AJAS / SPAS

African Journal of Applied Statistics
Vol. 10 (2), 2023, pages 1503 - 1515.
DOI: https://dx.doi.org/10.16929/ajas/2023.1503.279

ISSN 2316-0861

A note on the statistical similarity of English Version Gospels

Soumaila Dembelé [1,3,*], **Mohamed Cheikh Haidara** [2,3]

[1] Faculté des Sciences Économiques et de Gestion (FSEG). Université des Sciences Sociales et de Gestion de Bamako (USSGB) Mali.
[2] Département des Techniques Quantitatives. Faculté des Sciences Juridiques et Économiques. Université de Nouakchott Al Aasriya (UNA) Mauritanie.
[3] Imhotep International Mathematical Center (imho-imc), https://imhotepsciences.org

Received November 30, 2023; Accepted on December 28, 2023; Published online on July 20, 2024

Abstract. Here, the English version Gospels (John, Luke, Mathew, Mark) are studied with respect to their similarity in order to make sense to the existence of a common version named *QUELLE* that supposedly influenced them. We adapt the existing results in the French versions to the English versions. We find equivalent results that confirmed our positive response, the existence of *QUELLE*. Moreover, our statistical results are also shown to be quick and efficient in the context of big data.

Key words: Jacquard similarity; Rajaraman-Ullman algorithm; shingling; bible; English Gospel; French Gospel.
AMS 2010 Mathematics Subject Classification Objects : 60-07;60C05; 60F15

[*] Corresponding author: Soumaila Dembelé (dembele.soumaila@ugb.edu.sn, soumailadembeleussgb@gmail.com)
Mohamed Cheikh Haidara: cheikh.haidara@fsje.e-una.mr, mcheikhhaidara@gmail.com, chheikhh@yahoo.fr

Résumé (Abstract in French) Ici, les Évangiles en version anglaise (Jean, Luc, Matthieu, Marc) sont étudiés en fonction de leur similarité afin de donner un sens à l'existence d'une version commune nommée *QUELLE* qui les aurait influencées. Nous adaptons les résultats existants des versions françaises aux versions anglaises. Nous trouvons des résultats équivalents qui confirment notre réponse positive à l'existence de *QUELLE*. De plus, nos résultats statistiques se révèlent rapides et efficaces dans le contexte du big data.

Presentation of authors.

Soumaila Dembele, PhD, is professor of Statistics at *Université des Sciences Sociales et de Gestion de Bamako* and a regular researcher at the *Imhotep International Mathematical Center (imho-imc), https://imhotepsciences.org*.

Mohamed Cheikh Haidara, PhD, is professor of Statistics at *Université de Nouackchott* Al Aasriya and a regular researcher at the *Imhotep International Mathematical Center (imho-imc), https://imhotepsciences.org*.

1. Introduction

1.1. The aim of the paper

The study of similarity on canonical Gospels in French was conducted in Dembélé and Lo (2015), as an application to a broader study of the similarity of big and massive texts transformed in collection of *minhashings*, where a modification of the Rajaraman and Ullman (2010)'s algorithm is used. In Dembélé and Lo (2017), we can find the theoretical background beneath the method that uses the characterization of the probability law of the exact similarity between texts and the approximations of the similarity index that results from the Rajaraman and Ullman (2010)'s method. The results obtained by comparing the four gospels in French gave stunning results with approximated similarity around 50%.

However, English version Gospels are more read and used by far at a worldwide stage. So it is logical to have the same study for them, what we will do in this paper. Since all the methods are already fully detailed in Dembélé and Lo (2015) and Dembélé and Lo (2017), we only need to have simple recalls and to explain how works the algorithms. The most important part of this note are stating the results and the comments on them for comparing gospels in both languages.

We still need to say a few words on the general topic of similarity measures, a topic which has been studied and is still studied by researchers from a variety of disciplines: (see e.g. Stein and Essen (2006), Gionis *et al.* (1999), Chen (2003) for visual similarity, Cha(2007) for the use of density functions in similarity detection, Gower and Legendre (1986) and Zezula *et al.* (2006) for the metric space

approach, Strehl *et al.* (2000), Formica (2005) in the context of information sciences, Bilenko and Mooney (2003), Theobald *et al.* (2008) for focus similarity on large-web collections and (see e.g. de França (2014)) on the Iterative Universal Hash Function Generator for Minhashing, and the resemblance and containment of documents (see e.g. Broder(1997)). Now, we are going to have a simple introduction to the topic of similarity measurement.

