Are the major bibliographic databases reliable?

Cyril Labbé, Dominique Labbé

Are the major bibliographic databases reliable?

Cyril Labbé[1] Dominique Labbé[2]

[1] - Laboratoire d'Informatique de Grenoble, Université Joseph Fourier Cyril.Labbe@imag.fr
[2] - PACTE, Institut d'Etudes Politiques de Grenoble Dominique.Labbe@iep-grenoble.fr

Technical report: Appendices and data available on request.

Abstract : It is generally assumed that fee paying bibliographic databases guarantee the quality of the data they sell. This technical report shows that this assertion can be challenged. It demonstrates a software method of detecting duplicate and fake publications. The charged-for services (such as IEEE Xplore) accept and index these kinds of publications.

Keywords : Bibliographic Databases ; Scientific Conferences ; Data Quality ; Fake Publications ; Text-Mining ; Inter-Textual Distance ; Google Scholar ; Scopus ; WoK.

Newspeak was founded on the English language as we now know it, though many Newspeak sentences, even when not containing newly-created words, would be barely intelligible to an English-speaker of our own day.

George Orwell (1949), The Principles of Newspeak. Appendix of 1984.

1 Introduction

Editorial databases such as Scopus (Elsevier) or ISI-Web of Knowledge (WoK Thomson-Reuters) require access fees. They are generally regarded as reliable tools for bibliographic research and for the evaluation of researchers, laboratories, and universities. This is mainly because they store only publications in journals and conferences in which peer selection is supposed to guarantee the quality of the indexed publications. Data quality would also seem to be secured by the publisher of journals and of conference proceedings. Therefore, to our knowledge, the quality of these bibliographic databases has never been questioned.

Text-mining tools are presented and used to detect problematic or questionable papers such as meaningless publications. The method has enabled the identification of dozens of bogus scientific papers.

2 Corpora and methods

All the texts used are present in bibliographic databases (*Scopus* and *WoK*). They are either available from the conferences' web sites, or from the publishers' web sites, like the Institute of Electrical and Electronic Engineers (IEEE) or Association for Computing Machinery (ACM) websites.

Three a priori above-reproach corpora: Texts, available online, from three recent conferences were chosen:

- Corpus X: The texts were downloaded from the site ACM Digital Library: http://portal.acm.org/. Organizers state : 126 "Full papers" (10 pages), 165 "short papers (4 pages), 20" demo papers "(2 pages) or a total of 311 documents. Announced acceptance rates for this edition and the various categories of publications are 13.3%, 17.5% and 52% respectively.
- Corpus Y: The texts were downloaded from the site of the IEEE: http://ieeexplore.ieee.org/. This corpus comprises 150 "regular papers" (4 pages). Official acceptance rate for this edition is 28%.
- Corpus Z: The texts were downloaded from the site of the Conference. Organizers count: 159 documents, which are divided into 3 tracks: one track with 58 documents (acceptance rate 18.4%), 33 documents for the second track (acceptance rate 16.1%), 30 + 6 documents for the last track (acceptance rate unknown), and 32 demonstrations (36% acceptance rate). The used corpus for this conference comprises 153 documents (121 articles and 32 "demos").

Automatically generated, deliberately faked texts: A specific corpus "Antkare" is added. It contains 100 documents automatically generated using the software SCIgen (http://pdos.csail.mit.edu/scigen/). This software, developed at MIT in 2005, generates random texts without any meaning, but having the appearance of research papers in the field of computer science, and containing summary, keywords, tables, graphs, figures, citations and bibliography.

For the Antkare experiment, SCIgen was modified so that each article had references to the 99 others– creating a link farm. Thus, all these texts have the same bibliography. Google Scholar retrieved these faked online articles and, as a result, Ike Antkare's H-index reached 99, putting him in the 21st position of the most highly cited scientists [8].

Texts Processing: Pdf files are converted to plain text files by the program "pdftotxt" (free software unix and windows version 3.01) that extracts the text from pdf files. During this operation, figures, graphs and formulas disappear, but the titles and captions of these figures and tables remain. To prevent the 100 identical references in the corpus Antkare from disturbing the experiments, the bibliographies have been removed from all texts in all corpora.

The texts are segmented into word-tokens using the Oxford Concordance Program commonly used for English texts [7]. In fact, the word-tokens are strings of alphanumeric signs separated by spaces or punctuation.

