Text similarity in academic conference papers

Jun-Peng Bao
Xi'an Jiaotong University, People's Republic of China

James A. Malcolm
University of Hertfordshire, UK

Abstract

If we are to use electronic plagiarism detectors on student work, it would be interesting to know how much similarity should be expected in independently written documents on a similar topic. If our measure is coarse, the answer should be zero, but a finer grained analysis (such as would be needed to detect inadequate paraphrasing) is likely to detect some background noise. How much background noise should there be?

We would like to determine this, but it is hard to publish research based on analysis of student work, because we cannot know whether any particular pair of students worked completely independently or not, and in any case the results might attract unwelcome publicity.

To get an estimate of an appropriate level of this background noise, we analysed submissions to an international conference using the Ferret plagiarism detector developed by Lyon *et al.* (2001). Ferret provides very fast and fine-grained similarity detection in moderately large collections of documents.

This was an exercise intra-corporal or collusion detection rather than comparison to Web sources. For this purpose the Ferret algorithm is well suited. There were 483 files; scanning the files took about 50 seconds, and calculating the similarity statistics took about 10 seconds.

There were 116403 file pairs. Of these pairs, only 116 (0.1%) had more than 99 common triples, and of these only 19 pairs (0.016% of the total) had over 200 matching triples (200 is about 10% of the typical size of the smaller files). There should be NO plagiarism here, as these are published conference papers, but in fact the top few are all pairs of papers with common authors, and they have re-used text. A simple MS Word file compare between one of the top ranking pairs is sufficient to make the similarities obvious (though Word does not highlight all the similarities by any means). Nevertheless, as expected, most document pairs showed very low similarity measures, and this was consistent across the vast majority of pairs.

As noted, there was a surprisingly large degree of similarity in just a few cases. We accordingly investigated these pairs more carefully. The worst case was of an author who had submitted two papers. Each paper reported the results of a single experiment, but the background material for both experiments was very much the same and he had simply reproduced the same text in both papers.

We present also the other cases where similarity was high, and ponder the implications for routine scanning of student work.

Corresponding author: James.A.Malcolm, Division of Computer Science, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB.
Email: j.a.malcolm@herts.ac.uk

Introduction

If we are to use electronic plagiarism detectors on student work, it would be interesting to know how much similarity should be expected in independently written documents on a similar topic. If our measure is coarse, the answer should be zero, but a finer grained analysis (such as would be needed to detect inadequate paraphrasing) is likely to detect some background noise. How much background noise should there be?

We would like to determine this, but it is hard to publish research based on analysis of student work. There would be two problems with such an approach: firstly, we cannot know whether any particular pair of students worked completely independently or not, and secondly the results might attract unwelcome publicity.

To get an estimate of an appropriate level of this background noise, we analysed submissions to an international conference using the Ferret plagiarism detector developed by Lyon et al. (2001). Ferret provides very fast and fine-grained similarity detection in moderately large collections of documents. As part of a previous study (Bao *et al*., 2004), we had available a collection of papers which had been submitted to an international conference, and these had already been converted from PDF to plain text format.

Research Questions

We wanted to address the following research question: how similar are independently written documents on similar topics? But as the study progressed, a number of subsidiary questions were raised:

- What causes high similarity between two documents (other than one copying from the other, or both copying, possibly indirectly, from a common source)?

- The documents we used for our study are all on the Web, and in many cases represent the results of fairly small-scale research projects such as might be attempted by an MSc student. Does TurnitinUK find all these documents? Does Google find them all?

- How much plagiarism is there in the set of papers we investigated?

- When two papers have a common author, how much of the similarity that we find is because a common author implies a common subject, how much is because a common author implies a common style of writing and turn of phrase, and how much because text has been copied from one paper to the other. Of course the common author may not have had anything to do with the actual writing, particularly perhaps in the case of a supervisor of a PhD student.

- Do authors always self-plagiarise, or not always?

We do not yet have answers to all of these questions: this is a work in progress. As an aside, we are also able from this data set to look at patterns of co-authorship and multiple submissions to the same conference.

