HAL open science

Detection of Hidden Intertextuality in the Scientific Publications

Cyril Labbé, Dominique Labbé

To cite this version:

Cyril Labbé, Dominique Labbé. Detection of Hidden Intertextuality in the Scientific Publications. 11th International Conference on Textual Data Statistical Analysis, Jun 2012, Liège, Belgium. pp.537-551. halshs-00709018

HAL Id: halshs-00709018

https://shs.hal.science/halshs-00709018v1

Submitted on 16 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Detection of Hidden Intertextuality in the Scientific Publications

Cyril Labbé[1] Dominique Labbé[2]

1 - Laboratoire d'Informatique de Grenoble, Université Joseph Fourier Cyril.Labbé@imag.fr

2 – PACTE (CNRS – IEP Grenoble) Dominique.Labbe@iep-grenoble.fr

Abstract

Intertextuality is the presence of one text contained within another. Hidden intertextuality is a problem for scientific publications. We propose a detection method combining calculation of intertextual distance and several classifications with the technique of the "sliding window" (to pinpoint any duplicated excerpts). This method is tested using a group of texts extracted from the IEEE bibliographic database.

Résumé

L'intertextualité est la présence d'un texte dans un autre. L'intertextualité dissimulée est un problème pour la publication scientifique. On propose une méthode de détection combinant le calcul de la distance intertextuelle, la classification et la technique de la fenêtre glissante. Cette méthode est testée à l'aide d'un groupe de textes dupliqués tirés de la base bibliographique de l'IEEE.

Mots-clés : intertextuality ; scientific literature ; intertextual distance ; tree-classification ; plagiarism.

1. Introduction

In literary analysis, inter-textuality is defined as the presence - either explicit or hidden – of one text inside another. Explicit intertextuality plays a legitimate role in scientific publications (quotations of the original publications on the same topic, of the related works, references, aknowledgments ...). However, hidden intertextuality is a perennial problem (Bouville 2008), and it is not a new idea that statistics can help fighting it (Ottenstein 1976).

We present a new set of procedures able to detect this hidden intertextuality in the scientific literature, and to pinpoint the phenomenon in the texts concerned, and to measure its importance. These procedures are developed with a number of actual cases, taken from one of the largest bibliographic databases online which is introduced at the beginning of this communication. The method is tested in the following sections.

2. A large corpus for experiments

The IEEE (Institute of Electrical and Electronic Engineers) is, alongside the ACM (Association for Computing Machinery), the leading association of electronic and computer scientists. Its bibliographic database (fee-paying) is the largest in electronics, information technology and related fields. In this data base, a number of texts are preceded by a caveat like the one reproduced in the Fig. 1 below (the quoted texts are used later in this paper).

The terms used by the IEEE define the phenomenon to be studied: **duplication of a significant proportion of one or several original text(s), without giving the references of this (or these) original(s) and without permission**. If it is the case, the IEEE requires that the assumed references to the derived text are replaced by references to the original(s), that is to say that it has been decided to declare a kind of authorship (re)attribution.

Notice of Violation of IEEE Publication Principles

"Estimating neutral divergence amongst Mammals for Comparative Genomics with Mammalian Scope" by Anup Bhatkar and J.L. Rana in the Proceedings of the 9th International Conference on Information Technology (ICIT'06)

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEE's Publication Principles.

This paper contains significant duplication of original text from the papers cited below.

The original text was copied without attribution (including appropriate references to the original author(s) and/or paper titles) and without permission.

Due to the nature of this violation, reasonable effort should be made to remove all past references to this paper, and future references should be made to the following articles:

"Distribution and intensity of constraint in mammalian genomic sequence" by Gregory M. Cooper, Eric A. Stone, George Asimenos, Eric D. Green, Serafim Batzoglou, and Arend Sidow in Genome Research, Jul 2005; 15, pp 901 – 913, Cold Spring Harbor Press.
and

"Quantitative Estimates of Sequence Divergence for Comparative Analyses of Mammalian Genomes" by Gregory M. Cooper, Michael Brudno, Eric D. Green, Serafim Batzoglou, and Arend Sidow in Genome Research, May 2003; 13, pp 813 – 820, Cold Spring Harbor Press

Fig. 1. *Example of notice preceding a paper found to be in violation of IEEE publication principles*

In the IEEE statement, one criterion is problematic: at which point can one consider that there is a "significant duplication"? It is proposed that this is adjudicated on the advices of experts - whom IEEE consults when it receives a complaint -, by examining the cases where they have decided that there is a clear violation of the principles of scientific publication.

A search through the entire IEEE database - more than 3 million references (according to the latter) - reveals the presence of more than three hundred papers preceded by this warning. It is therefore proposed to study these texts to determine the nature and the threshold of "significant duplication" and to test software tools able to detect such cases.

