Optimal clustering method in ultrametric spaces

Said Fouchal
University of Paris 8, CHArt & Laisc
41, rue Gay Lussac
75005 Paris
email : said.fouchal@laisc.net

Murat Ahat
EPHE, Laisc
email : murat.ahat@laisc.net

Ivan Lavallée
University of Paris 8, CHArt & Laisc & CNRS
UMI ESS 311010 UCAD Dakar BP 5005
email : ivan.lavallee@gmail.com

***Résumé*—We propose in this paper a novel clustering algorithm in *ultrametric spaces*. It has a computational cost of *O(n)*. This method is based on the *ultratriangle inequality* property. Using the order of ultrametric space we demonstrate that we can deduce the proximities between all data in this space with just a few informations. We present an example of our results and show the efficiency and the consistency of our algorithm compared with another.**

Keywords-clustering ; ultrametric ; complexity

I. INTRODUCTION

The goal of clustering is to organize objects into groups whose members are similar according, most often, to some proximity criteria defined by introducing distances [19].
There are several approaches of clustering, hierarchical, partitioning, density-based, ... , which are used in a large variety of fields, such as astronomy, physics, medicine, biology, archaeology, geology, geography, psychology, and marketing [18].

The clustering aims to group objects of a data set into a set of meaningful subclasses, it can be used as a stand-alone tool to get insight into the distribution of data [2] [18].
The clustering of high-dimensional data is an open problem encountered by clustering algorithms in different areas. Since the computational cost increases with the size of the data set, the feasability can not be fully guaranteed.
We propose in this paper a fast clustering in ultra-metric spaces. It aims to show rapidly the inner structure of the data set by providing a general view of proximities in the data set. The computational complexity of our algorithm is in order of $O(n)$ (n is the size of data). Thus, it guarantees the clustering of high-dimensional data in ultrametric space.

We can find the ulrametric spaces in many kinds of data sets : genealogy, library, information and social sciences data, to name a few.
This paper is organized as the following : In section 2 we present a brief overview of the clustering strategies. In section 3, we introduce the notions of metric and ultra-metric spaces, distance and balls. Our approach (contribution) is presented in section 4. We present an example in section 5. Finally, in section 6 we give our conclusion and future work.

II. RELATED WORK

Clustering strategies can be widely classified into the following :
Hierarchical clustering is either agglomerative ('bottom-up") or divisive ('top-down"). The agglomerative approach starts with each element as a cluster and merges them successively until forming a unique cluster (i.e. the whole set) (e.g. *WPGMA* [8], *UPGMA*). The divisive begins with the whole set and divides it iteratively until it reaches the elementary data. The outcome of hierarchical clustering is generally a dendrogram which is difficult to interpret when the data set size exceeds a few hundred of elements. The complexity of these clustering algorithms is at least $O(n^2)$ [20].
Partitional clustering creates clusters by dividing the whole set into k subsets. It can also be used as divisive algorithm in hierarchical clustering. Among the typical partitional algorithms we can name K-means and its variants K-medoids, PAM, CLARA and CLARANS. The results depend on the k selected data in this kind of algorithms. Since the number of clusters is defined upstream of the clustering, the clusters can be empty.
Density-based clustering , where the clusters are regarded as a dense regions leading to the elimination of the noise. DBSCAN, OPTICS and DENCLUE are typical algorithms based on this strategy [2], [3], [6], [18].
The major clustering algorithms calculate similarities between all data before using the *adapted* clustering algorithm (for all types of similarity measure used). Consequently, the computational complexity is increased to $O(n^2)$ before the execution of the clustering algorithm. Our approach consist in avoiding these mutual calculations reduce the global computational cost. Thus, we propose an $O(n)$ clustering algorithm for the *ultrametric spaces*.

III. DEFINITIONS

Definition 1: A metric space is a set endowed whith distance between its elements. It is a particular case of a *topological space*.

Definition 2: We call a distance on a given set E, an application $d : E \times E \to \mathbb{R}^+$ which has the following properties for all $x, y, z \in E$:

1) (Symmetry) $d(x,y) = d(y,x)$,

2) (Positive Definiteness) $d(x,y) \geqslant 0$, and $d(x,y) = 0$ if and only if $x = y$,
3) (Triangle Inequality) $d(x,z) \leqslant d(x,y) + d(y,z)$.