Pilot Study

This was a study in intra-corpal plagiarism (Culwin and Lancaster, 2001) or collusion, rather than comparison to Web sources. For this purpose the Ferret algorithm is well suited.

Although Ferret is fast, the volume of data to be examined is large, and we were going to have to look at quite a number of pairs of documents to see why the automated tool suggested that there was a high degree of similarity in any particular pair of documents. So, as a pilot study, and to determine the likely scale of our more detailed investigation, we looked at just the first 100 papers (4950 pairs).

Ferret works by building an index of where, in the collection of documents, each triple may be found, and so two similarity metrics naturally suggest themselves. One is a simple count of the number of common triples -- this is good for spotting relatively small chunks of copying in large documents. The alternative is the Jaccard or Tanimoto coefficient, which is calculated as the number of common triples in this pair of documents, divided by the total number of different triples in the pair (Manning and Schutze, 1999). This is the resemblance measure which we refer to as R and which we shall use in the rest of the paper. R can be expressed either as a fraction or as a percentage.

Figure 1 shows the resemblance, R, for the 4950 pairs of papers.

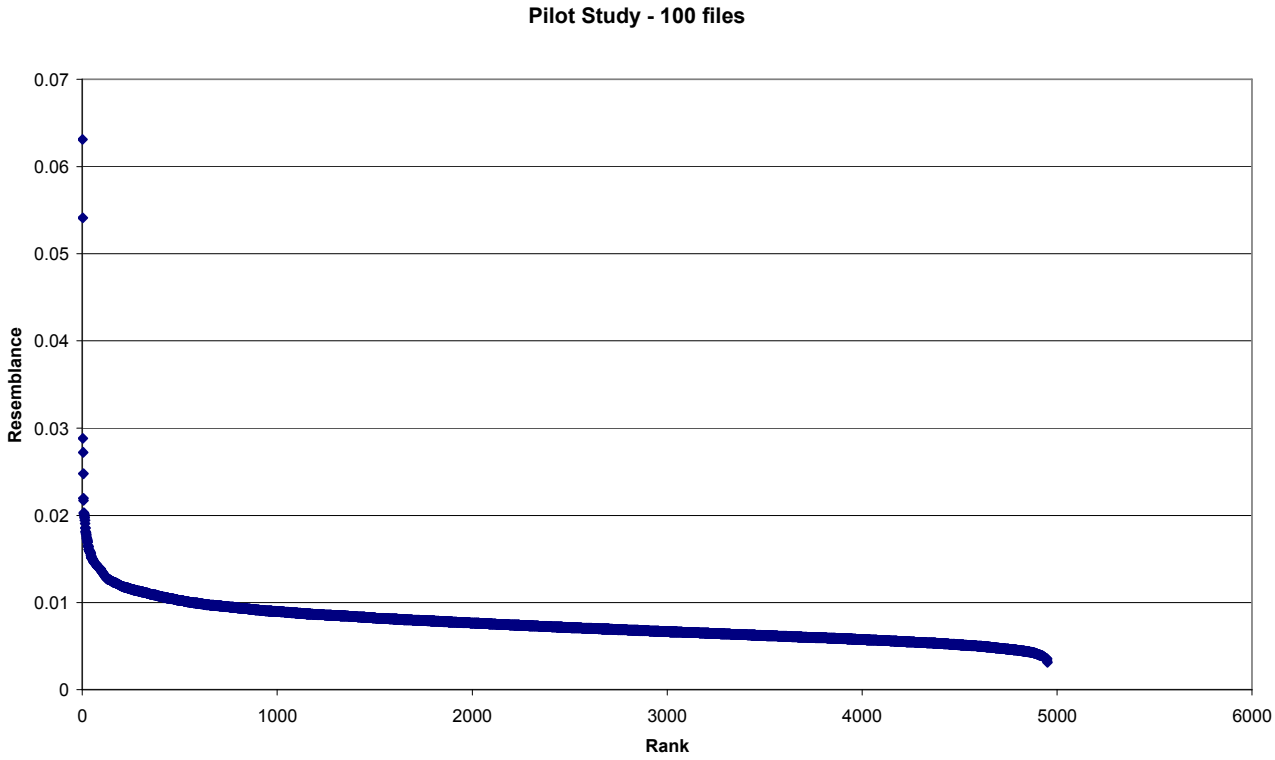

Figure 1: All 4950 pairs from the pilot study in rank order of resemblance.