Within the limited scope of this paper, the method is presented with the help of a sample of 14 cases (set D) drawn at random out from the derived papers detected by the IEEE in its database and, consequently, preceded by such a warning (Appendix 1). These 14 texts are derived from 23 original papers (set O). In this set O, we have added an extra paper – when available - on the same topic by the same author(s). Therefore, this preliminary experiment focus on 42 original texts and 14 derived texts.

We proceed in three steps: calculation of the distances between these texts, identification of texts with abnormal proximities, and identification of the duplicated passage(s).

3. Text processing and intertextual distance calculation

Pdf files are converted into plain text files by the program "pdftotxt" (free software unix and windows version 3.01). During this operation, figures, graphs and formulas disappear, but the titles and captions of these figures and tables remain. To prevent the bibliographies from disturbing the experiments, the reference sections are removed from all texts.

The texts are segmented into word-tokens, using the procedure of the Oxford Concordance Program commonly used for English texts (Hockey & Martin, 1998), and the word-types are counted. In fact, the word-tokens and the word-types are strings of alphanumeric signs separated by spaces or punctuations. This procedure could be even further improved, for example by replacing all the abbreviations and inflections of a single word with a unique spelling convention (infinitive of verbs, singular masculine of adjectives…)

Then, the distances between one text and the others are measured using the following method (Labbé & Labbé 2001; Labbé & Labbé 2011).

Given two texts A and B, let us consider:
• N_A and N_B: the number of word-tokens in A and respectively B, ie the lengths of these texts;
• V_A and V_B: the number of word-types in A and respectively B, ie the vocabularies of the texts;
• F_{iA} and F_{iB}: the numbers of occurrences (absolute frequency) of a word-type *i* in texts A and respectively B;
• $|F_{iA} - F_{iB}|$ the absolute difference between the absolute frequencies of a word-type *i* in A and respectively B;
• $D_{(A,B)}$: the inter-textual distance between A and B is the sum of the absolute differences between the absolute frequencies of all the word-types of A and B ($V_{(A,B)}$):

$$D_{(A,B)} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - F_{iB} \right| \quad with\ N_A = N_B \qquad (1)$$

The distance index (or relative distance) is:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - F_{iB} \right|}{N_A + N_B} \qquad (2)$$

If the two texts are not of the same lengths ($N_A < N_B$), B is "reduced" to the length of A:
• $U = \dfrac{N_A}{N_B}$ is the proportion used to reduce B in B'.
• $E_{iA(u)} = F_{iB}.U$ is the theoretical absolute frequency of a word-type *i* in B'.

In the formula (1), the absolute frequency of each word-type in B is replaced by its theoretical absolute frequency in B':

$$D_{(A,B')} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right|$$

The formula (2) becomes:

$$D_{rel(A,B')} = \frac{\sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right|}{N_A + N_{B'}} \qquad (3)$$

This index varies evenly between 0 – the same vocabulary is used in both texts (with the same frequencies) – and 1 (the two texts share no word-tokens). This index has the three properties of a Euclidean distance (identity, symmetry, triangular inequality) and it can be interpreted as the proportion of words that are different in both texts. A distance of 0.5 means that the texts share 50% of their words-types, i.e. more or less half of their content.

In order to make this measure fully interpretable, one must bear in mind that:
• the texts must be sufficiently long (at least more than 1000 word-tokens). In the test corpus, the shortest text is 1597 word-tokens long;
• for short texts (those less than 3000 word-tokens), values of the index can be artificially high and sensitive to the length of the texts. In the test corpus, this is the case for the 10 shorter papers (out of a total of 54) that may appear to be a little more distant from the others than they would be actually;
• the lengths of the compared texts should not be too different. In any case, for English texts, the ratio of the smallest to the longest must be less than 1:7 (Labbé 2007). In the test-corpus, the longest length is 8697 word-tokens, i.e. six times the shortest length.

Inter-textual distance depends on four factors. In order of decreasing importance they are: genre, author, subject and epoch. In the corpora presented in Appendix 1, all texts are of the same genre (scientific papers) and are contemporary. Thus only the authorial and thematic factors remain to explain some anomalies detected by the calculus and the classifications. An unusually small inter-textual distance suggests striking similarities and/or texts by the same author(s). A large number of experiments and blind tests lead to the conclusion that for texts written in the same genre by contemporaneous writers, authorship is almost always the dominant factor (Labbé, 2007). Thus, the inter-textual distance offers a useful tool for "non-traditional authorship attribution" (Love 2002).

4. Detection of the anomalies

The anomalies within the test corpus are detected using two methods.

4.1 Calibration of two confidence intervals

Leaving aside the derived texts (set D), the distances between the original texts are grouped into two sets (Table 1). As mentioned above, to ensure that the calculations cover roughly the same number of texts, the set O was supplemented with an additional paper by each author (or group of authors) of the original texts.