3 Text mining tools

Distances between each text and all the others (inter-textual distances) are computed. Then these distances are used to determine which texts, within this large set, were closer to each other and may thus be grouped together (cluster analysis).

3.1 Inter-textual distance

The distance between two texts A and B is measured using the following method (for a detailed presentation, see [9, 10, 13, 12]). 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;
- F_{iA} and F_{iB} : the absolute frequencies of a type i in texts A and respectively B;
- $|F_{iA} - F_{iB}|$ the absolute difference between the frequencies of a type i in A and respectively B;
- $D_{(A,B)}$: the inter-textual distance between A and B is as follows:

$$D_{(A,B)} = \sum_{i \in (A \cup B)} |F_{iA} - F_{iB}| \quad with \quad N_A = N_B \tag{1}$$

The distance index (or relative distance) is as follows:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A \cup B)} |F_{iA} - F_{iB}|}{N_A + N_B} \tag{2}$$

This index can be interpreted as the proportion of different words in both texts. A distance of 0.5 means that the texts share 50% of their words-types.

If the two texts are not of the same lengths in tokens ($N_A < N_B$), B is "reduced" to the length of A:

- $U = \frac{N_A}{N_B}$ is the proportion used to reduce B in B'
- $E_{iA(u)} = F_{iB}.U$ is the theoretical frequency of a type i in B'

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

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

Putting aside rounding-offs, the sum of these theoretical frequencies is equal to the length of A. The Equation (2) becomes:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A \cup B)} |F_{iA} - E_{iA(u)|}}{N_A + N_{B'}}$$

This index varies evenly between 0 – the same vocabulary is used in both texts (with the same frequencies) – and 1 (both texts share no word-tokens).

In order to make this measure fully interpretable:

- the texts must be long enough (at least more than 1000 word-tokens),
- one must consider that, for short texts (less than 3000 word-tokens), values of the index can be artificially high and sensitive to the length of the texts, and
- the lengths of the compared texts should not be too different. In any case, the ratio of the smallest to the longest must be less than 1:10.

Inter-textual distance depends on four factors. In order of decreasing importance, they are as follows: genre, author, subject and epoch. In the corpora presented above, all texts are in 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 classification. An unusually small inter-textual distance suggests striking similarities and/or texts by the same author.

3.2 Automatic clustering

The inter-textual distances allow clustering according to similarities between texts (classifications) and graphical representations of their proximities [21, 3, 17, 18].

An automatic cluster analysis is performed on the inter-textual distance matrix, using the following method. The algorithm proceeds by grouping the two texts separated by the smallest distance and by recomputing the average (arithmetic mean) distance between all other texts and this new set, and so on until the establishment of a single set.

These successive groupings are represented by a dendrogram with, in ordinates, the relative distances corresponding to the different levels of aggregation (see Figure 2 and 3).

To correctly analyze these figures, it must be also remembered that:

- whatever their position on the horizontal axis, the proximity between two texts or groups of texts is measured by the height at which the vertices uniting them converge, and
- the technique sometimes results in "chain effects": some similarities between texts are indistinguishable because the vertices connecting them are erased by aggregations performed at a lower level.

4 Detection of forgeries and duplicates in the three corpora

Intra-corpus distances: For each corpus, distances are ranked by ascending values and distributed in equal interval classes. Fig. 1 shows these four distributions.