The results show that only 555 pairs had more than 1% resemblance. From the pair with rank 556 (R = 0.01) onwards the resemblance drops slowly, with a slightly more rapid fall off on

the last 50 or so pairs. There is no obvious cause for the lower resemblance on the least similar pairs.

In this subset of the data, the most similar pair had a similarity of R = 0.063. The most similar pairs after that showed rapidly falling values for the resemblance metric R: 0.054, 0.029, 0.027, 0.025, 0.022 ...

This can be seen more clearly if we look at only the 29 most similar pairs, as shown in Figure 2.

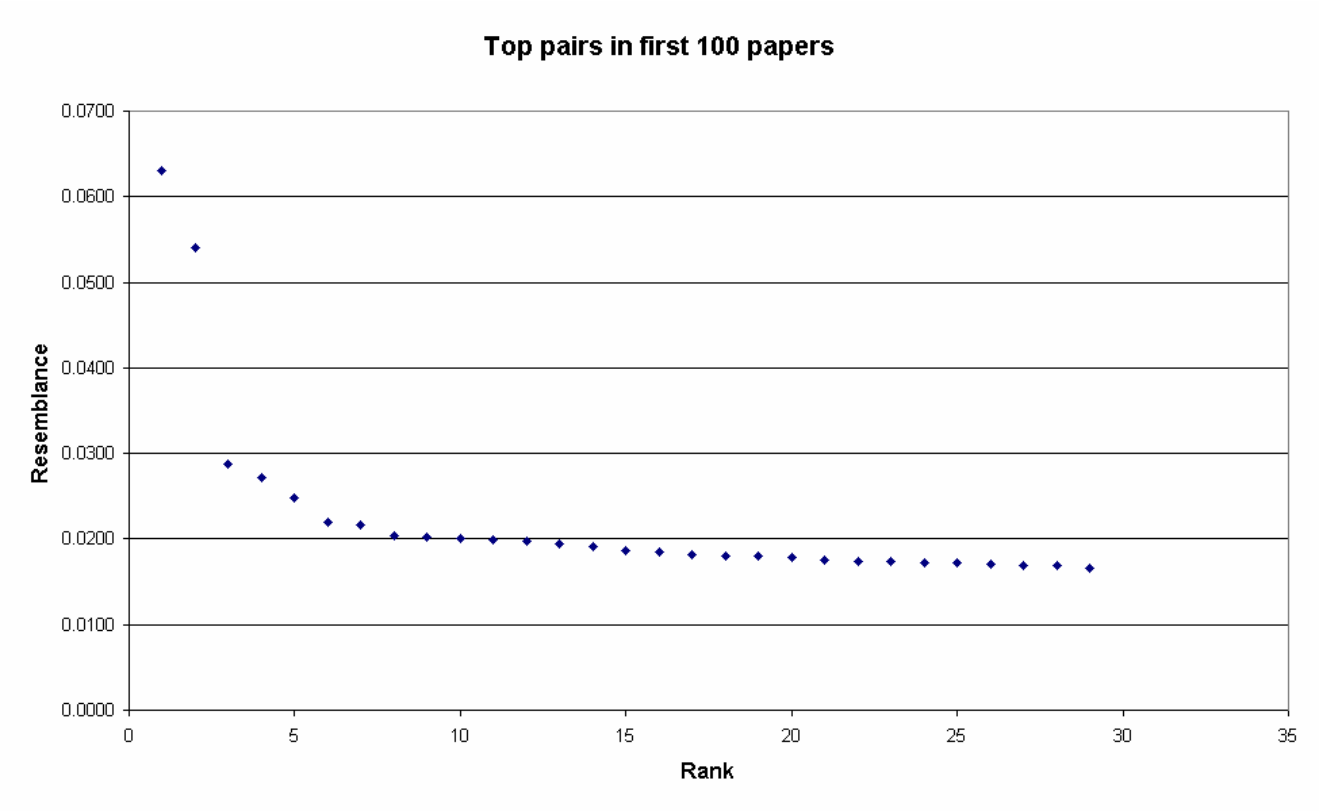

Figure 2: The 29 most similar pairs from the first 100 papers.

The first two pairs show vastly greater similarity than the remainder, and although the third pair (in rank order of resemblance value) is much lower, there does seem to be a knee in the curve after the seventh pair, around the slightly lower value of R = 0.02. This can be seen better if we plot just the top pairs; in Figure 1 the leftmost part of the curve seems quite smooth.

We looked at these top 10 pairs individually to see whether there was any difference between those pairs with higher R values, and those with lower resemblance.

The most similar pair (documents 23 and 24, R = 0.063) had 2 common authors (out of 3 in each case), and the two papers were found to contain whole matching or substantially matching paragraphs. To be fair, the authors of the papers were using similar methods to solve two different problems, but this does seem to be a case of ‘self-plagiarism’.

On the other had, the next pair (documents 96 and 97, R = 0.054) had all 3 authors in common, and a similar subject, but the matching triples were scattered throughout the documents indicating that plagiarism probably had not taken place.

There is not a big difference in R value or number of common triples between these two pairs, so we are reminded that human judgement is always necessary to decide whether plagiarism may have taken place: the computer just helps us to know which cases should be examined more closely 'by hand'. Nevertheless we are developing metrics that will take into account the relative ordering of triples in the two documents in order to better identify documents that need further examination. But any such metric needs to be simpler than finding the longest common subsequence, as the run time for that algorithm is proportional to the product of the document lengths.