Example 1: The most familiar metric space is the *Euclidean* space of dimension n, which we will denote by $\mathbb{R}^n$, with the standard formula for the distance :
$d((x_1,\cdots,x_n),(y_1,\cdots,y_n)) = ((x_1-y_1)^2+\cdots+(x_n-y_n)^2)^{\frac{1}{2}}$ (1).

Definition 3: Let (E,d) be a metric space. If the metric d satisfies the strong triangle inequality :

$$\forall\ x,y,z \in E,\ d(x,y) \leqslant max\{d(x,z),d(z,y)\}$$

then It is called ultrametric on E. The couple (E,d) is an ultrametric space [10], [11], [22].

Definition 4: We name open ball centered on $a \in E$ and has a radius $r \in \mathbb{R}^+$ the set $\{x \in E : d(x,a) < r\} \subset E$, it is called $B_r(a)$ or $B(a,r)$.

Definition 5: We name closed ball centered on $a \in E$ and has a radius $r \in \mathbb{R}^+$ the set $\{x \in E : d(x,a) \leqslant r\} \subset E$, it is called $B_f(a,r)$.

Remark 1: Let d be an ultrametric on E. The closed ball on $a \in E$ with a radius $r > 0$ is the set : $B(a,r)=\{x \in E : d(x,a) \leqslant r\}$

Proposition 1: Let d be an ultrametric on E, the following properties are true [10] :

1) If $a,b \in E, r > 0$, and $b \in B(a,r)$, then $B(a,r) = B(b,r)$;
2) If $a,b \in E, 0 < i \leqslant r$, then either $B(a,r) \cap B(b,i) = \emptyset$ or $B(b,i) \subseteq B(a,r)$. This is not true for every metric space ;
3) Every ball is clopen (closed and open) in the topology defined by d (*i.e* every ultrametrizable topology is zero-dimensional). Thus, the parts are disconnected in this topology. Hence, the space defined by d is homeomorphic to a subspace of countable product of discrete spaces (*c.f* Remark 1) (see the proof in [10]).

Remark 2: A topological space is ultrametrizable if and only if it is homeomorphic to a subspace of countable product of discrete spaces [10].

Definition 6: Let E be a finite set, endowed with a distance d. E is classifiable for d if : $\forall \alpha \in \mathbb{R}^+$ the relation on E :

$$\forall x,y \in E,\ xR_\alpha y \Leftrightarrow d(x,y) \leqslant \alpha$$

is an equivalent relation. Thus, we can provide a partition from E as [23] :

– $d(x,y) \leqslant \alpha \Leftrightarrow x$ and y belong to the same cluster, or,
– $d(x,y) > \alpha \Leftrightarrow x$ and y belong to two distinct clusters.

Example 2: x, y and z are three points of plan endowed with an Euclidean distance d, we have :

$$d(x,y)= 2,\ d(y,z)=3,\ d(x,z)=4.$$

The set E is not classifiable for $\alpha= 3$. The classification leads to inconsistency.

Proposition 2: A finite set E endowed with a distance d is classifiable if and only if d is an ultrametric distance on E [10].

Proposition 3: An ultrametric distance generates an order in a set, viewed three to three they form isoscel triangles. Thus, the representation of all data is fixed whatever the angle of view is (see Fig. 1).

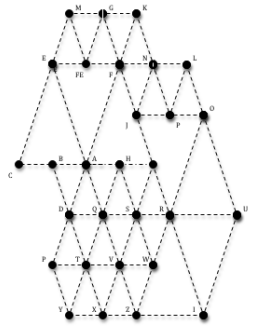


FIGURE 1. Pseudo sutucture of an ultramertic space

Proof: Let (E, d) an ultrametric set, for all x, y and $z \in E$: Consider : $d(x,y) \leqslant d(x,z)$ ....(1)
$d(x,y) \leqslant d(y,z)$ .....(2)
(1) and (2) $\Rightarrow$ (3) and (4)
$d(x,z) \leqslant max\{d(x,y),d(y,z)\} \Rightarrow d(x,z) \leqslant d(y,z)$ ....(3)
$d(y,z) \leqslant max\{d(x,y),d(x,z)\} \Rightarrow d(y,z) \leqslant d(x,z)$ ....(4)
(3) and (4) $\Rightarrow d(x,z) = d(y,z)$. ■