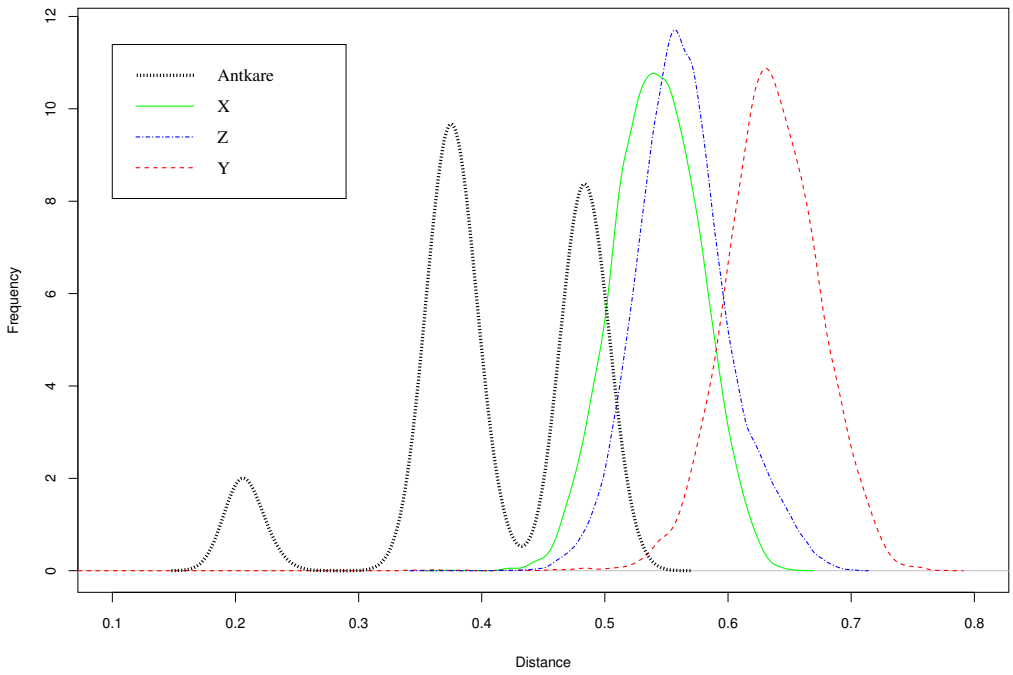

Figure 1: Distribution of intra-corpus distances.

The X, Y and Z corpora have the classic bell curve profile suggesting the existence of relatively homogeneous populations (here a large number of contemporary authors writing in a similar genre and

on more or less similar themes). X and Z have a comparable mean/mode and a similar dispersion. In contrast,

- Y has a high average distance and a higher dispersion around this mean, indicating heterogeneity of papers, but also suggesting the presence of anomalies (these two explanations are not mutually exclusive);
- On the left of the graph, the curve with three modes is the distribution of distances between the 100 faked texts by Ike Antkare. This trimodal distribution suggests the existence of two different populations within the texts generated by SCIgen: a small group with very low internal distances are centered on 0.2 - these are short texts (about 1600 word-tokens) - and the other group, with a greater number of texts, containing longer texts (about 3000 word-tokens): Their internal distances are centered on 0.38. The third mode is the distance between these two groups. Therefore it can be inferred that SCIgen uses at least two different ways to generate texts. This is confirmed by the classification (see Figures 2 and 3).

Classification of corpora: The classification and its representation by a dendrogram (Figure 2) show 4 main groups:

- In the center, a large body (C) includes all texts Z and almost all X texts.
- On the right (D) and on the extreme left (A), the texts of the Y Conference meet at the higher levels, confirming the heterogeneity of this conference.
- There is very little intermingling between on one side X, Z and Y on the other side. In other words, most of the papers presented at the Y conference are not of the same nature as those presented at the other two conferences.
- All the chimeras generated by SCIgen for Ike Antkare are grouped in B into two homogeneous groups and connected at a very low level. Thus, SCIgen texts are not "close" to natural language and are distinct from the scientific papers they are supposed to emulate.

Four "genuine-fake" texts: In the dendrogram in Figure 2, the number (1) branches are four Y texts that are clustered within the corpus Antkare. These four texts are "genuine" publications because they have, at least formally, been selected by peer reviewers. They are available (on payment) and referenced by sites of scientific publishers (WoK, Scopus, IEEE). But these texts are fake publications because they have the characteristics of the texts generated using SCIgen: absurd titles and figures, faked bibliographies, mixture of jargon with no logic.

Duplicated publications: Number **(2)** branch is a zero distance (0.006) between two Y papers. Only the titles are different. It reveals that an identical text was presented (and accepted) twice.

Smallest distances (without SCIgen texts): The branches of the dendrogram numbered 2 to 8 are the texts with the smallest distances (less or equal to 0.35) all sharing a common subset of authors and very similar topics.