1.2. Jacard similarity and approximation

For better explain the notion of similarity, let us consider two sets S_1 and S_2, whose total cardinality is n. The Jaccard similarity between S_1 and S_2 is defined by:

$$p=\frac{\#(S_1 \cap S_2)}{\#(S_1 \cup S_2)} =: sim(S_1, S_2). \tag{1}$$

We can also compute the similarity between m subsets of S: $S_1, ..., S_m$. These sets can be represented as in Table 1, that we will call the *representation matrix* of $S_1, ..., S_m$. The similarity between two columns S_h and S_ℓ is given by :

$$sim(S_h, S_\ell) = \frac{\#\{i, 1 \leq i \leq n, S_{ih} = S_{i\ell} = 1\}}{\#\{i, 1 \leq i \leq n, (S_{ih} + S_{i\ell} = 1) + (S_{ih} = S_{i\ell} = 1)\}}. \tag{2}$$

Elements	S_1	S_2	...	S_h	...	S_l	...	S_m
1	1	0	...	0	...	1	...	1
2	0	0	...	1	...	0	...	0
...	0	...	...	...	...	...	...	...
i	1	0	...	1	...	1	...	1
...	...	...	...	...	...	...	...	...
n	0	0	...	0	...	0	...	1

Table 1. Representation matrix of $S_1, \ldots, S_m$

For comparing texts, a common method is to eliminate are common words (that are repeaded in all texts, as conjunctions *and*, *or*, articles *the*, *a*, $\cdots$) and, for $k \in \{3, \cdots, 6\}$ for examples, to transform the texts into all successive subtexts of k-letters, called k-*shinglings*. Interesting discussions on how to fix the value of k are given in Dembélé and Lo (2015).

The idea of the Rajaraman and Ullman (2010)'s algorithm is the following. For lengthy texts, Glivenko-Cantelli's theorem of Statistics, says that we may draw at random n_1 elements of S_1 and n_2 from S_2 and get good approximations of the similarity index as described in tables like 1. However, this would be quite easy in the theoretical frame. However, we study similarity indexes in the context of big data, especially in online environments. We have the following limitations:

1. the cardinalities of sets S_1 and S_2 are big and this make the computations lengthy ;
2. the number of permutations of $S_1 \cup S_2$ can be extremely big and, in the course of the computations, we have two options: **either** having all data in the access memory of a computer may lead to overloading the system and make it stuck **or** having the data in hard discs and opening and closing them at each reading leads to a great consumption of time.

Let us explain the different ways of computing the exact similarity index.

(1) Exact similarity: the RAM method. We charge the sets S_1 and S_2 in the Random Access Memory (RAM). Then we compute the similarity index.

(2) Exact similarity: the FILE method. The two files A and B are in hard drives. To avoid to have both files in the *RAM*, we set a method that requires to open one file only all along the procedure. We open File A and read the first line L_j, $j = 1$, that we uploaded in the RAM and close it. Next, we open the second file and compare all the lines of File B. We add up all the cardinalities of the intersections between L_j and the line of File B and keep that sum in the RAM and close File B. At that time none the files is open. We iterate the same procedure with L_2 of File A. At the end, we get the cardinality of the whole intersection between texts A and B and then exact similarity index. Obviously, we do not charge the *RAM* too much. But as said before repeating the openings and closings lead to consuming time.

Now, we may use approximated methods described as follows.

(3) The Glinenko-Cantelli Method. We open each file separately and draw random numbers n_1 and n_2 of line in the *RAM*. Next, the similarity between these sets of lines from both files is taken as the approximated index. From the Glivenko-Cantelli or simply the Strong Law of Large number, the average of the computed index after a number of BB samples is close to the true similarity index. In practice, we found that setting $BB = 50$ gives good results.

(4) The RUM Algorithm. That method reduces the full signature matrix by another with a small number k of lines. The method is based on k minhashing functions of the form

$$i \mapsto h_\alpha(i) = A_\alpha \times i + B_\alpha, \quad 1 \leq \alpha k.$$

The set $\{1, \cdots, n\}$, where n is the cardinality of the union of all sets S_j. The *RU* algorithm replaces the signature matrix to a matrix like in Table 3 in Dembélé and Lo (2017), page 1202. The way of forming this reduce signature matrix is explained in the same cited paper as well in Dembélé and Lo (2015).

It is amazing that we could set k as low as 5, 10, ..., etc, and compute the similarity index between two columns $t(S_h)$ and $t(S_\ell)$ as a good approximation of the similarity

index between S_h and S_k. The *RU* algorithm has been modified in the **RUM** for the following reason. We depart of a signature matrix where we list all elements of the two compared sets. So the intersection is put twice. This allows to go faster. At the index we used the formula between the similarity indexes *simRU* and *simRUM* from the two algorithms to find *simRU*.