IV. PROPOSED APPROACH

We propose here a new approach of clustering, in ultrametric spaces, which has a computational cost of $O(n)$, thus it makes it *treatable* to cluster very large databases. Our method provides an insight of proximities between all data.
The idea consists in using the ultrametric space properties, in particular, the order induced by the *Ultratriangule Inequality*. Since the structure of an ultrametric space is frozen (*c.f* proposition 3), we do not need to calculate all mutual distances. The calculation of distances compared just to one element are sufficient to determine the clusters, as clopen balls. Consequently, we avoid the computation of $O(n^2)$. However, our objective is to propose a solution to the problem

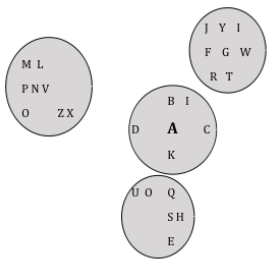


FIGURE 2. Distribution of data according to *A*

of computational cost, notably the feasability of clustering algorithms, in large data base.

A. Algorithm

Hypothesis 1: Consider E a set of data endowed with an ultrametric distance d.

Our method is composed of the 3 following steps :

Step 1: Choose randomly a data form E (see Fig 2) ;

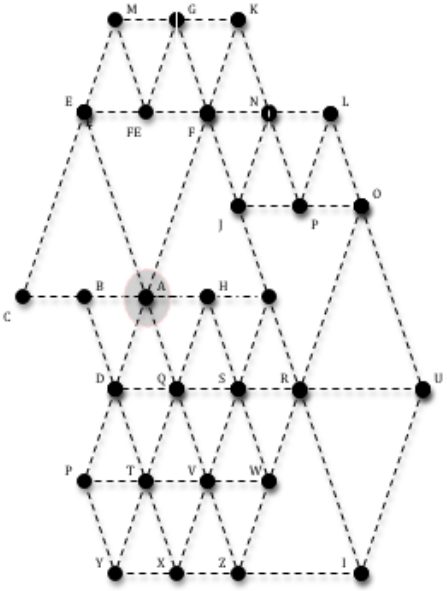


FIGURE 3. Step1 : Choosing randomly one element A

Step 2: Calculate distances between the chosen data and all others.

Step 3: Define thresholds and represent the distribution of data "*according*" to the chosen one, the thresholds and the calculated distances (see. Fig 4).

Proposition 4: The clusters are distinguished by the closed balls (or thresholds), around the chosen data.

Proof: An ultrametric space is a classifiable set (cf. Proposition 2). Thus, comparing mutual distances with a given threshold shows the belonging or not to the same cluster. Then, the distribution of data around any data in the set reflects entirely their proximity. ■

Proposition 5: The random choice of initial point, does not affect the resulting clusters.

Proof: Since an ultrametric distance generates an order in a set, the representation of all data is fixed whatever the angle of view. Hence, for every chosen data the same isosceles triangle are formed. ■

Proposition 6: The computational cost of this algorithm is equal to $O(n)$.

Proof: The algorithm consists in calculating distances between the chosen data and the $n-1$ data. ■

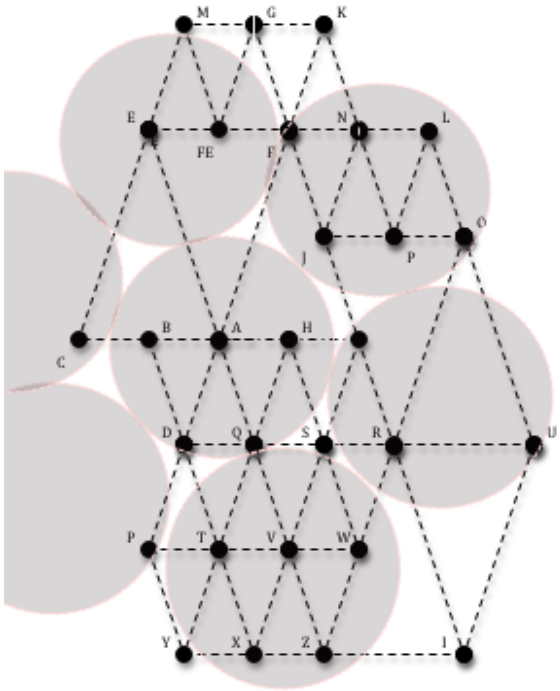


FIGURE 4. Representation of all elements around A, according to the calculated distances