	Texts by the same authors	Texts by different authors
Mean distance	0.3755	0.6092
Standard deviation	0.0389	0.0353
Confidence intervals: $\alpha = 0.05$	0.2994 – 0.4517	0.5401 – 0.6784
$\alpha = 0.01$	0.2761 – 0.4750	0.5189 – 0.6996

Table 1. *Mean distances between original texts(set O) by the same authors and by different authors, and confidence intervals.*

Among the 253 distances between O texts by different authors, only 4 are lower than 0.5189. The two lowest are: O0013 - O0021 (0.4727) and O0013 - O0017 (0.4770). These results are logical if one considers that O0013 is a survey paper on the topic covered by the two others (O0017 & O0021); then come O0017 - O0021 (0.51221) and O0001 – O0010 (0.5134), for the same reasons (same topic and very close reasoning). Leaving aside the case of the survey paper, it is an unlikely event that two papers – the lengths of which being between 1500 and 9000 word tokens - by different authors can be separated by distances of less than 0.500. Yet it is the case for all the duplicated texts in Appendix 1 (Table 2). Each of these duplicated paper is correctly associated to its respective original text(s).

Original	Detected	Distance
O0022	D0013	0.1345
O0015	D0012	0.1623
O0018	D0014	0.1684
O0002	D0004	0.2345
O0020	D0010	0.3018
O0019	D0010	0.3081
O0010	D0007	0.3110
O0024	D0008	0.3220
O0016	D0009	0.3452
O0009	D0006	0.3482
O0011	D0002	0.3753
O0014	D0009	0.3867
O0001	D0005	0.3933
O0021	D0011	0.4039
O0012	D0003	0.4109
O0005	D0001	0.4205
O0004	D0001	0.4297
O0017	D0011	0.4304
O0007	D0001	0.4326
O0006	D0001	0.4408
O0003	D0002	0.4606
O0013	D0011	0.4859
O0023	D0014	0.4876

Table 2. *Papers presented by different authors, but abnormally close together*

This first operation is completed with some classifications. The inter-textual distances allow clustering according to similarities between texts and graphical representations of their proximities (Sneath & Sokal, 1973; Benzecri, 1980; Roux, 1985; Roux, 1994). The best classification is the one that minimizes the distances between texts in a same cluster and maximizes the distances between these clusters.

A "nearest neighbor" classification - k-nn classification with k=1 (Cover & Hart 1967; Meyer et al. 2008) - is used to test the feasibility of automatic detection of hidden intertextuality. For this experiment, the original articles are first classified by authors. Then a 1-nn classification is done to assign each D paper to the class of its nearest neighbor. Using this method, all D papers listed in Appendix 1 are correctly classified with their real hidden author(s) (Table 2).

Two other methods are used: clustering analysis and tree classification (Felsenstein 2004a, 2004b; Luong 1988). In the present experiments, the two methods lead to the same conclusions. Due to the lack of space, clustering analysis is not displayed in this paper.

4.2 Tree-classifications

The tree below (Fig. 2) is drawn following Luong's formulae: "valued" trees and "grouping" method (Luong 1988). These formulae, methods and algorithms are fully explained in: Rulhman (2003).

