IMA: Identification of Multi-author Student Assignment Submissions Using a Data Mining Approach

Kathryn Burn-Thornton
Data Mining Group
Brunel University
Uxbridge, UK
e-mail: Kathryn.thornton@brunel.ac.uk

Tim Burman
Dept. Computing and Engineering Science
Durham University
DURHAM, UK
e-mail: tim.burman@dur.ac.uk

***Abstract*—In this paper, we describe a novel application of data mining techniques which can be used to identify multi-authorship contained within student submissions. We show that by regarding the pages of the submission as a set of Cascading Style Sheets, CSS type files, which we call author signature styles (ASSs), and accompanying information, it is possible to identify the number of author signature styles contained within the page, or document, irrespective of the number of pages concerned. We also describe how, as a by-product of this work, a set of author signature styles (ASSs) can be created during investigation of each submission and hence be used as a library, containing increasing membership, for comparison with future submissions by the same student. The implications of the use of ASSs for identification of future suspect submissions, and for comparison with future submissions by the same student, are discussed.**

Keywords-plagiarism; data mining.

I. Introduction

Government cuts in Higher Education funding have provided the driver for larger university class sizes, both face-to-face and online [19]. For class sizes greater than 50 this can mean that those marking essay style submissions may be unaware of the written style of the students and, in many cases, unable to put a name to a face [20-21]. For online students, the lecturer, or marker, may not ever meet the student [9, 13].

This lack of knowledge on the student puts the marker at a great disadvantage and provides a window of opportunity for those who are aware of the situation and who are keen to reuse material which may have been created by others i.e., those who are willing to plagiarize existing material. Such activity is readily facilitated by the virtual society which now makes it possible for students to access material from all over the world and with which the marker may not reasonably be expected to be aware [21].

Approaches to ameliorate this problem include continual assignment subject changes but it is not possible to ensure that they do not overlap with others set somewhere else in the world [20]. However, identification of whether the student's submission contains a duplication of information which may be found elsewhere on the superhighway is an approach to solving the plagiarism problem, which may be ideally suited to a software tool [2].

Although identical duplicate documents to those of student submission, or paragraphs, which may be readily found by making use of a simple search engine [24], submissions which contain modification of documents from various sources are harder to detect by this approach and a more sophisticated approach must be used to identify these. This has resulted in a 100 fold growth, over the last ten years, in published papers which outline approaches, and software tools, which may be used to provide aid in the detection of student plagiarism by universities [23].

However, the results from the use of plagiarism tools are often hard to follow using formal university procedure because of their determination of degree of commonality between the student submissions and other documents which are available [24-27]. In addition, the tools do not necessarily provide the user/investigator with an indication of whether, or not, the submission is individual original work. An approach which has not been used to identify 'suspect' student submissions, which may emanate from more than one author, is document signature style. This approach has an added advantage in that it is easier to follow up the results obtained using formal university procedures if required.

This paper describes a novel data mining approach, which enables documents to be identified which contain more than one document signature style. The first section describes current approaches which are used to identify 'suspect' student submissions.. This is followed by a discussion of two possible solutions which would enable document signature styles to be determined and a description of techniques which may be employed in order to achieve each of the potential solutions. Algorithms which may be gainfully employed in achieving the chosen solution are then outlined. The remaining sections discuss the investigations which were carried out in order to determine the effectiveness of the approach and the results of the investigations for the CSS type solutions – the ASS based solution. Conclusions regarding the results of the investigations are then drawn with future profitable avenues for investigation being discussed.

II. Exisiting Approaches and Tools

The vast majority of tools, in common use in a university environment, which enable the investigation of submission of non-original work such as TurnitinUK and Viper [23, 27] appear to make the simplest assumption that the submission of non original work by a student falls into the category of potential plagiarism. With this prior assumption that determination of plagiarism may achieved by comparing the student's submission with all other submissions, and documents, which are available throughout the world.

The process is readily suited to current pattern matching algorithms, and methods, especially if paragraphs similarity between documents is to be considered. It is a type of pattern matching engine which underpins plagiarism tools which are commonly used in a university environment to identify potential plagiarized submission [23, 27].