The remaining pairs have distinctly lower R values.

The next 4 pairs also have one or more authors in common; documents 2 and 52 (R = 0.029) have a similar subject, but were judged not to have plagiarised, whereas documents 21 and 26 not only deal with the same subject as each other, but also have some very similar text. This is a borderline case. The next two pairs (documents 48 and 49, R = 0.025, and documents 6 and 28, R = 0.022) have little similarity.

The next pair (documents 76 and 89, R = 0.022) has little obvious similarity to a human reader, but the final three pairs of our top ten (documents 14 and 67, documents 13 and 72, and documents 6 and 83, all with R = 0.020) do at least deal with similar topics.

This shows us that while common authors and common topics can make similarity seem more likely, it is also possible for this level of similarity to happen apparently by chance. Some of the reasons for this apparently chance similarity are discussed later.

It is also worth reminding ourselves that there are another 4940 pairs of documents with lower R values that we can assume have little or no evidence of plagiarism and do not even have to look at.

The full set of documents

Then we went on to examine the full set of 483 documents. Scanning these 483 files took about 50 seconds, and calculating the similarity statistics took about 10 seconds. These timings differ from those presented by Lane et al. (2006) because this work was done on an older machine using a different version of Ferret. Another Ferret implementation took 24 seconds to scan all the data, and just 4 seconds to calculate the resemblance measures for each pair.

Prechelt et al. (2002) point out that when looking for collusion within a group we need to consider the number of pairs in the dataset (in this case 116403 pairs). Of this large number, only 210 pairs were above the 2% resemblance level. Of the 210 pairs above 2% similarity, many had common authors.