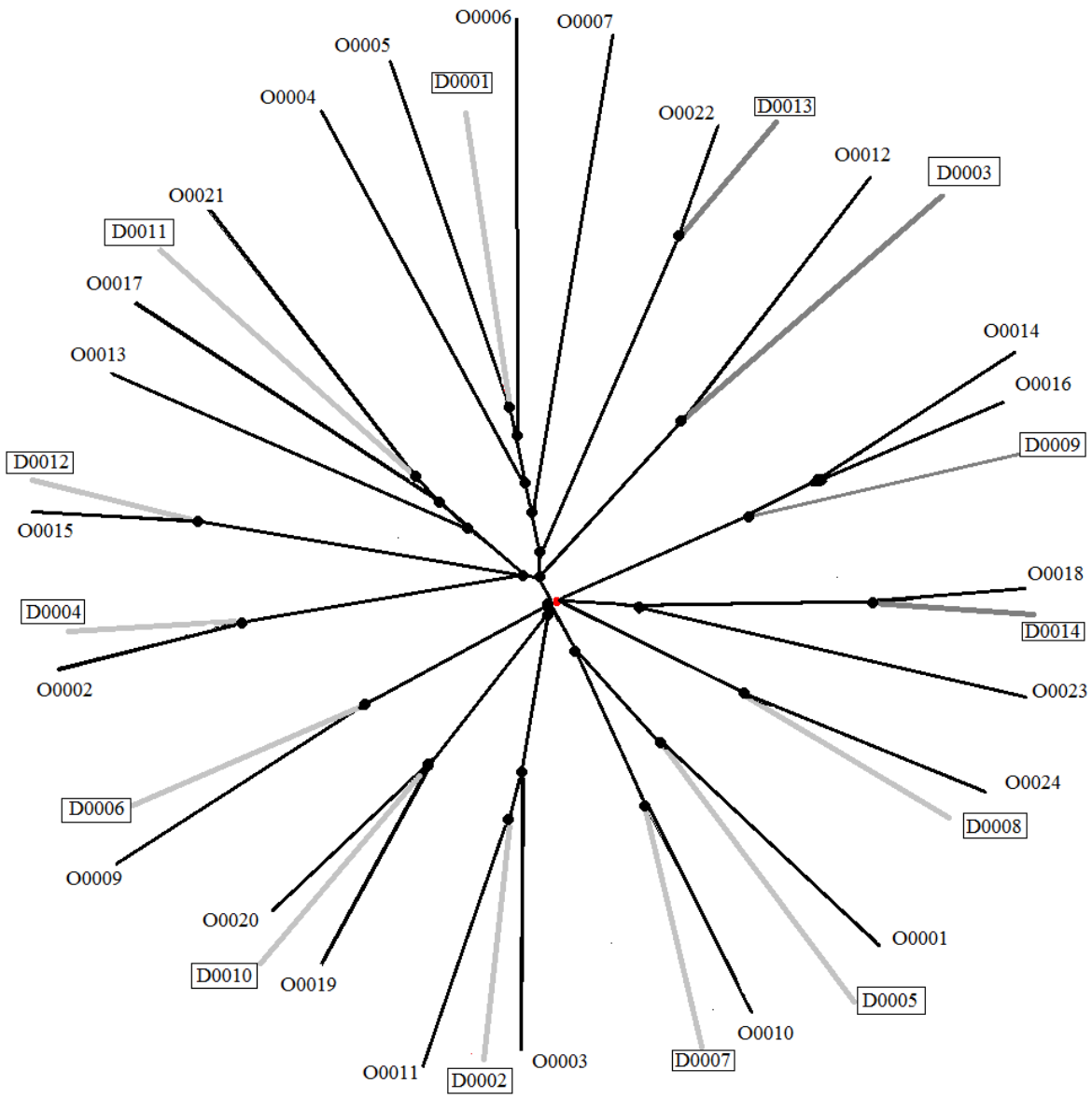

Fig 2. *Tree classification on the whole corpus (gray lines: duplications; bold lines: original texts)*

For example, O0022 and D0013 (top of the tree), are adjacent, as also are O0010 and D0007 (bottom of the tree). They form two sets of "neighbours" and these two groups are opposed. The edges ("stems" or "branches") link those four leafs to centrally located nodes which are created by the algorithm. Their relative positions are calculated in order to create edges proportional to the original distances. A leaf of the tree is linked to another by a path. The longer the path, the farther apart are the two texts.

This graph is "valued". That is to say that the path lengths are positive and proportional to the original values in the corresponding cells of the distance matrix. This calculation is very complex because this tree must represent the lengths of 666 different links.

A measure of quality is proposed (Labbé & Labbé 2008). The quality of a tree, such as that presented in Fig. 2, can be evaluated by comparing the 666 original indices with all the corresponding path lengths on the tree. If all these paths are exactly equal to their corresponding distance indices, the quality index will be equal to 1. This index is calculated for each path, each node and the whole tree. For the Fig. 2, all quality indexes for the paths

and the nodes are higher than 0.90; for the whole tree, this index is equal to 0.981. In other words, 98,1 % of the information contained in the distance matrix is faithfully represented on this tree.

This tree assigns all the cases detected by the IEEE with the original work(s) from which they are "inspired":
- 8 couples report simple "intertextualities" (the “inspiration” comes from a single original);
- 4 triplets: each of these D papers comes from a mix of two original texts;
- 1 quadruplet comes from the mix of three original papers (D0011);
- D0001 was actually created by mixing four originals, and this “chimera” is clustered in the middle of this group;

South of the graph, the two original texts (O0003 & O0011 quoted in the IEEE warning which is reproduced at the beginning of this paper) have the same derived (D0002). Both are by the same authors, on the same topic. Yet they are more distant between them that the couple (D0002-O0011) formed by one of the two originals with the text detected by the IEEE as a duplication of these two originals. The total length of these two originals is 9408 word tokens; the derived text is 2094 tokens long. Thus, one can ask how to locate precisely, in these three papers, the excerpts that have been duplicated?

5. Location of duplicated excerpts

This experiment is only on the texts, by different authors, with abnormally low distances highlighted by the classifications. To pinpoint the duplicated parts, it is proposed to divide each text into small windows of equal lengths in tokens and to compare each of these windows to all the other ones (Fig. 3). This method is fully discussed in (Labbé 2007). A similar technique is used in (Brixtel & Al. 2009).