The mathematical background and the justification of the *RUM* has been properly done in Dembélé and Lo (2017), while those four methods were implemented and applied to French versions of Gospels in Dembélé and Lo (2015). The main conclusions were:

(1) RAM method. In the example of the four Gospels in French, the number letters are from 55 000 to 109 000. For such sizes, it is possible to upload the files in the *RAM* and the execution time is between 500s (8.5mn) and 1430s (23.88mn). We can imagine the big time to wait for far bigger files.

(2) FILE method. In the same context, the execution times run from 24.24mn to 51.33mn. So trying to save the *RAM* obviously increase time consumption.

(3) Glivenko-Cantelli method. By taking samples of one thousand (10000) 3-*Shinglings* and $BB = 50$ resamplings. We get sound approximations of the index after 27 to 31s. By using confidence interval based on the Glivenko-Cantelli rate of convergence, we have discrepancies of the approximation with respect to the true index ranging from 14 to 16%.
(4) RUM algorithm. We have similar execution times. By using confidence interval based on the Bernouilli law, we have discrepancies of the approximation with respect to the true index ranging from 13.20 to 14.20% for the use of 15 *minhasing* functions and from 8.90 to 10.7% for the use of 20 *minhasing* functions.
In conclusion, the *RUM algorithm* is a very good method, better than the Glivenko method, given good approximations of the similarity index required low time consumptions around 30s.

Here, we wish to confirm the results on the study on the similarity of French version of Gospels when we shift to English versions.

2. English versions of Gospels

We are going to use the same theoretical frame that we described above and the same computer package. All the tables are given in the Appendix from page 1510.

2.1. global results

The results for the similarity study of French version Gospels are very similar when we shift to English version Gospels. This is a new argument of the existence of the common core of the Gospel even when translated. However, when the Matthew Gospel is concerned, we see a bigger discrepancy in the indexes in

French and English versions. We see this in Tables 3-6.

In reading these tables, the following the abbreviations used.

(1) *Time* : is the time required to complete the computation of the similarity.

(2) *Kc*: is the number of k-shinglings of the concerned Gospel in Tables 3 and 4.

(3) *deviation* : is the difference in execution time between an estimation method and the direct method (*direct RAM method*).

3. Comparison by pairs

The conclusion we already drew in the above subsection is more illustrated in Tables 7-12. The most significant thing that is striking is that the worse comparison concerns Mathew and Mark Gospels and at the same time Mathew Gospel shows the biggest drift in the comparison when passing from French to English.

4. Conclusion

The English version Gospels (John, Luke, Mathew, Mark) have been studied with respect to their similarity in order to make sense to the existence of a common version named *QUELLE* that supposedly influenced them. We adapted the existing results in the French versions to the English versions. We find equivalent results that confirmed our positive response the existence of *QUELLE*. Moreover, our statistical results are also shown to be quick and efficient in the context of big data.

Acknowledgment. First we thank our mentor Professor Gane Samb Lo who strongly assisted us in this work. Next we thank the anonymous referee, the associated editor and the members of the editorial board of the journal who helped to significantly improve the paper.

References

Rajaraman and Ullman J.U. (2011). *Mining of Massive Datasets*. California. http://mmds.org/ (visited on August 12, 2024)

Leskovec J., Rajaraman A. and Ullman J.D. (2011). *Mining of Massive Datasets*. California. http://mmds.org/ (visited on August 12, 2024)

Stein. B. and zu Eissen S. M.(2006). Near Similarity Search and Plagiarism Analysis. In: *Spiliopoulou et al. (Eds.): From Data and Information Analysis to Knowledge Engineering Selected Papers from the 29th Annual Conference of the German Classification Society (GfKl) Magdeburg*, pp. 430-437, Springer.

Gionis A., Indyk P. and Motwani R.(1999). Similarity Search in High Dimensions via Hashing. In: *Proceedings of the 25th VLDB Conference*, pp. 518-529 , Eds: Edinburgh, Scotland.

Chen D.Y, Xiao-Pei T., Yu-Te S. and Ouhyoung M.(2003). On Visual Similarity Based 3D Model Retrieval, *Eurographics 2003, P. Brunet and D. Fellner (Guest Editors)*, 22(3), pp. 223-232.

Cha S.H. (2007) Comprehensive Survey on Distance Similarity Measures between Probability Density Functions. *International journal of mathematical models and methods in applied sciences*, 4(1), pp. 300-307.