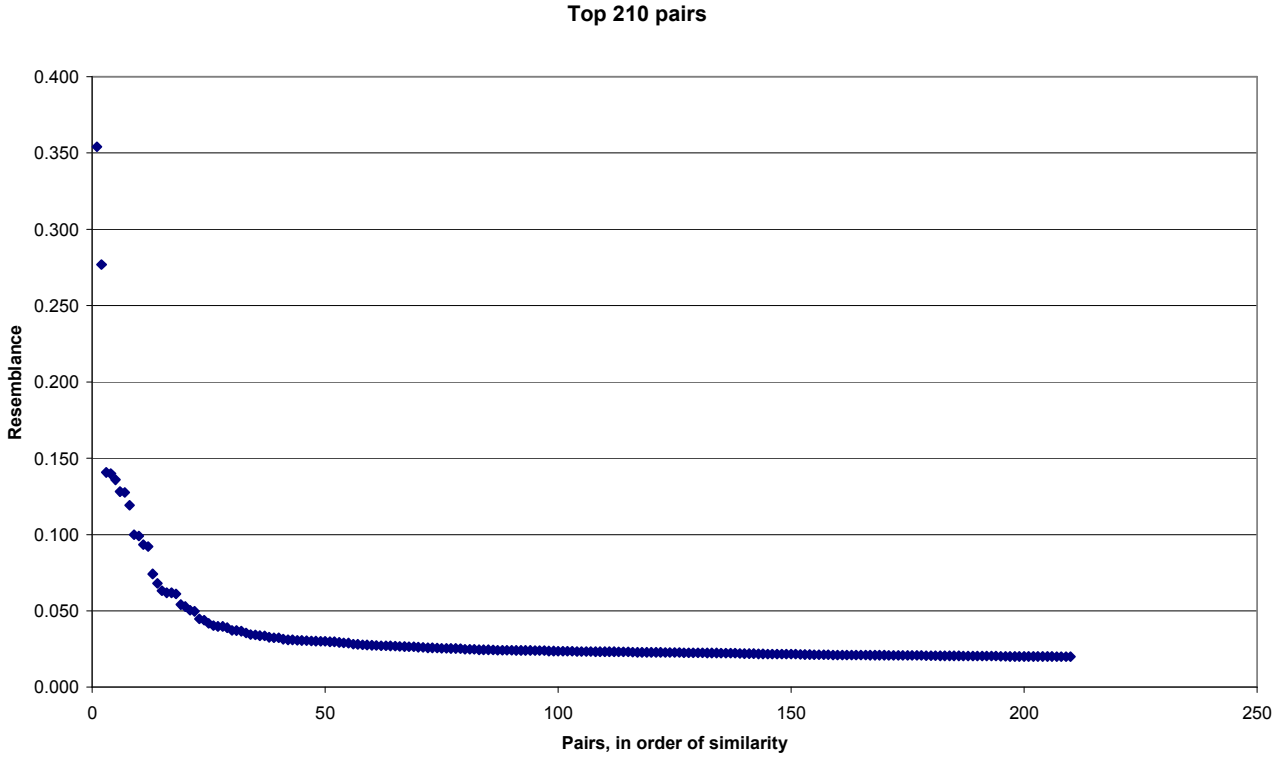

Figure 3: The 210 most similar pairs from the full 483 papers (all those of 2% similarity or more).

We choose to present here in Figure 3 only the 210 pairs with R values of 0.02 or greater; the graph of the other 116,193 pairs is not very interesting.

The most important thing to notice here is that the scale on the Y-axis has changed. The most similar pair in the full set of data has R = 0.354: more than 5 times greater than the maximum R found in the pilot study. This pair of papers has the same four authors and a very similar topic: related approaches to solving the same problem. Part of the reason for the very high similarity is that much of both papers is taken up with background material, so it is hardly surprising to find large chunks of matching text in the two papers. The authors should have noted the advice given to University of Hertfordshire students about writing a final year project report: only put in background material that is relevant to your project, and if you do put in background material show how it is relevant by giving examples of how you applied the ideas to your project.

Pairs with high similarity

In the top 30 pairs of papers in ranked order of R value (from 35% down to below 4%) all but three pairs have one or more authors in common. Two pairs, 13 and 16, do not have common authors, but in each case the authors of the two papers do come from the same institute. In fact there are several other papers similar to these, and lots of self-plagiarism!