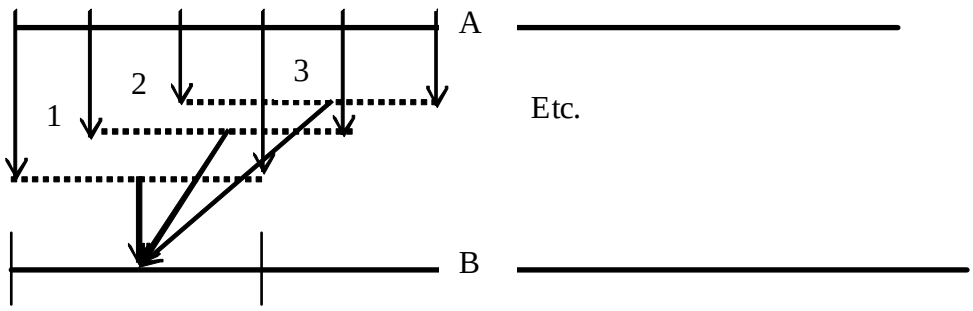

Fig. 3 *The "sliding window" method*

The procedure is the same as in Section 1: determination of a confidence interval and detection of anomalies. Here is an example: using a window of 250 token length and a pace of 125 tokens:
- mean distance between all the windows split in all original texts by same author(s): 0.687;
- standard deviation around the mean of all these distances: 0.040;
- lower limit of the confidence interval ($\alpha = 0.01$): 0.582.

It can be concluded, with less than 1% risk of error, that a distance of less than 0.58 – between two 250 token windows drawn from different texts by different authors - indicates excerpts to examine closely. The Table 3 gives a summary of these results for the three texts cited in the warning reproduced at the beginning of this paper,.

Slice	Duplicate (D0002)	Original papers	Original portions	smaller distance
1	0 - 250	O0011	0 - 750	0.38
2	250 - 500	O0011	500 - 1250	0.42
3	500 - 750	O0011	1000 - 1250	0.52
4	750 - 1000	O0003	6750 - 7250	0.50
5	1000 - 1500	O0003	7250 - 7750	0.25
6	1500 - 1600	O0011	2000 - 2250	0.54

Table 3 *Detection of the duplicated portions with the help of the sliding window*

The slice n°5 demands particular attention. Both texts can be read in parallel in Appendix 2. Of course, in this case, duplication is particularly rough, but it is interesting to note that the combination of the sliding window with the intertextual distance allows one to pinpoint the problem, and put the relevant passages in parallel.

6. Conclusions

This preliminary experiment was designed to test the method, carefully checking each text and controlling all the parameters. A larger sample is being set up. Subject to the future results, it is possible already to draw three conclusions.

• First, the cases found in the IEEE database seem relatively simple: large excerpts from one or several original text(s) have been imported into a subsequent text with little modification. There are certainly more hidden cases. For example, the translation into another language or the adoption of original ideas without using the same vocabulary (Alzahrani 2011), not to mention a related problem: the same author(s) duplicating the same paper with few cosmetic modifications…

• Second, this preliminary experiment suggests that combining intertextual distance with classification provides an effective tool for the detection of hidden intertextuality in scientific literature, which is logical since these tools are able to recognize texts by the same author. Finally, the technique of the sliding window pinpoints the passages that may have been significantly duplicated – according to the IEEE standards.

These data-mining tools would be useful for decision making, especially for detecting duplications and for allowing conference organizers, journal editors and database managers to counter these practices. Of course, automatic procedures are only an aid and not a substitute for careful reading. One must keep in mind that cases like the review article – mentioned in this paper - can still occur. This kind of “false positive” is possible and only a manual control can definitively rule out this possibility.

These tools may also be configured to scan the web in search of new scientific publications, comparing and contrasting them with those already known, detecting hidden intertextuality but also the guenine original contributions.

• Thirdly, Our purpose is not to stigmatize individuals. However, it is necessary to use actual cases - and to give the references - in order to allow the reader to check our findings, and the researchers to develop software and to calibrate them on the expert practice. We chose not to use words like "plagiarism", “fraud” or "copy", etc. These notions convey moral or legal connotations that are far beyond the statistical approach. It seems better to use the concept of "significant duplication" (as defined by the IEEE) and the concept of “hidden intertextuality”.

This problem should be seriously considered. When a scientist addresses a new topic, the first step is to find the original publications on this topic and the related articles in the field.

Establishing the origins of ideas, algorithms, data is not a moral issue, it is an important condition for the advancement of knowledge and for sharing concepts, tools and data between researchers. Hidden intertextuality complicates the research and, most importantly, it undermines the confidence between researchers.

References

Alzahrani S., Salim N. & Abraham A. (2011). "Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods". *IEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, (PP 99).

Arnold D. N. (2010). Integrity Under Attack:The State of Scholarly Publishing. *SIAM News. Journal of the Society for Industrial and Applied Mathematics*. 42-10, Décembre 2010 (consultable en ligne sur le site de SIAM).