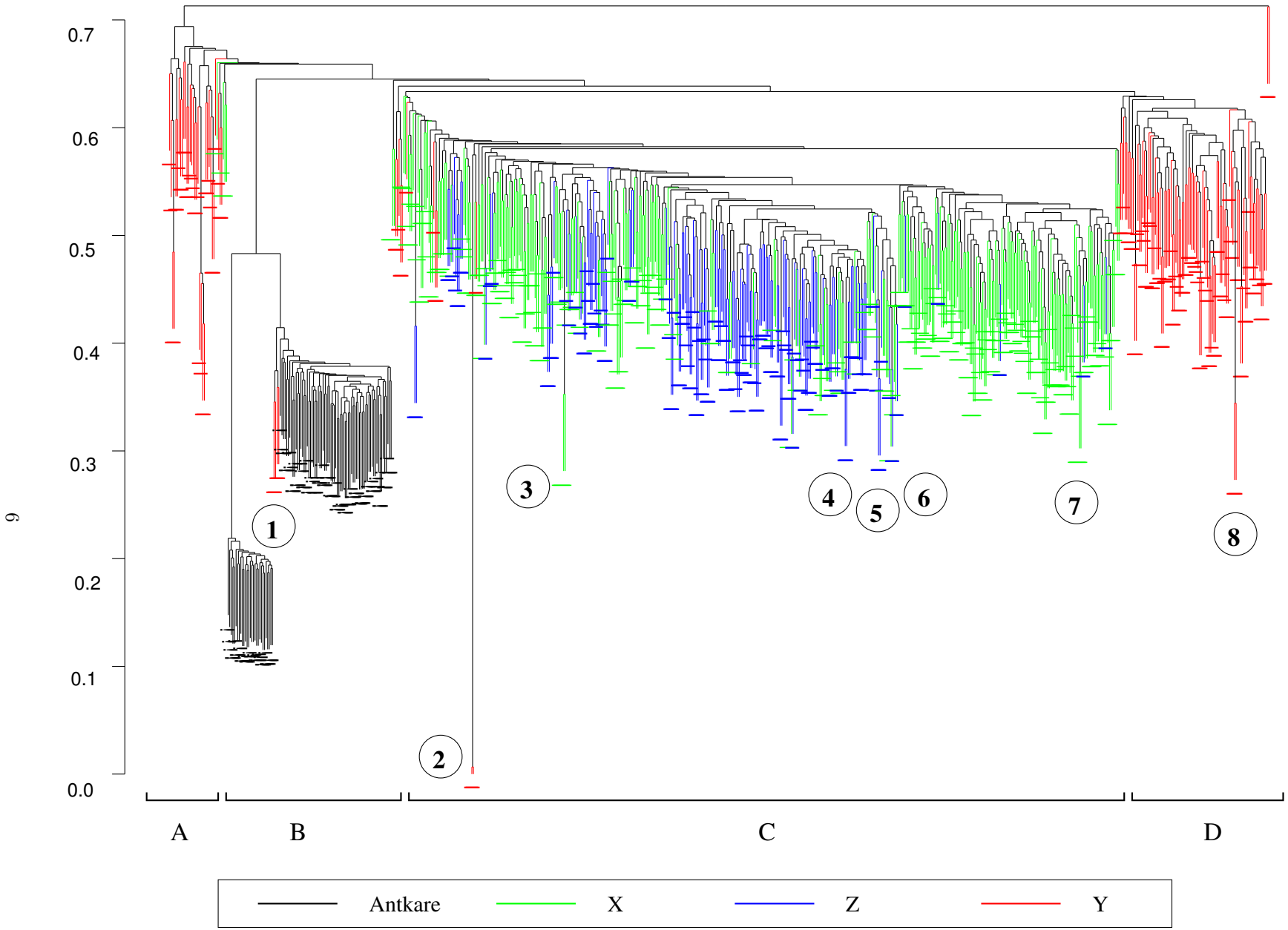

Figure 2: Dendrogram for automatic cluster analysis of corpora *Antkare* (black), *X* (green), *Z* (blue), *Y* (red). Main clusters : group A (corpus Y), group B (corpus Ike Antkare), group C (corpora Z et X), group D (corpus Y). Main remarkable points : (1) four Y texts are classified with Ike Antkare fake documents, (2), (3), (4), (5), (6), (7), (8) smallest distances (without SCIgen texts).

5 Other SCIgen texts in IEEE

The search engine http://ieeexplore.ieee.org/ offers a functionality ("More Like This") to research texts similar to a chosen paper. We applied it to 3 pseudo papers out of the four found in corpus Y. On the day of the experiment (April 22, 2011), this functionality returned 122 different documents. We call this new corpus More Like This "MLT" and we applied to it the same tools. To make this cluster analysis readable, the dendrogram, reproduced in Figure 3, relates only the comparison of this new corpus with the Antkare texts (to detect some new pseudo texts) and with those of Z (containing only genuine texts).

The corpus MLT includes:

- 82 new pseudo papers grouped with Ike Antkare documents (Group C Figure 3).