Gower J.C and Legendre, P.(1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3, pp. 5-48.

ZEzula P., Dohnal V. and Amato G.(2006) *Similarity Search The Metric Space Approach*, Springer.

Strehl A., Ghosh J. and Mooney R.(2000) Impact of Similarity Measures on Web-page Clustering. *American Association for Artificial Intelligence*, pp. 78712-1084.

Formica A.(2005) Ontology-based concept similarity in Formal Concept Analysis, *Information sciences*, pp. 2624-2641.

Bilenko M. and Mooney R.J.(2003). Adaptive Duplicate Detection Using Learnable String Similarity Measures. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD)*, pp.39-48, Washington DC.

Theobold M., Siddharth J. and Paepcke A.(2008) SpotSigs: robust and efficient near duplicate detection in large web collections. In: *31st Annual ACM SIGIR 08 Conference*, Singapore.

de França, F.O.(2014). Iterative Universal Hash Function Generator for Minhashing. arXiv:1401.6124.

Broder, A.(1997) On the resemblance and containment of documents. In : *Compression and Complexity of Sequences, Proceedings*, pp. 21-29.

Guyon, I. and Elisseeff, A.(2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, Vol. 3, 1157-1182.

Guyon, I. (2006) *Feature extraction: foundations and applications*, Vol. 207, Springer.

Lawrence, N.D. (2008) *Tutorial, Dimensionality reduction the probabilistic way*, ICML Tutorial.

Dembélé S. and Lo, G.S. (2015) Probabilistic, statistical and algorithmic aspects of the similarity of texts and application to Gospels comparison. *Journal of Data Analysis and Information Processing*, Vol. 3, 112-127.

Dembélé S. and Lo, G.S. (2017) The exact probability law for the approximated similarity from the Minhashing method. *Afrika Statistika*, Vol. 12, 1199-1218.

Appendix.

(A) Numbers of letters by Gospel.

	John	Luke	Mark	Matthew
NR-E	1916	2535	1518	2457
NR-F	2534	3442	628	1319
NLBE-E	98 179	133 360	77 939	123 112
NLBE-F	96 269	129 458	76 543	149 747
NLAE-E	74 692	101 464	77 939	123 112
NLAE-F	69 316	94 766	55 555	108 555

Table 2. Total numbers of rows and letters of the synoptic Gospels English and French. Abbreviations.NR: number of rows, NLBE: number of letters before editing; NLAE: number of letters after editing; -E: English, -F:French

(B) Exact similarity index.

	Luke	Mark	Matthew
John	Sim= 55.9 % Kc= 63162 Time= 642 s	Sim= 56.81% Kc= 48717 Time= 508 s	Sim= 57.53 % Kc= 61772 Time= 753 s
Luke		Sim=53.95 % Kc= 56504 Time= 579 s	Sim=72.74 % Kc= 82498 Time= 792 s
Mark			Sim=57.28 % kc= 56173 Time= 386 s

Table 3. Computation durations of the exact Jaccard similarity beetwen two Gospels by the **RAM** method

	Luke	Mark	Matthew
John	Sim= 55.9 % Kc= 63162 Time= 2853 s	Sim= 56.81% Kc= 48717 Time= 1674 s	Sim= 57.53 % Kc= 61772 Time= 2398 s
Luke		Sim=53.95 % Kc= 56504 Time= 1938 s	Sim=72.74 % Kc= 82498 Time= 2673 s
Mark			Sim=57,28 % kc= 56173 Time= 1453 s

Table 4. Computation durations of the exact Jaccard similarity beetwen two Gospels by **FILE** method

	Luke	Mark	Matthew
John	Sim= 48.08 % Time= 29 s	Sim= 47.51 % Time= 28 s	Sim= 48.17 % Time= 29 s
Luke		Sim=53.18 % Time= 28 s	Sim=54.85 % Time= 27 s
Mark			Sim=52.95 % Time= 29 s

Table 5. Glivenko-Cantelli's method

John	Luke Sim= 59.9 % deviation= 11.06	Mark Sim= 61.3 % deviation= 10.13	Matthew Sim= 58.9 % deviation= 10.30 s
	Time= 26 s	Time= 26 s	Time= 26 s
Luke		Sim= 62.3 % deviation= 10.91	Sim= 57.6 % deviation= 10.64
		Time= 27 s	Time= 25 s
Mark			Sim= 63.4 % deviation= 10.97
			Time= 27 s

Table 6. Outputs by **RUM**'s algorithm for pp = 20 minhashing functions

(C) Comparison of similarity indexes in French and English versions.

