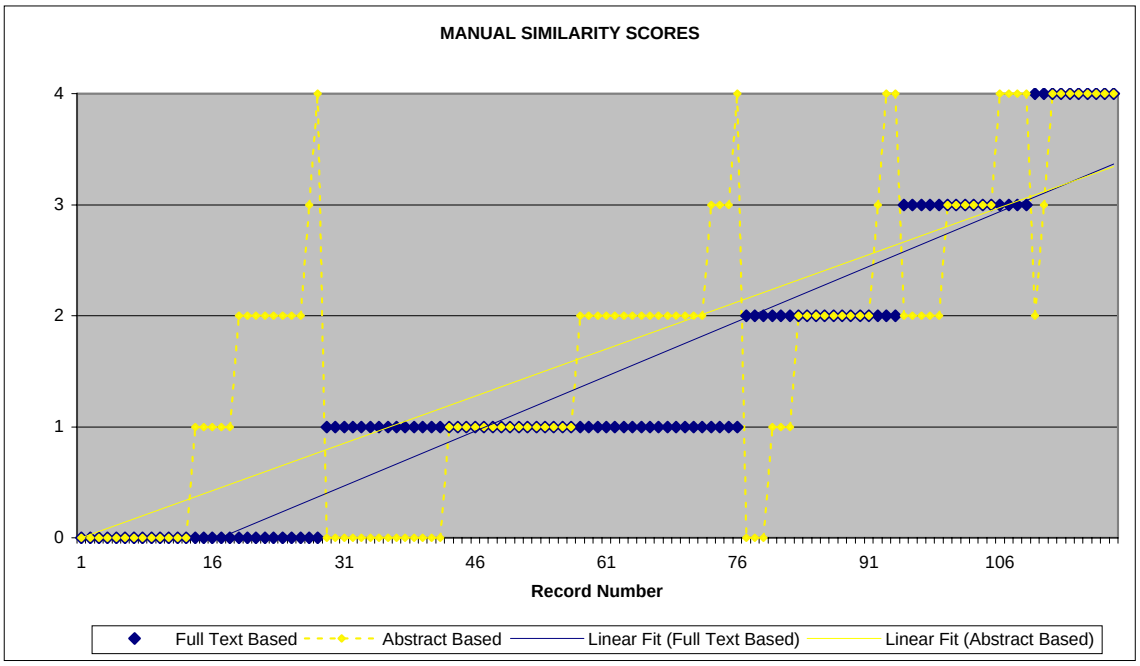
Results from this analysis showed a moderate correlation between Abstract scores and full text scores, which can be seen in Figure (10). The plot uses the record number as the x-axis metric (an arbitrary index in this case) and the manual score as the y-axis. The data are sorted first according to full 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 more significantly changed 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. These records are shown in the plot as yellow points outlined in blue (signifying the overlap of scores).

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 (10) – 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 2 scored records, 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 publications 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.

5. REFERENCES

Atlam ES, Fuketa M, Morita K, Aoe J. Documents Similarity Measurement Using Field Association Terms. Information Processing & Management 39 (6): 809-824 Nov 2003

Bailey BJ. Duplicate Publication in the Field of Otolaryngology-Head and Neck Surgery. Otolaryngology-Head and Neck Surgery 126 (3): 211-216 Mar 2002

Bansal N, Blum A, Chawla S. Correlation Clustering. Machine Learning 56 (1-3): 89-113 Jul-Sep 2004

Bao JP, Shen JY, Liu XD, Liu HY, Zhang XD. Finding Plagiarism Based on Common Semantic Sequence Model. Advances in Web-Age Information Management: Proceedings Lecture Notes in Computer Science 3129: 640-645 2004

Benedetto D, Caglioti E, Loreto V. Language trees and zipping. Physical Review Letters 88 (4): Jan 28 2002.