  C1 contains 17 texts very similar to those of Ike Antkare, but slightly "distorted" to pass the peer selection. Sometimes the titles are appropriate to the subject of the conference, some abstracts are more or less coherent, and few figures have been changed, but most of the writing remains SCIgen.

  C2 contains 65 twins from those of Ike Antkare–the texts generated by SCIgen were accepted, without any change, by the conference organizers.

  C3 and C4 : twice, identical pseudo publications were presented under a different title, by the same authors to two different conferences!

- 41 genuine papers are classified into two groups (A and B).

After verification, in these 41 texts at list one bibliography in one of these papers comes from SCIgen.

The smallest distances: groups A1, A2, A3, B1 and B2. A2 branches are three very similar papers by the same authors in three different conferences. A1, A3 and B2 are pairs of texts by the same authors on very related topics. B1 is a pair of papers by the same authors on the same topic at two different stages: first presented in a conference then enriched and published in a scientific journal.

A "nearest neighbor" classification (knn classification [4, 15] with k=1) was tested to verify the feasibility of automatic detection of pseudo papers. For this experiment, the 100 documents of the Ike Antkare corpus and the 121 articles of the corpus Z respectively represent the "fake" and "genuine" papers class. For each text of the corpus "More Like This" the distances to the 221 reference texts are computed and the text is assigned to the group of its nearest neighbor (pseudo or genuine).

Texts examination confirms that, using this method, all pseudo items (SCIgen texts and modified SCIgen texts) are classified with the corpus Antkare.

6 Conclusion

We selected three international conferences with scientific committees, 614 papers in all. These three conferences show stringent selection rates and two of them (X and Z) are considered among the most rigorous in their fields. In a second step, we added 122 texts selected by the IEEE as "close" to the 4 pseudo texts detected in Y.

This very limited trial gives:

- At least 86 SCIgen papers in IEEExplore.
- More than 20 different conferences have been "infected" between 2008 and 2011. For the two most affected there was respectively more than 20 and more than 10 fake papers published.

Acknowledgment

We thank the anonymous referees for theirs valuable comments on a previous version of this document but also Tom Merriam, Jacques Savoy, and Edward Arnold for their careful readings of previous versions of this paper, Eric Gaussier, Marie-Christine Rousset for useful discussions about the topic.