Despite their speed of document comparison most tools of this type present the user(investigator) with a problem. Namely, that unless the submission is a 'simple' combination of existing work many of the current plagiarism tools do not provide sufficiently large a percentage match between the student's submission and documents which may be available on the web in order to pursue further the investigation of the lack of originality in the submission using formal university approaches [29]. However, another approach to detection of non-original work in the student submission could prove profitable when the pattern matching approach fails, that of the author signature style [6, 12, 14, 23] (ASS) since all student submissions should emanate from one student so should contain only one ASS or a variant on the same ASS.

III. Document Signature Style

Document signature style makes the assumption that each individual has a unique writing style which is characterized by their individual use, and combination, of nouns, verbs and a other features which include referencing [7, 12, 14, 23]. If the document signature style were to vary throughout a document's paragraphs, pages and chapters this could provide an indication that the submitted document originated from more than one author and was not the submission from one individual.

Such variation in style could be used as a basis for a formal university approach as the student submissions profess that the word is their individual work, in other words from only once source. This approach could, if sufficiently accurate, prove to enable the task to be achieved faster, and hence enable more student submissions to be checked, because all the information in cyberspace is not be trawled for each submission. The following section outlines how such a solution can be achieved.

A. *Extraction of Signature Style*

In order to determine the unique author signature(s) present in the electronic submissions it necessary to determine which key elements of written documents will be used to determine the unique documents signature created by each author. Initial analysis of over 300 submissions from this university [29] suggested that the key elements of the signature required in order to determine whether, or not, a document emanates from one author, may be reduced to number of words in a sentence, number of lines in a paragraph, paragraph formatting, degree and use of grammar, type of language used, word spelling and referencing style. These key signature features are concomitant with those proposed at ICADPR for instance those in [7] and [10].

The first two elements of the signature are self explanatory but the others may require some clarification. Degree and use of grammar is taken to include the manner in which infinitives are used; use of, and types, of punctuation; use of plurality. Type of language is taken to mean language style which includes different types of English for instance UK and US. However, word spelling includes not only language spelling differences such as those found between UK & US, for example as in counsellor and counselor, but also frequency of typographical errors and spelling mistakes. The referencing style required by different bodies, and institutions, vary and can provide an indication of material which originates from more than one source.

A solution to this problem will be an approach which will enable extraction of the key signature elements, and their values, from paragraphs, and pages, and compare them with others in the same document and with those extracted from other documents. It would also be useful if the approach taken could show how the document would have appeared if written by a sole author if additional proof of multi-authorship was required for use in a formal university process.. The following section describes two possible solutions.

IV. Possible Solutions

Both of the solutions suggested in this section make use of approaches which we used in our web site maintainability tool [1]. The approaches make use of Cascading Style Sheets (CSS) or a combination of the eXtensible Markup Language (XML) in combination with the eXtensible Style Language (XSL) [17]. The approaches which we suggest make use of information extraction and representation. Some commonality can be observed between the first steps of the approaches, which are described in the next section, and that of Ghani [8] and Simpson [15].

A. CSS

If a CSS –based approach were used, a named author signature style (ASS) could be defined which would describe the values assigned to the key signature features. Once the ASS files were created, the signature of style of the author could not only be compared with others within the same document but it could also be applied to any document section and the output compared with that contained within the current, or other, submitted document. By using this approach the speed of investigation of submitted documents could be minimized by the reduction in the size of file which is required in order achieve comparison [16].

In practice, each section of the document being investigated could be converted directly to a section of ASS containing the feature values. Such an approach would require the use of a measure of uncertainty when mapping the samples of document and related ASS code to named signature styles. Figure 1 provides an example of how a page of text may be converted using such an approach.

Data mining would appear to provide a solution to this problem using clustering techniques.

The only drawback to this approach is that a library of assignable values for each key signature feature will need to be defined initially. However, this library may be updated as each submission is investigated.

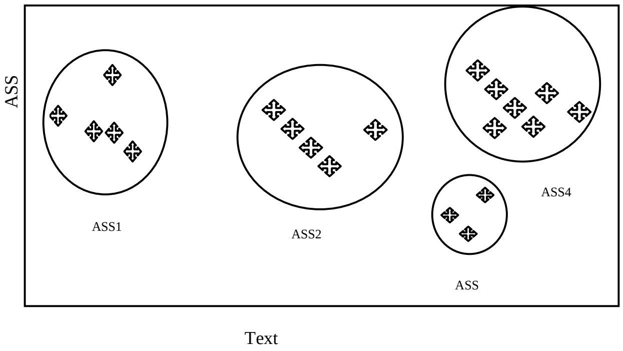

Figure 1. Clustering.