Braumoeller BF, Gaines BJ. Actions Do Speak Louder Than Words: Deterring Plagiarism with the Use of Plagiarism-Detection Software. PS-Political Science & Politics 34 (4): 835-839 Dec 2001

Chang Y, Kim M, Ounis I. 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 2004

Chen X, Francia B, Li M, Mckinnon B, Seker A. Shared Information and Program Plagiarism Detection. IEEE Transactions on Information Theory 50 (7): 1545-1551 Jul 2004

Cook DE, Mellor L, Frost G, Creutzburg R. Knowledge Management and the Control of Duplication. Engineering and Deployment of Cooperative Information Systems, Proceedings Lecture Notes in Computer Science 2480: 396-402 2002

Dobrynin V, Patterson D, Rooney N. Contextual Document Clustering. Advances in Information Retrieval, Proceedings Lecture Notes in Computer Science 2997: 167-180 2004

Doherty M. Misconduct of Redundant Publication. Annals of the Rheumatic Diseases 55 (11): 783-785 Nov 1996

Gilbert FJ, Denison AR. Research Misconduct. Clinical Radiology 58 (7): 499-504 Jul 2003

Gwilym SE, Swan MC, Giele H. 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 Jul 2004

Hoad TC, Zobel J. Methods for Identifying Versioned and Plagiarized Documents. Journal of the American Society for Information Science and Technology 54 (3): 203-215 Feb 1 2003

Hui SC, Fong ACM. Document Retrieval from a Citation Database Using Conceptual Clustering and Co-Word Analysis. Online Information Review 28 (1): 22-32 2004

Kostoff RN, Eberhart HJ., and Toothman DR. Database Tomography for information retrieval. Journal of Information Science, 23:4. 1997.

Kostoff RN. The underpublishing of science and technology results. The Scientist. 14:9. 6-6. 1 May 2000.

Kostoff, RN, Shlesinger M, and Malpohl G. Fractals roadmaps using bibliometrics and database tomography. Fractals. 12:1. 1-16. March 2004.

Jefferson T. Redundant Publication in Biomedical Sciences: Scientific Misconduct or Necessity? Science and Engineering Ethics 4 (2): 135-140 Apr 1998

Leuski A, Allan J. Interactive Information Retrieval Using Clustering and Spatial Proximity. User Modeling and User-Adapted Interaction 14 (2-3): 259-288 Jun 2004

Li WY, Ng WK, Lim EP. 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 2004

Maderlechner G, Suda P, Bruckner T. Classification of Documents by Form and Content. Pattern Recognition Letters 18 (11-13): 1225-1231 Nov 1997

Mojon-Azzi SM, Jiang XY, Wagner U, Mojon DS. Redundant Publications in Scientific Ophthalmologic Journals - the Tip of the Iceberg?. Ophthalmology 111 (5): 863-866 May 2004

Monostori K, Finkel R, Zaslavsky A, Hodasz G, Pataki M. Comparison of Overlap Detection Techniques. Computational Science-ICCS 2002, Pt I, Proceedings Lecture Notes In Computer Science 2329: 51-60 2002

Muresan G, Harper DJ. Topic Modeling for Mediated Access to Very Large Document Collections. Journal of the American Society for Information Science and Technology 55 (10): 892-910 Aug 2004

Pecorari D. Good and Original: Plagiarism and Patchwriting in Academic Second-Language Writing. Journal of Second Language Writing 12 (4): 317-345 Dec 2003

Prechelt L, Malpohl G, Philippsen M. Finding plagiarisms among a set of programs with JPlag. Journal of Universal Computer Science. 2002. 8(11). 1016-1038.

Rasmussen E. Clustering Algorithms. In W. B. Frakes and R. Baeza-Yates (eds.). Information Retrieval Data Structures and Algorithms. 1992. Prentice Hall, N. J.

Salton G, Buckley C. Text Matching for Information-Retrieval. Science 253 (5023): 1012-1015 Aug 30 1991

Schein M, Paladugu R. Redundant Surgical Publications: Tip of the Iceberg? Surgery 129 (6): 655-661 Jun 2001

Shin K, Han SY, Gelbukh A. Advanced Clustering Technique for Medical Data Using Semantic Information. Micai 2004: Advances in Artificial Intelligence Lecture Notes in Computer Science 2972: 322-331 2004

Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Technical Report #00--034. 2000. Department of Computer Science and Engineering. University of Minnesota.