	John-Luke in English	John-Luke in French
Exact similarity by direct method	Sim= 55.9 % Time= 642 s	Sim= 57.62 % Time= 755 s
Exact similarity by file method	Sim= 55.9 % Time= 2853 s	Sim= 57.62 % Time= 2312 s
Approximated similarity by GC	Sim= 48.08 % Time = 29 s	Sim= 47.50 % Time= 20 s
Approximated similarity by RUM	Sim= 66 % deviation= 17.08 % Time= 17 s	Sim=60 % deviation= 20.76 % Time= 26 s

Table 7. Comparison between the John and Luke Gospels both in French and English versions

	John-Mark in English	John-Mark in French
Exact similarity by direct method	Sim= 56.81 % Time= 508 s	Sim= 57.53 % Time= 816 s
Exact similarity by file method	Sim= 56.81 % Time= 1674 s	Sim= 57.53 % Time= 1376 s
Approximated similarity by GC	Sim= 47.51 % Time = 28 s	Sim= 46.46 % Time= 34 s
Approximated similarity by RUM	Sim= 65.6 % deviation= 23.17 % Time= 18 s	Sim=58 % deviation= 22.1 % Time= 25 s

Table 8. Comparison between the John and Mark Gospels both in French and English versions

	John-Matthew in English	John-Matthew in French
Exact similarity by direct method	Sim= 57.53 % Time= 753 s	Sim= 51 % Time= 510 s
Exact similarity by file method	Sim= 57.53 % Time= 2398 s	Sim= 51 % Time= 3021 s
Approximated similarity by GC	Sim= 48.17 % Time = 29 s	Sim= 46.04 % Time= 29 s
Approximated similarity by RUM	Sim= 64.4 % deviation= 20.50 % Time= 18 s	Sim=59.2 % deviation= 20.38 % Time= 22 s

Table 9. Comparison between the John and Matthew Gospels both in French and English versions

	Luke-Mark in English	Luke-Mark in French
Exact similarity by direct method	Sim= 53.95 % Time= 579 s	Sim= 54.12 % Time= 640 s
Exact similarity by file method	Sim= 53.95 % Time= 1938 s	Sim= 54.12 % Time= 3080 s
Approximated similarity by GC	Sim= 53.18 % Time = 28 s	Sim= 50.79 % Time= 19 s
Approximated similarity by RUM	Sim= 61.2 % deviation= 22.05 % Time= 18 s	Sim= 56.4 % deviation= 19.87 % Time= 22 s

Table 10. Comparison between the Luke and Mark Gospels both in French and English versions

	Luke-Matthew in English	Luke-Matthew in French
Exact similarity by direct method	Sim= 72.74 % Time= 792 s	Sim= 69.55 % Time= 1430 s
Exact similarity by file method	Sim= 72.74 % Time= 2673 s	Sim= 69.55 % Time= 1457 s
Approximated similarity by GC	Sim= 54.85 % Time = 27 s	Sim= 50.26 % Time= 22 s
Approximated similarity by RUM	Sim= 68.4 % deviation= 21.57 % Time= 21 s	Sim=63.2 % deviation= 22.03 % Time= 22 s

Table 11. Comparison between the Luke and Matthew Gospels both in French and English versions

	Mark-Matthew in English	Mark-Matthew in French
Exact similarity by direct method	Sim= 57.28 % Time= 386 s	Sim= 48.74 % Time= 508 s
Exact similarity by file method	Sim= 57.28 % Time= 1453 s	Sim= 48.74 % Time= 1552 s
Approximated similarity by GC	Sim= 52.95 % Time = 29 s	Sim= 52.28 % Time= 27 s
Approximated similarity by RUM	Sim= 62.4 % deviation= 21.29 % Time= 18 s	Sim= 59.2 % deviation= 22.38 % Time= 27 s

Table 12. Comparison between the Mark and Matthew Gospels both in French and English versions

Important updates.

1. We used the algorithm in Rajaraman and Ullman (2010), authored by Anand Rajaraman, Jeffrey D. Ullman, (Subsubsection 3.3.4, page 65). This explains why we refer to it as Rajaraman-Ullman algorithm (RU). Meanwhile, an other version of the book including Jure Leskovec as a third author (Leskovec et al. (2011)) has been released. The same algorithm is given in page 83. So, we still keep the acronym (RU) for the algorithm.

2. In Dembélé and Lo (2015), the authors have forgetten to defined *kc*, found in the tables, as the number of shinglings for each gospel. As well the used *Ecart* rather than **deviation** to mean the discrepancy between the exact similarity and the approximated similarity.