V. EXAMPLE

We present in this section a comparison of our algorithm with $WPGMA$. The Weighted Paired Group Method using Averages ($WPGMA$) is a hierarchical clustering algorithm developped by McQuitty in 1966, it has a computational complexity of $O(n^2)$ [8] [12].

We have tested the two methods on sets of 34 and 100 data respectively. Let us consider the ultrametric distance matrix of 34 data (see Fig. 5), calculated from similarity matrix based on the *Normalizd Information Distance NID* [5] [8]. The resulting

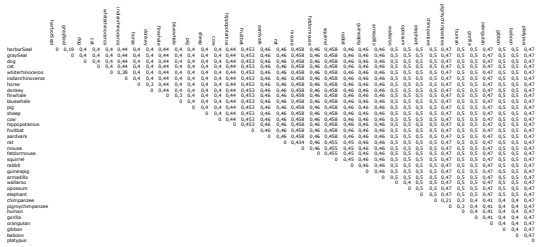


FIGURE 5. Ultrametric distance matrix of 34 data (see Appendix for clarity)

dendrogram with *WPGMA*, on 34 data, is shown in Fig. 6.

The results with our algorithm (TABLE I and TABLEII),

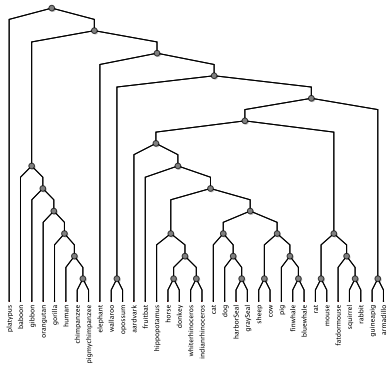


FIGURE 6. Clustering of the 34 data with WPGMA

using two differents data as starting points, show that the proximities in the data set are similar to those of the *WPGMA* clustering (Fig. 6).

TABLE I
RESULTS USING *finwhale* AS THE CHOSEN ORIGIN

Thresholds	Clusters
0.19 – 0.2	
0.2 – 0.29	
0.29 – 0.32	bluewhale
0.32 – 0.35	
0.35 – 0.39	
0.39 – 0.41	
0.41 – 0.42	
0.42 – 0.43	
0.43 – 0.44	pig sheep cow
0.44 – 0.443	harborSeal graySeal dog cat whiterhinoceros indianrhinoceros horse donkey
0.443 – 0.445	hippopotamus
0.445 – 0.45	
0.45 – 0.453	
0.453 – 0.454	fruitbat
0.454 – 0.457	aardvark
0.457 – 0.46	rat mouse fatdormouse squirrel rabbit
0.46 – 0.463	
0.463 – 0.466	guineapig armadillo
0.466 – 0.469	wallaroo opossum elephant
0.469 – 0.47	chimpanzee pigmychimpanzee human gorilla orangutan gibbon baboon
0.47 – 0.9	platypus

The figure 7 summarizes the results of the two methods on the same dendrogram, it shows that the generated clusters with our algorithm (balls) are similar to those of WPGMA.