Bouville M. (2008). Plagiarism: Words and ideas. *Science and Engineering Ethics*, 14(3), 311-322.

Brixtel R., Lesner B., Bagan G. & Bazin C. (2009). De la mesure de similarité de codes sources vers la detection de plagiat : le “Pomp-O-Mètre”. MajecSTIC 2009.

Cover T. M. & Hart P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*. 13, 21–27.

Felsenstein J. (2004a). *Inferring Phylogenies*. Sunderland: Sinauer Ass.

Felsenstein J. (2004b). *Package of Programs for Inferring Phylogenies (PHYLIP)*. Seattle: University of Washington.

Hockey S. & Martin J. (1988). *OCP Users’ Manual*. Oxford: Oxford University Computing Service.

Labbé C. & Labbé D. (2001). Inter-Textual Distance and Authorship Attribution Corneille and Moliere. *Journal of Quantitative Linguistics*. 8(3), 213–231.

Labbé C. & Labbé D. (2008). Peut-on se fier aux arbres ? *9e Journées internationales d’analyse statistique des données textuelles*. Lyon: Presses Universitaires de Lyon, 635–645.

Labbé C. & Labbé D. (2011). La classification des textes. Comment trouver le meilleur classement possible au sein d’une collection de textes ? *Images des mathématiques. La recherche mathématique en mots et en images*, 28 mars 2011.

Labbé D. (2007). Experiments on Authorship Attribution by Inter-Textual Distance in English. *Journal of Quantitative Linguistics*. 14(1), 33–80.

Love H. (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Luong X. (1988). *Méthodes d’analyse arborée. Algorithmes, applications*. PhD thesis. Paris: Université de Paris V.

Meyer D., Hornik K. & Feinerer I. (2008). *Text Mining Infrastructure in r*. 25(5), 569–576.

Ottenstein K. J. (1976). "An Algorithmic Approach to the Detection and Prevention of Plagiarism." vol. 8: *ACM*, 1976, pp. 30-41.

Aknowledgments

We thank Tom Merriam for his help in the development of software, for his valuable comments and for his careful reading of a previous version of this paper.

Appendix 1. The corpus

	Duplicated		Originals
D0001	T. J. Hammons, "Status of International Interconnections and Electricity Deregulation in Africa"	O0004	P Naidoo, L. Musaba, W Balet & A Chikova, "Toward Developing a Competitive Market for Regional Electricity Cross Border Trading : the Case of the Southern African Power Pool"
		O0005	A. Majeed, H A Karim, N.H Al Maskati, S. Sud, "Status of Gulf Co-Operation Council (GCC)

			Electricity Grid System Interconnection"
		O0006	Ahmed Zobaa, "Status of International Interconnections"
		O0007	Raymond Johnson, "Impact of Privatization and Deregulation on Infrastructure Development in Africa"
D0002	Anup Bhatkar & J.L. Rana, "Estimating neutral divergence amongst Mammals for Comparative Genomics with Mammalian Scope"	O0003	Gregory M. Cooper, Eric A. Stone, George Asimenos, NISC Comparative Sequencing Program, Eric D. Green, Serafim Batzoglou and Arend Sidow, "Distribution and intensity of constraint in mammalian genomic sequence"
		O0011	Gregory M. Cooper, Michael Brudno, NISC Comparative Sequencing Program, Eric D. Green, Serafim Batzoglou, and Arend Sidow, "Quantitative Estimates of Sequence Divergence for Comparative Analyses of Mammalian Genomes"
D0003	Krzysztof Szafranski, "Analysis of Hemodynamics of Intercranial Saccular Aneurysms"	O0012	Yiemeng Hoi, Hui Meng, Scott H. Woodward, Bernard R. Bendok, Ricardo A. Hanel, Lee R. Guterman, and L. Nelson Hopkins, "Effects of Arterial Geometry on Aneurysm Growth: Three-dimensional Computational Fluid Dynamics Study"
D0004	David I. Eromon, "High Temperature Superconducting (HTS) Generator Field Coil with Influence of Thermal AC Losses"	O0002	NMagnusson and M Runde, "The influence of thermal gradients on AC losses in high-temperature superconducting coils"
D0005	Rahul Choudhari, Ajay Choudhari, R. D. Choudhari, "Increasing Search Engine Efficiency using Cooperative Web"	O0001	Jie Xu Qinglan Li Huiming Qu Alexandros Labrinidis, "Towards a Content-Provider-Friendly Web Page Crawler"
D0006	Hong Fei, Liu Rui, Bai Yu, “Performance Evaluation of the Burstiness Impact with a Realistic IP Structure Model"	O0009	Chloé Rolland, Julien Ridoux, Bruno Baynat, Vincent Borrel, “Using LiTGen, a realistic IP traffic model, to evaluate the impact of burstiness on performance”
D0007	Umesh Sehgal, Kuljeet Kaur, Pawan Kumar, “The Anatomy of a Large-Scale Hyper Textual Web Search Engine”	O0010	Sergey Brin, Lawrence Page, “The anatomy of a large-scale hypertextual Web search engine”
D0008	Baolin Sun, Hua Chen. “An Intrusion Detection System for AODV”	O0024	Yang Tseng, Poornima Balasubramanyam, Calvin Ko, Rattapon Limprasittiporn, Jeff Rowe & Karl Levitt. “A Specification-based Intrusion Detection System for AODVC”
D0009	HuaiKou Miao and JunFeng Wu. “Applying Formal Methods to Compositionality Description of Web Service”	O0016	M. Solanki, A. Cau & H. Zedan. “Introducing Compositionality in Web Service Descriptions”
		O0014	M. Solanki, A. Cau & H. Zedan. "Augmenting Semantic Web Service Description with Compositional Specifications"
D0010	M. Aruna, M.P. Suguna Devi & M. Deepa. “Measuring the Quality of Software Modularization using Coupling-Based Structural Metrics for an OOS System”	O0020	Santonu Sarkar, Girish Maskeri Rama & Avinash C. Kak. “API-Based and Information-Theoretic Metrics for Measuring the Quality of Software Modularization”
		O0019	Santonu Sarkar, Avinash C. Kak & N. S. Nagaraja. "Metrics for Analyzing Module Interactions in Large Software Systems“