B. XML

For an XML approach all content information would be contained in an XSL file with its companion XML file containing the ASS feature information which would be recursively applied to the XSL document.

Using the example from Figure 1 this approach would result in the production of a XML file containing a section of text that would be marked up as a reference name, and the XSL file would contain a template which could be applied reference names in that document. Such an approach would readily facilitate comparison of documents because it would be relatively easy to target comparison of documents by investigation of specific signatures, ASSs.

Rigid definitions do not exist for XML tags which means that any appropriately defined names will have to be used in the XML file as well as a library of attributable values of the signature features, as in the CSS approach. However, a major drawback of this approach would be the need of consistency for XML tags and the possibility of ongoing modification to a centrally accessed XML tag dictionary.

Both requirements for the XML/XSL approach suggest that the CSS based solution may be more accurate to carry out comparison of signature styles in documents because even a alight variation is XML tags could result in a large discrepancy in ASSs and hence identify a document as containing information from more than one author when it does not.

The following section provides an introduction to data mining, the basis of the CSS , or ASS, approach.

V. Data mining

Data mining finds novel, potentially useful and ultimately understandable patterns from mountains of data [3] and has been used to mine data from diverse domains including the medical domain [5], pharamaceutical [4] and, as such, appears to be the ideal solution for finding the patterns of information contained within the files extracted from (and contained within) the student submissions. This is an approach which we used in our web maintainability tool [1].

Data mining determines the patterns by clustering the data according to variable values contained in the data [11]. Figure 1 shows how clustering could be carried out using pre-determined CSS, or ASS, files and unclassified student submissions. In this example – each sample in the classifier is marked with a cross indicating the document page giving rise to the sample, and the CSS section (ASS section) that was generated from the document. Samples that have similar values, appear to have been produced by the same author, are given the same classification.

In clustering, each CSS section would be classified in turn. If it is sufficiently similar to other previously classified sections, ASS, it is added to the same classification (class) as these other sections. If it is not sufficiently similar to another section, a new classification, a new ASS, is created.

There are many different classes of Data mining algorithm which can perform clustering with each class possessing different properties. It is these different properties which make each class suitable for analyzing different types of data [11]. The class of algorithms which appear to be particularly appropriate for mining the type of data of which CSS files are composed belong to the statistical & machine learning classes of algorithms. More information regarding this may be found from the results of the STALOG project [23]. These classes of algorithms are described, briefly, in the following section.

A. Suitable data mining Algorithms

The suitability of algorithms chosen from the statistical & machine learning classes, namely: - k nearest neighbours, linear (k-NN), quadratic & logistic discriminants, k means, rule based, decision trees and Bayesian classifiers are described and their appropriateness for the task in hand. These are the same algorithms which were discussed for the task of web site maintainability [1].

The most appropriate algorithm for the conversion from student document to CSS, ASS, from those listed above, is the k-NN algorithm. The other algorithms are not appropriate because they either require too many samples with which to build an effective model from which to work effectively in this application (decision trees, Bayesian classifiers), require numerical data (Fisher’s linear discriminants), or require prior knowledge of the classes (K Means). However, k-NN can work effectively with a small number of samples, can work with categorical data given an appropriate function to compare two samples, and does not require any prior knowledge of the number of classes, or authors.

The following sections describe the implementation of the CSS solution which has been described in this section.

VI. CSS Solution

In order to implement the k-NN algorithm some means of finding a numeric difference between two samples of student document and ASS is required. This can be achieved by determining the percentage of elements of the code in one sample which is not present in the other sample. In order to determine the difference between the two samples of ASS signature features, and their values, present in each sample signature needs to be investigated. A visual representation of this process may be seen in Figure 2.

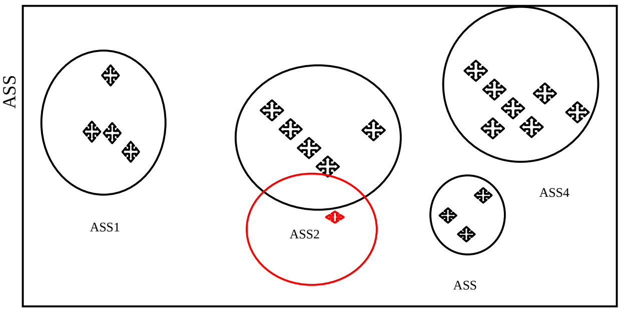

Figure 2. Clustering using K-NN.