The exception in the top 30 is the pair of papers ranked 21st for similarity, which had R = 0.05 but did not have a common author or institute. They did have very similar topics, but the

Ferret display shows that similarity is caused by use of similar terminology and volumes of mathematics with similar notation rather than any evidence of copying.

How many pairs with common authors were below the 2% level? There is some difficulty (or rather tedium) in cleaning the data, but by a combination of extracting the authors' names by program, and manual data cleaning, we determined that there are approximately 389 pairs of papers sharing one or more common author, so less than half of paper pairs with one or more joint authors showed evidence of self-plagiarism. Further work is needed to complete and check this analysis.

What was the top pair in our pilot study of the first 100 documents is now ranked only 15th. But number of pairs has increased from 4950 to 116403 pairs (a factor of 23) so this drop in rank is not as surprising as might seem at first sight.

Causes of Document Similarity

As well as looking at the most similar pairs, we also wanted to find out what features made a document similar to other documents in the set without actually showing evidence of plagiarism on closer examination. We looked to see which documents were most often found in the top 210 pairs. The document which was similar to the greatest number of other documents was number 110 which showed similarity to 11 other documents at resemblance over 0.02. It was most similar to document 253 with an R value of 0.03. These two documents had a very similar topic (time delay control with varying parameters), but there was also a substantial amount of similarity caused by a number of common factors which appeared in many document pairs.

Firstly, there are always a few innocuous common phrases: ‘in this paper’, ‘the simulation results show’, etc. This is what we are trying to measure.

Secondly, the conference title and copyright notice, which appears on every page of every document, together with a small set of other triples which are artefacts of the standard formatting, contribute 15 common triples to every pair. This is a systematic source of error.

Thirdly, there were many common triples caused by diagrams and mathematics not properly treated by the conversion utility. This could be considered as noise, as it will be different in every pair; sometimes more, sometimes less, but possibly significant with maybe 20 or more common triples from this source in a typical pair.

Similar explanations apply for document 78 (R = 0.026 when compared to document 110) but here there were also a clutch of common triples caused by the text describing the Fuzzy subsets used. This text, or something similar, is quite common in Fuzzy logic papers as it seems to be a routine part of the technique, and is described in Fuzzy logic tutorials.

Other documents similar to 110 (and the R values calculated) were 240 (R = .020), 189 (R = .021), 390 (R = .021), 82 (R = .021), 199 (R = .021), 209 (R = .022), 89 (R = .023), 193 (R = .023), 360 (R = .024) as well as 78 and 253 mentioned above (11 documents in all).

Useful further work would be to separate out this set of 12 documents and carefully remove all sources of error (such as common text and diagrams) in order to find the true level of common triples in independently written text.

Dissimilar Papers

In a couple of sample pairs of dissimilar papers (at a middle rank of about 50,000 out of the 116,403 total pairs) there were around 30 common triples.

Those not caused by the sources of error identified above were in both cases very general; in the first file these 13:

according to the
and future work
and so on
based on the
belongs to the
in the table
indicate that the
is defined as
of the whole
should be considered
shows that the
the number of
to compute the

Another pair gives the following 16 common triples (apart from 'boilerplate'):

as a result
implemented in our
in figure a
in order to
in recent years
in the same
in this paper
is to find
it is more
paper we present
the process of
the system to
this paper we
used to find
we can get
widely used in

Note that none of these triples reveal the subject of the paper in any way: the remaining common triples (after removal of triples from the conference title which are common to all documents) were often very neutral.

The paper from the keynote speaker did not have the same format and headings as the others, so we ignore that exception in stating that the minimum resemblance of any pair was 0.0027 (about one quarter of a percent). For pairs with this low R value, the only similarities apart from the conference title were just 2 or 3 common triples. More generally about 0.5% similarity can be explained by the various sources of error identified above.

Application

Care should be taken over any application of the results of this study to the similarity figures produced by TurnitinUK.