D0011	Dong Lingxun, Dou Lihua & Feng Heping. "Hybrid Time-Optimal Predictive Control for Mechanical Systems with Backlash Nonlinearity"	O0017	Mario Vasak, Mato Baoti´c, Ivan Petrovi´c & Nedjeljko Peri´c. "Hybrid Theory Based Time-Optimal Control of an Electronic Throttle"
		O0013	Mattias Nordin & Per-Olof Gutmanin. "Controlling mechanical systems with backlash — a survey"
		O0021	P. Rostalski, T. Besselmann, M. Bari, F. Van Belzen & M. Morari. "A hybrid approach to modelling, control and state estimation of mechanical systems with backlash"
D0012	K. Inderjeet, T. Kamal, M. Kulkarni, G. Daya & A. Prabhjyot. "Adaptive OFDM Vs Single Carrier Modulation with Frequency Domain Equalization"	O0015	Andreas Czylwik. "Comparison between Adaptive OFDM and Single Carrier Modulation with Frequency Domain Equalization"
D0013	H.M. Khodr, Zita A. Vale & Carlos Ramos. "Optimal Cost-Benefit for the Location of Capacitors in Radial Distribution Systems"	O0022	H.M. Khodr, F.G. Olsinab, P.M. De Oliveira-De Jesus & J.M. Yustad. "Maximum savings approach for location and sizing of capacitors in distribution systems"
D0014	Dejia Shi, Li Wang & Jing He. "The Design of Multi-agent System in IDAPS Microgrid"	O0018	Victoria M. Catterson, Euan M. Davidson & Stephen D. J. McArthur. "Issues in Integrating Existing Multi-agent Systems for Power Engineering Applications"
		O0023	M. Pipattanasomporn, H. Feroze & S. Rahman."Multi-agent Systems in a Distributed Smart Grid: Design and Implementation"

Appendix 2 Detection of duplicated passages with the sliding window method

Duplicate paper	**Original paper**
(D0002 : from the 1000th to the 1500th word-token)	(O0003 : from the 7250th to the 7750th word-token)
[armadillo, horse, cow, sheep, indianmunjtak, pig, rabit, galago, lemur, mouse-lemur, marmoset, dusky-titi, squirrel-monkey, vervet, baboon, macaque, oraguntam, gorilla, chimp, wallaby, monodelphis, opussum. The reference human sequence for this targeted region corresponds to NCBI build 35, i.e., human chromosome 7, 115404472—117281897. For all of our analyses] we treat the first human base in this region as position 1. This region contains 10 RefSeq genes, 40.2% repetitive DNA, and 38.4% G+C. Gene annotations for the human sequence were obtained from the UCSC Genome Browser (http://genome.ucsc.edu/) using the RefSeq gene track ([7]); this includes 151 unique exons (in which an exon consisting of both UTR and coding sequence is split into separate "unique" exons) totaling 36,959 bases.	we treat the first human base in this region as position 1. This region contains 10 RefSeq genes, 40.2% repetitive DNA, and 38.4% G+C. Gene annotations for the human sequence were obtained from the UCSC Genome Browser (http://genome.ucsc.edu/) using the RefSeq gene track ([7]); this includes 151 unique exons (in which an exon consisting of both UTR and coding sequence is split into separate "unique" exons) totaling 36,959 bases.
Alignment	**Alignment**
We used a combination of both global and local techniques to construct a multiple sequence alignment of set of these sequences. This strategy ensures that rearrangement events, identified as high-scoring local alignments, are properly captured and placed in the context of a global alignment. First, we compared each nonhuman sequence to the human using the program Shuffle-LAGAN ([8]). Shuffle-LAGAN is effective at the identification of	We used a combination of both global and local techniques to construct a multiple sequence alignment of set of these sequences. This strategy ensures that rearrangement events, identified as high-scoring local alignments, are properly captured and placed in the context of a global alignment. First, we compared each nonhuman sequence to the human using the program Shuffle-LAGAN ([8]). Shuffle-LAGAN is effective at the identification of rearrangements, such as translocations and inversions, in the context of a global, pairwise alignment. The nonhuman sequences are subsequently reordered and reoriented (i.e., shuffled) so that the local alignment chains are monotonic with respect to the human sequence. In this process, regions of the nonhuman sequences that lack