Von Elm E, Poglia G, Walder B, Tramer MR. Different Patterns of Duplicate Publication - An Analysis of Articles Used in Systematic Reviews. Jama-Journal of the American Medical Association 291 (8): 974-980 Feb 25 2004

Wise MJ. String similarity via greedy string tiling and running Karb-Rabin matching. ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps, 1992. Dept. of CS, University of Sidney.

6. 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) – GREEDY STRING TILING (GST)

Greedy String Tiling clustering is a method of grouping text or text character documents (files) by similarity. All documents to be grouped are placed in a database. Each pair of documents is compared by GST, an algorithm originally used to detect plagiarism [Wise, 1992; Prechelt et al, 2002], and a similarity score is assigned to the pair. Then, hierarchical aggregation clustering [Rasmussen, 1992; Steinbach, 2000] is performed on all the documents, using the similarity score for group assignment.

Greedy String Tiling computes the similarity of a pair of documents in two phases. First, all documents to be compared are parsed and converted into token strings (words or characters). Second, these token strings are compared in pairs for determining the similarity of each pair. During each comparison, the GST algorithm attempts to cover one token string (document) with sub-strings ('tiles') taken from the other string. These sub-

strings are not allowed to overlap, resulting in a one to one mapping of tokens. The attribute greedy stems from the fact that the algorithm matches the longest sub-strings first.

A number of similarity metrics can be defined once the tiling is completed. One similarity metric is the percentage of both token strings that is covered. Another similarity metric is the absolute number of shared tokens. A third similarity metric is the mutual information index. Depending on the purpose of the matching, additional weightings can be used for the similarity matrix to increase the ranking precision. For example, if plagiarism is one study objective, additional weighting could be given to shared string length. All similarity metrics have positive and negative features, and the choice of metric is somewhat influenced by the study objectives and the structure of the database.

Once the document similarity matrix has been generated, myriad clustering techniques can be used to produce a classification scheme (taxonomy). In the present study, multi-link hierarchical aggregation was used. Three clustering variants were actually generated, although the extension to other clustering schemes is straight-forward. Single-link, average-link, and complete-link variants are implemented. The variants differ in how the decision of merging to clusters is made. Single-link requires that the similarity of at least two documents is higher than a certain threshold, while complete-link requires that the similarity between all documents in both clusters be higher than a threshold. Average-link requires that the average pair-wise similarity between the documents of both clusters exceed the threshold. For the present study, complete-link appeared to give good results, and was the clustering method used.

APPENDIX (3) – COPYFIND ALGORITHM

A variant of the freely available "WCopyfind" software (v. 2.2) was run on the full collection of Abstracts. Parameters were assigned as follows. The minimal phrase match, that is the minimum string length considered to be a match, was set to 6. Additionally, a tolerance for imperfections was used to account for typos, misspellings, and very minor modifications. This tolerance is assigned in numerical form as (A) the maximum number of non-matching words that Copyfind will allow between perfectly matching portions of a phrase and (B) the minimum percentage of perfectly matching words required for a pair of phrases to be considered matching. For the

present study, a moderate value of both parameters was used: (A) was set at 2 words and (B) at 80%. The similarity index obtained was simply the percentage of matching words, as contained in matching phrases, divided by the average number of total words in the two documents being compared. A similarity index of 0 indicates no matching words while an index of 100 indicates that the two documents are identical.