In order to achieve this each section of student document, submission need to be represented by equivalent signature features and their values. In the same manner as presentation tags in HTML code these can be represented as signature tags. It is these adjacent signature tags which form clusters of tags and can be represented by a single ASS.

The first stage of the implementation of the k-NN algorithm is to create the signature tags from the original document and then each cluster of signature tags is converted to a ASS sample using a set of rules that are defined in a data file. This can be changed by the user as the ASS evolves, but a standard set of rules.

Each line is in the format:

Tag-name ASS-equivalent value

After each cluster is converted to an ASS the algorithm iterates through each sample and compares it to any that have already been classified. At the start of the loop, none will have been classified. Otherwise, a list of the other classified samples is created and ordered by difference to the new sample. If no sample is within a threshold distance, it is assumed that the new sample is not sufficiently similar to any previous classification, and so the user is prompted for a new classification for this sample. Otherwise, the closest k samples are taken from this list and the new sample is assigned the same classification as the majority of these k samples. An appropriate value of k can be found through trial and error during the implementation.

For the final conversion of the classifications to a style sheet, an arbitrary sample from each classification is used to supply the definition of the style, and the name assigned to the classification is used as the name of the style. As each sample in the class should be very similar, it should not matter which sample is used for the style definition.

A slight modification was made to the k-NN class so that it could be used to create an example document from an existing signature style. This modification was that a new author signature is not created if no close match among the previously classified samples is found. The contents of the style sheet are read in and set as the classified samples to provide the classification. The same approach is used for finding groups of pages with the same style. The major differences in this case is that the methods used to represent each page, and the differences between them – as well as the automatic naming procedure of a process which is to all intents and purposes completely unsupervised.

Each page, or paragraph, is represented by a set of feature information, including a list of the number of times each one is used, and the distribution of the feature tags throughout the page or paragraph. The combination of this set of information gives a good overall impression of the written signature style of the author.

The difference between two sets of information is found by the number of features, and values, that are not present in one set of information and is present in another, or those where the font or style is used more than twice as many times in one than in the other. The table distributions are compared using the chi-squared test. Each distribution is composed of 100 values, indicating the number of signature tags in that 100th of the section. The chi-squared value is calculated as the sum of the squares of the differences of each of these values, as given by the formula:

$$\chi^2 = \sum_{i=1}^{100} \frac{(x_i - y_i)^2}{y_i} \qquad (1)$$

where x is distribution of table tags in Section 1
y is distribution of table tags in Section 2.

The set of this information provides an overall value for the difference between the two pages. This can then be directly compared to the value for any other two pages. Again, if the page being classified is not sufficiently similar to any previously classified section, a new classification, or ASS, is created for it.

The following section describes investigations which were carried out in order to determine the effectiveness of the CSS methods to facilitate comparison of author signature styles (ASS) in the paragraphs comprising the students

VII. Investigations

In order to determine the effectiveness of the solution, a set of metrics were defined which enabled the effectives of the solution to be determined on a wide range of submitted documents.. This section describes the metrics used and the wide range of documents used.

A. Measures of Effectiveness: Metrics Used

The effectiveness of the solution was determined by the ease, and effectiveness, of extraction of file information from the source page into a separate author signature style sheet and the degree to which the content of the original pages remained unaltered once it has been re-produced by use of the style sheet.

Two metrics were also used to determine the effectiveness of the solution. Firstly, relative time for comparison of author signature styles in paragraphs contained within student submission by tool with that taken by a human carrying out the same task. Secondly, the number of author signature styles produced and number of differences between the signature style features in the original page, or paragraph, in the submission and that created using the ASS Documents Investigated

These submissions were chosen as examples of their wide range of document pages to which the algorithms, which represent the algorithm of the tool, can be applied because they represent a cross section of the variation in author styles contained with documents submitted at this university.

Sample	**Student Origin**	**First Language**	**Number**	Authorship
1	UK	English	20	Known Single
2	UK	English	20	Assumed Single
3	EU & UK (50:50)	50 English: 50 2nd English	20	Assumed Single
4	Not EU	Not English	20	Assumed Single
5	UK	English	20	Known Multi

Figure 3 provides examples of the wide range of student submissions which were investigated.