rearrangements, such as translocations and inversions, in the context of a global, pairwise alignment. The nonhuman sequences are subsequently reordered and reoriented (i.e., shuffled) so that the local alignment chains are monotonic with respect to the human sequence. In this process, regions of the nonhuman sequences that lack detectable similarity to any region of the human are clipped and deleted. These rearranged sequences are thus orthologously collinear with the human sequence. We then aligned the rearranged sequences using MLAGAN, a global multiple sequence aligner previously shown to be effective and accurate for multiple alignment of mammalian genomic sequences ([9]). The tree supplied to MLAGAN for this step is similar to the topology shown in Figure 1(on the following page), but it includes a small number of topology changes designed to align longer branch groups later; this step is necessary because alignment accuracy is best when species with the greatest sequence similarity are aligned first. We used parameters similar to those used to generate alignments of the human, mouse, and rat genome sequences ([10]). Entire tree construction and estimation of the neutral rate for the entire tree. We extracted all of those regions from the uncompressed alignment corresponding to the highest-scoring constrained elements in the human sequence yielding an alignment of 97,274 columns. Using a species topology as shown in Figure 1 defined ([11];[12]), we obtained the maximum likelihood branch lengths using PHYLIP (http://evolution.genetics.washington.edu/phylip.html), with the HKY 85 model of nucleotide substitution ([13]). The ML-tree is shown in Figure 1. Given the relative branch-length tree (Figure 1), we estimated the neutral rate for the entire tree as follows. Briefly, we estimate the neutral divergence among closely related species (ranging from 3% to 10% difference; Table 1), and subsequently extrapolated these rate estimates over the entire relative branch-length tree. As a source of aligned neutral DNA, we began with the uncompressed global alignment and excluded all of those alignment regions containing unambiguously constrained elements in the human sequence (the complement of the alignment used to determine the relative branch-length tree	detectable similarity to any region of the human are clipped and deleted. These rearranged sequences are thus orthologously collinear with the human sequence. In this process, regions of the nonhuman sequences that lack detectable similarity to any region of the human are clipped and deleted. These rearranged sequences are thus orthologously collinear with the human sequence. We then aligned the rearranged sequences using MLAGAN, a global multiple sequence aligner previously shown to be effective and accurate for multiple alignment of mammalian genomic sequences (Brudno et al. 2003a). The tree supplied to MLAGAN for this step is similar to the topology shown in Figure 1B, but is rooted on the marsupial branch and includes a small number of topology changes designed to align longer branch groups later; this step is necessary because alignment accuracy is best when species with the greatest sequence similarity are aligned first. We used parameters similar to those used to generate alignments of the human, mouse, and rat genome sequences (Brudno et al. 2004). **Tree construction and estimation of the neutral rate** We extracted all of those regions from the uncompressed alignment corresponding to the highest-scoring constrained elements in the human sequence yielding an alignment of 97,274 columns. Using a species topology previously defined (Madsen et al. 2001; Murphy et al. 2001), we obtained the maximum likelihood branch lengths using SEMPHY (Friedman et al. 2002), with the HKY 85 model of nucleotide substitution (Hasegawa et al. 1985; Fig. 1B). Given the relative branch-length tree (Fig. 1B), we estimated the neutral rate for the entire tree essentially as previously described (Cooper et al. 2003). Briefly, we estimate the neutral divergence among closely related species (ranging from 3% to 10% difference; Table 3), and subsequently extrapolated these rate estimates over the entire relative branch-length tree. As a source of aligned neutral DNA, we began with the uncompressed global alignment and excluded all of those alignment regions containing unambiguously constrained elements in the human sequence (the complement of the alignment used to determine the relative branch-length tree above). Divergence estimates were then made for each closely related group of species (neutral rate between 0.03 and 0.10 subs/site) using baseml of the PAML (Yang 1997) software package with the HKY 85 model of nucleotide substitution (Hasegawa et al. 1985).