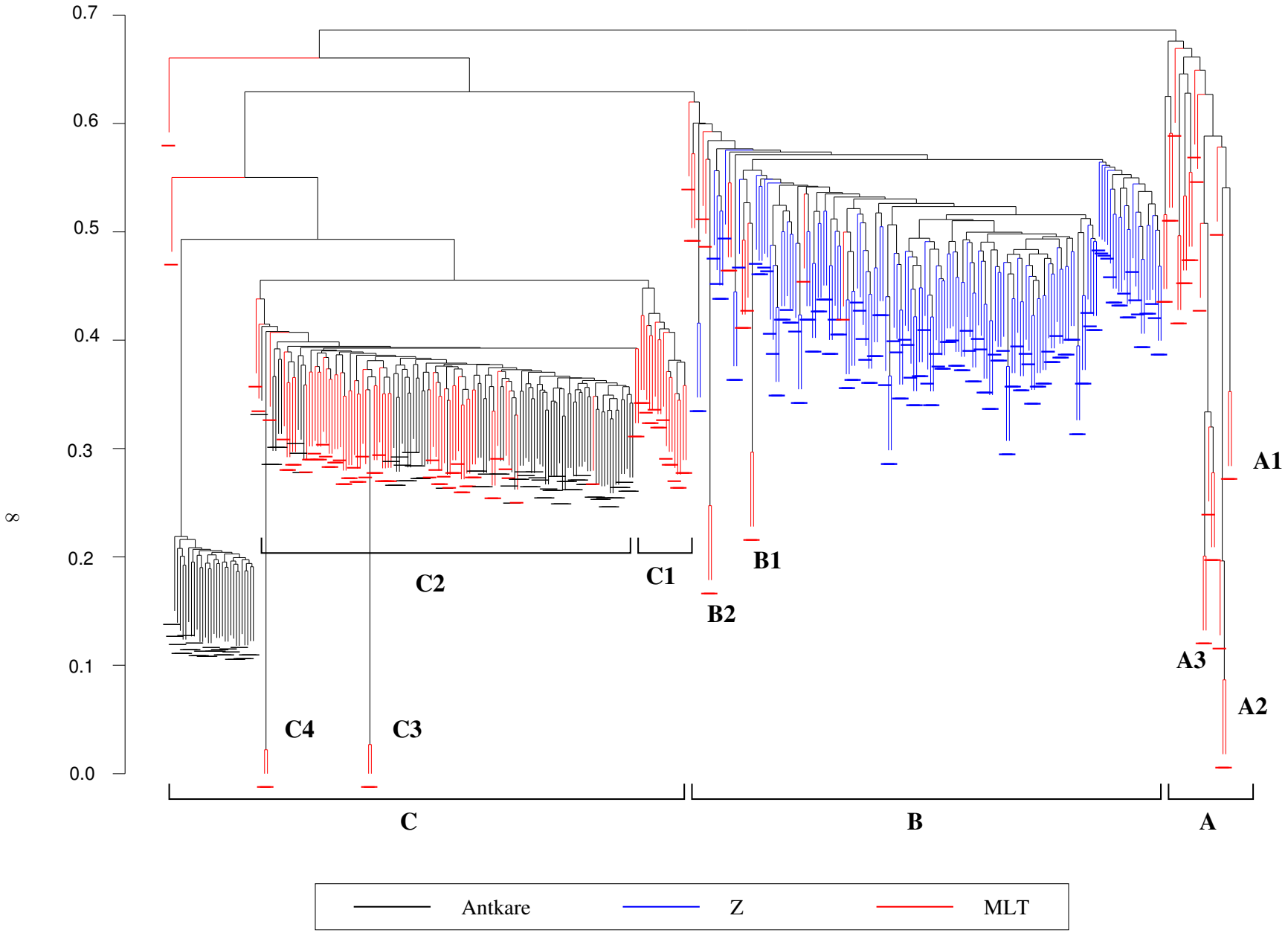

Figure 3: Dendrogram for cluster analysis of corpora *Antkare* (black), *Z* (blue), *MLT* (red).
Main clusters : C (Antkare and MLT pseudo texts), B (Z and MLT genuine), C(MLT genuine).
main remarkable points : C3 C4 (SCIgen texts published many times). A1, A2, A3, B1, B2 (smallest distances).

References

[1] P. Ball. Computer conference welcomes gobbledegook paper. *Nature*, 434, 946, 21 avril 2005.

[2] J. Beel and B. Gipp. Academic search engine spam and google scholar's resilience against it. *Journal of Electronic Publishing*, 13(3), 2010.

[3] J.-P. Benzecri. *L'analyse des données*. Dunod, 1980.

[4] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[5] J. Felsenstein. *Inferring Phylogenies*. Sunderland : Sinauer Ass., 2004.

[6] J. Felsenstein. *Package of Programs for Inferring Phylogenies (PHYLIP)*. Seattle : University of Washington, 2004.

[7] S. Hockey and J. Martin. *OCP Users' Manual. Oxford*. Oxford University Computing Service, 1988.

[8] C. Labbé. Ike antkare, one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter*, 6(2):48–52, June 2010.

[9] C. Labbé and D. Labbé. Inter-textual distance and authorship attribution corneille and moliere. *Journal of Quantitative Linguistics*, 8(3):213–231, 2001.

[10] C. Labbé and D. Labbé. La distance intertextuelle. *Corpus*, 2:95–118, 2003.

[11] C. Labbé and D. Labbé. Peut-on se fier aux arbres ? In *9e Journées internationales d'analyse statistique des données textuelles*, pages 635–645, mars 2008.

[12] C. Labbé and D. Labbé. 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.

[13] D. Labbé. Experiments on authorship attribution by intertextual distance in english. *Journal of Quantitative Linguistics*, 14(1):33–80, April 2007.

[14] X. Luong. *Méthodes d'analyse arborée. Algorithmes, applications*. PhD thesis, Université de Paris V, 1988.

[15] D. Meyer, K. Hornik, and I. Feinerer. Text mining infrastructure in r. 25(5):569–576, 2008.

[16] D. L. Parnas. Stop the numbers game. *Commun. ACM*, 50(11):19–21, November 2007.

[17] M. Roux. *Algorithmes de classification*. Masson, 1985.

[18] M. Roux. *Classification des données d'enquête*. Dunod, 1994.

[19] M. Ruhlman. Analyse arborée. représentation par la méthode des groupements. Technical report, CERAT, 2003.

[20] J. Savoy. Les résultats de google sont-ils biaisés ? *Le Temps*, 9 février 2006.

[21] P. Sneath and R. Sokal. *Numerical Taxonomy*. San Francisco : Freeman, 1973.