Sample 1 containing documents known to have been written by one author. Sample 2 contains UK students whose first language is English whilst Sample 3 contains an equal mix of EU and UK students. Sample 4 contains non EU students who are required to take TOEFL and who have all passed the level required to be admitted to the university. Sample 5 contains documents which are known to contain multiple authorship.

This range of documents should enable the performances of the algorithm, and hence tool, on different written styles of pages to be determined.

The following section describes the results from applying the metrics to the wide range of test documents.

Figure 3. Examples of submission types.

VIII. RESULTS

Simple plots are used to visualize the results. Figures 4 to 6 show the results of investigation of the three metrics.

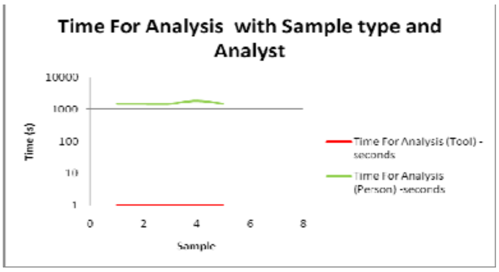

Figure 4. Relative time for paragraph authorship comparisons.

A. *Relative time for paragraph authorship comparisons*

The results of these investigations are shown in Figure 4. The figure shows that the tool was able to perform comparison up to 1000 times faster than the person carrying out the same task. The figure also shows that the time taken for the non-tool based comparison of the signature style in each paragraph varied from person to person and also from sample type to sample type. There was no difference in the comparison time for the tool because of the short time in which this was achieved, all within 1 second.

Half of the of people carrying out the task were unable to complete the comparison for any of sample size 4 because of the fluency in the written style of the student submissions.

B. *A count of the author signature styles produced*

The number of signature styles produced is dependent of the written content of each page. Figure 5 shows that, on average, two styles are produced from a page known to have been written by one author. The figure also shows that, on average, three styles are produced from a page of unknown authorship with the distribution of the number of styles produced being skewed towards the lower end. The tool accurately determined the number of the authors from the documents known to be multi-author. However, the figure shows that human determination was less accurate – especially for samples 3 and 4, those for which English was not a first language.

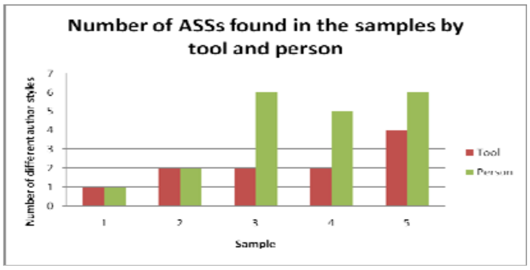

Figure 5. A count of the author signature styles.

C. *Information Differences*

These results shown in Figure 6 are consistent with that results of the ASS investigations in that information differences observed between the original, and key features of the, document are strongly correlated with the error in determining authorship number. Thus suggesting that if the ASSs contained in the document can be determined then it is it possible to reform key features of the original document for comparison with other student submissions and with future by the same student.

IX. CONCLUSIONS AND FUTURE WORK

We have described an approach for carrying out investigation of the plurality of the authorship of documents submitted by students, which is dependent upon Data mining-based clustering methods.

The results presented in section VIII show that this approach facilitates accurate investigation of the authorship of student document submission. Such results have the potential to be used in formal university procedures.

Is intended that further work will be carried out investigating the three key metrics in submission from other Faculties and universities. Work will also be carried out to modify the Data mining algorithm to maintain accuracy of Multi Author Determine across this new range of submissions.

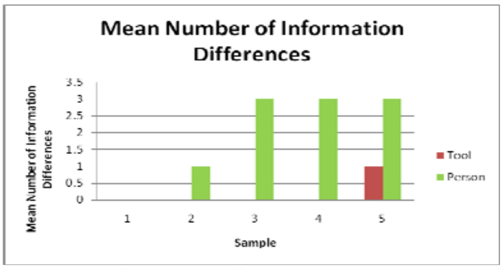

Figure 6. Information Differences between Original and Reformed Text produced.

ACKNOWLEDGMENT

Acknowledgment is made to Mark Carrington for his original project work in 2002 which led to development of this paper. The authors wish to thank the University of Durham for the provision of facilities to carry out this work.