These experiments were performed with Ferret, which provides very fine-grained detection of similarities, whereas TurnitinUK only identifies relatively long sequences of identical words. To avoid detection by TurnitinUK, it is believed to be necessary to change only about every fifth word, whereas with Ferret every third word would have to be changed. So a figure of 2% similarity in TurnitinUK is much more serious than 2% similarity in Ferret. An advantage of the (unpublished) TurnitinUK algorithm is that it is very unlikely to pick up accidental similarity. The relative ease of circumvention exactly mirrors the very low false positive rate. Any similarity it detects is very likely to be caused by copying from a common source (though possibly not the source it identifies -- one cannot tell which bits are the original, and which the copy (Christianson and Malcolm, 1997). So even 2% is too high.

Any figure over 2% very likely does represent a problem. Even with a fine-grained detector there are very few cases where over 2% similarity seems to have occurred by chance.

Whatever threshold we put; human intervention is essential. Note, you need to go back to the original source to check the format of the submission. Students occasionally cannot follow instructions, so may have used italics (lost in the processing by both TurnitinUK and by Ferret) to indicate quotation even if that is not the recommended technique for the submission in question.

Summary

With the help of our software, we examined 116403 document pairs. Of these pairs, only 116 (0.1%) had more than 99 common triples, and of these only 19 pairs (0.016% of the total) had over 200 matching triples (200 is about 10% of the typical size of the smaller documents). There was a surprisingly large degree of similarity in just a few cases. We accordingly investigated these pairs more carefully. Although there should be no plagiarism here, as these are published conference papers, in fact the most similar pairs of papers had common authors who had re-used text. A simple Microsoft Word file compare between one of the top ranking pairs is sufficient to make the similarities obvious (though Word does not highlight all the similarities by any means). Nevertheless, as expected, most document pairs showed very low similarity measures, and this was consistent across the vast majority of pairs.

Notes on contributors

James A. Malcolm is a member of academic staff in the School of Computer Science at the University of Hertfordshire where the Plagiarism Detection Research Group conducts research into both pedagogic and technical issues to do with plagiarism.

Jun-Peng Bao is in the Department of Computer Science and Engineering of Xi'an Jiaotong University, in the People's Republic of China. Document copy detection is one application of his research work on text feature extraction.

References

Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu, and Xiao-Di Zhang (2004).
Finding Plagiarism Based on Common Semantic Sequence Model.
Advances in Web-Age Information Management. LNCS Volume 3129, pp. 640 - 645

Bruce Christianson and James A. Malcolm (1997).
Binding Bit Patterns to Real World Entities.
Security Protocols. LNCS Volume 1361, pp. 105 - 113

Fintan Culwin and Thomas Lancaster (2001).
Visualising Intra-Corpal Plagiarism.
In *Proceedings of the International Conference on Information Visualisation*.

Peter C. R. Lane, Caroline M. Lyon & James A. Malcolm (2006).
Demonstration of the Ferret Plagiarism Detector.
In *Proceedings of the 2nd International Plagiarism Conference*.

C. M. Lyon, J. A. Malcolm, and R. G. Dickerson (2001).
Detecting short passages of similar text in large document collections.
In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.
SIGDAT Special Interest Group of the ACL.

C. D. Manning and H. Schutze (1999).
Foundations of Statistical Natural Language Processing.
Cambridge, MA: MIT Press.

Lutz Prechelt, Guido Malpohl, and Michael Philippsen (2002).
Finding Plagiarisms among a Set of Programs with JPlag.
Journal of Universal Computer Science 8(11), pp. 1016 - 1038

Authors' contact details:

Corresponding author: James A. Malcolm,
School of Computer Science,
University of Hertfordshire,
College Lane, Hatfield, Herts, AL10 9AB
UK

EMail: j.a.malcolm@herts.ac.uk

Co-author: Jun-Peng Bao,
Department of Computer Science and Engineering,
Xi'an Jiaotong University,
Xi'an 710049
People's Republic of China

EMail: baojp@mail.xjtu.edu.cn