APPENDIX (4) – DATA COMPRESSION CLUSTERING

The complete set of Abstracts was analyzed performing all binary possibilities of repetition or close Abstracts. The entropic algorithm based on zipping (compressing) files was applied using the following formula:

$$E(ab) = (\text{length}(\text{zip}(a+b)) - \text{length}(\text{zip}(a)) - \text{length}(\text{zip}(b+b)) + \text{length}(\text{zip}(b))) / \text{length}(b).$$

Where a and b are the Abstracts to be analyzed and zip indicates the zipped Abstract. First in order to have a complete automated algorithm, two other files from previous studies were analyzed, one composed for a previous set of Abstracts for the Fractals database (incomplete) and a second one of the Abstracts with a Mexican author published in 2001. With these two sets of Abstracts the entropy distributions were plotted (FIG. (11)). It was noted that the two distributions were very similar and could be adjusted by a Gaussian distribution in an interval of 12 times standard deviation around their mean value. The two distributions had approximately the same mean and standard deviation values. For values that were six standard deviations (6σ) less than the mean value a dramatic change in both distributions was observed. This value was determined to be indicative of possible repetitions and was the same for both previous analyses. Following a manual classification of Abstract pairs with entropy less than [mean - 6σ] in these two previous analyses the following classification was developed:

LEVEL 1 - IF $E(ab) < (\text{mean}-9\sigma)$ AND $E(ba) < (\text{mean}-9\sigma)$

LEVEL 2 - IF $E(ab) < (\text{mean}-9\sigma)$ AND $[(\text{mean}-9\sigma) < E(ba) < (\text{mean}-6\sigma)]$ or vice versa

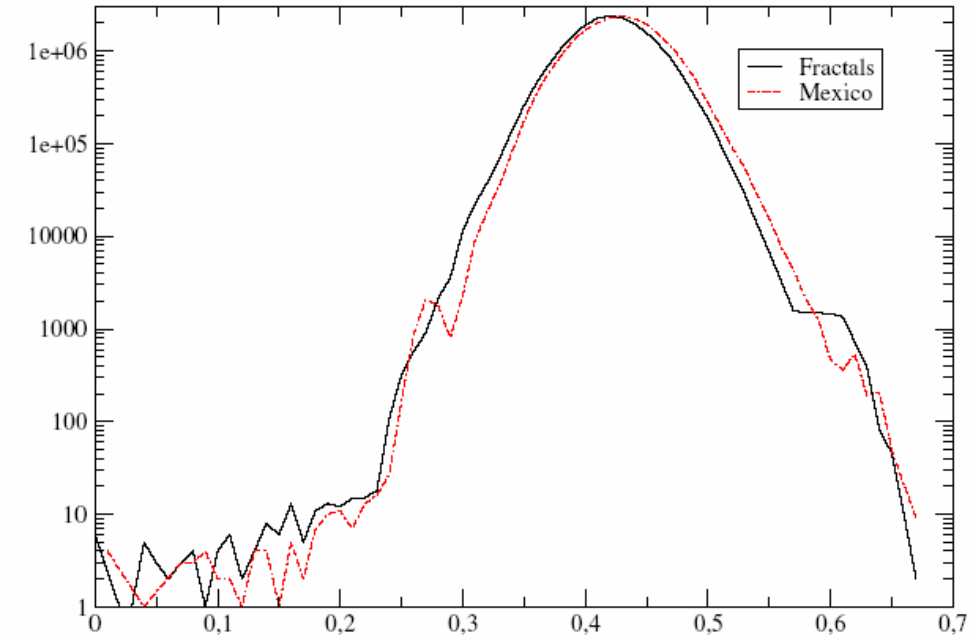
LEVEL 3 - IF $[(\text{mean}-9\sigma) < E(ab) < (\text{mean}-6\sigma)]$ AND $[(\text{mean}-9\sigma) < E(ba) < (\text{mean}-6\sigma)]$

LEVEL 4 - IF [E(ab) < (mean-6σ) AND E(ba) > (mean-6σ)] or vice versa

Additionally, a non-integer variation on these levels was developed for comparison with the other three computational techniques. It is given by the following formula and ranges from 0 (no similarity) to 100 (exact match):

EProd(ab) = (1-RE(ab))*(1-RE(BA))

FIGURE (11) – ENTROPY DISTRIBUTIONS FOR TWO ABSTRACT SETS



APPENDIX (5) – 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 (6) – 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.