REFERENCES

[1] K. E. B. Thornton, M. Carrington, and T. Burman, "A Data mining based method for web site maintenance," Intelligent Data Analysis, vol. 10, 2006, pp. 555-581.

[2] K.E. Burn-Thornton, D.M. Cattrall, and A. Simpson, "Polymorphic Functions for Data mining in A.T.M. Networks," Proc. 4th I.F.I.P. Conf., July 1996, pp.11-18, IIkely, UK. http://www.xent.com/summer96/0041.html. accessed 6/12/10.

[3] K.E Burn-Thornton., S.I. Thorpe, and J., A Attenborough,"Method for Determining Minimum Data Set Size Required for Accurate Domain Analysis." in Proc. PADD '00, International Data mining Conference, May 2000, pp. 161 –169, Manchester - ISBN 1 902426 08 8.

[4] K.E. Burn-Thornton and J. Bradshaw," Mining the organic compound jungle – a functional programming approach," chapter 11, IEE Practical Applications of Computing, March 1999, pp. 227-240.

[5] K.E. Burn-Thornton and L. Edenbrandt, "Myocardial Infarction - Pinpointing the Key Indicators in the 12 lead ECG Using Data mining," Computers and Biomedical Research, vol. 31, 1998, pp. 293-303..

[6] J. Cai, R. Paige and R. Tarjan,"More Efficient Bottom-Up Multi-Pattern Matching in Trees," Theoretical Computer Science, vol. 106, pp.21-60, 1992.

[7] C.E. Chaski,"Multilingual Forensic Author Identification through N-Gram Analysis," Paper presented at the annual meeting of the The Law and Society Association, TBA, Berlin, Germany 2010-06-04 from http://www.allacademic.com/meta/p177064_index.html, accessed 17/10/10.

[8] R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattery,"Data mining on symbolic knowledge extracted from the web," in Proc. of the Sixth International Conference on Knowledge Discovery and Data mining (KDD-2000), Workshop on Text Mining, pp. 21-29.

[9] J. Hewitt and C. Brett," The relationship between class size and online activity patterns in asynchronous computer conferencing environments," Computers and Education, vol. 49, pp., 1258-1271, 2007.

[10] B. Kövesi, J.M. Boucher, and S. Saoudi, "Stochastic K-means algorithm for vector quantization," Pattern Recognition Letters, vol. 22, pp. 603-610, 2001.

[11] D. Michie, D.J. Spiegelhalter, and C.C. Taylor (ed),"Machine learning, neural and statistical classification," New York: Ellis Horwood, 1994.

[12] A. Brink, L. Schomaker, and M. Bulacu,"Towards Explainable Writer Verification and Identification Using Vantage Writers," in. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 824-828, Parana, Brazil September 23- 26, ISBN: 0-7695-2822-8.

[13] N. Shadbolt, "Caught up in the web," Invited talk at the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02., 2002, pp. 317-334.

[14] I. Siddiqi and N.Vincent,, "Writer Identification in Handwritten Documents," Proc. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) in Proc. Document Analysis and Recognition, International, 2007, vol. 1, pp. 108-112.

[15] S.Simpson, http://www.comp.lancs.ac.ucomputing/users/ss/websitemgmt, accessed 6/12/10.

[16] I. Sommerville, "Software engineering," 5th ed., International computer science series, Wokingham, England : Addison-Wesley, 1996.

[17] Wilde E, Wilde's WWW. Technical foundations of the World Wide Web. London: Springer, 1999.

[18] ZigZag,www.zigzagdesign.co.uk/website_maintenance.htm, accessed 6/12/10.

[19] http://www.hefce.ac.uk, accessed 6/12/10.

[20] http://www.alluniversities.com/index.php, accessed 6/12/10.

[21] http://www.articlesnatch.com/Article/Uk-Academic-Writing-Service/1239456, accessed 6/12/10.

[22] http://www.euroscience.org/author-identification,28115,en.html, accessed 6/12/10..

[23] http://www.google.co.uk accessed 6/12/10

[24] www.scanmyessay.com accessed 6/12/10.

[25] cs.stanford.edu/~aiken/moss/ accessed 6/12/10.

[26] turnitinUK, www.submit.ac.uk/, accessed 6/12/10.

[27] http://www.brun.ac.uk , accessed 6/12/10