TABLE II
RESULTS USING *horse* AS THE CHOSEN ORIGIN

Thresholds	Clusters
0.19 – 0.2	
0.2 – 0.29	
0.29 – 0.32	donkey
0.32 – 0.35	
0.35 – 0.39	
0.39 – 0.41	
0.41 – 0.42	
0.42 – 0.43	whiterhinoceros indianrhinoceros
0.43 – 0.44	
0.44 – 0.443	finwhale bluewhale pig sheep cow hippopotamus
0.443 – 0.445	harborSeal graySeal dog cat
0.445 – 0.45	
0.45 – 0.453	
0.453 – 0.454	fruitbat
0.454 – 0.457	aardvark
0.457 – 0.46	rat mouse fatdormouse squirrel rabbit
0.46 – 0.463	
0.463 – 0.466	guineapig armadillo
0.466 – 0.469	wallaroo opossum elephant
0.469 – 0.47	chimpanzee pigmychimpanzee human gorilla orangutan gibbon baboon
0.47 – 0.9	platypus

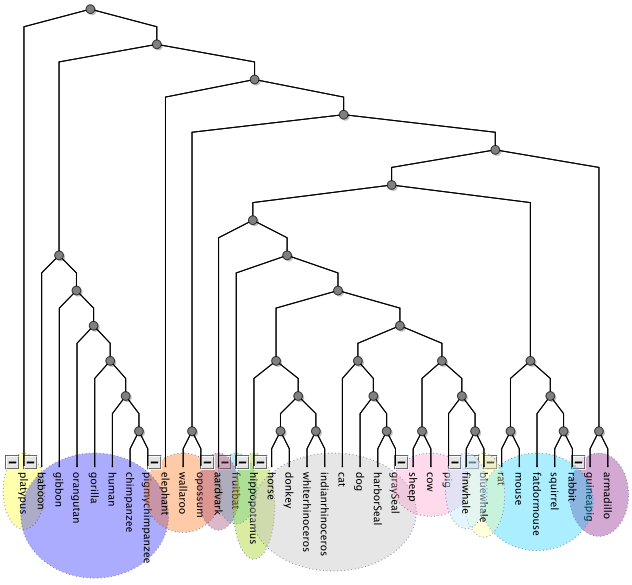


FIGURE 7. Clustering with our method and WPGMA

We have compared the two methods also on 100 words, chosen randomly form dictionary. The *WPGMA* results are shown in Fig. 8 :

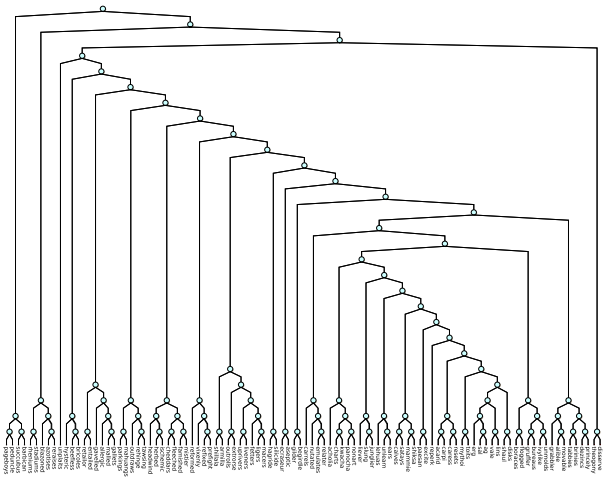


FIGURE 8. Clustering of 100 words with *WPGMA*

Our results are the same, whatever the chosen data and they are shown in Table III and Fig. 9 :

We see in the results of comparisons that, first, the results of our method similar to those of the hierarchical clustering (*WPGMA*), hence the consistency of our algorithm. Second, the clusters remain unchanged whatever the selected data.
NB : A few differences between results are due to the values of threshold chosen (arbitrarily), as if we obtain slightly different classes with different cuts in dendrograms.

Our first objective was to propose a method which provides a general view of data proximity, in particular in large databases. This objective is achieved, by giving a consistent result in just $O(n)$ iterations, thus allows the feasibility of the clustering in high-dimensional data.

TABLE III
CLUSTERING OF 100 WORDS, USING "TOITS" AS CHOSEN DATA, IN $O(n)$ OPERATIONS

Thresholds	Clusters
0 – 0.4	mythoi
0.4 – 0.48008	unmolds ivylike bureaus gruffer flogged boraces
	disks shaul lins vale ag sal erg resets caress carpi
	acarid reperk excite paesan shiksa mammie satays
	calves ears unlearn lehuas junglier slung liever
	nonart panocha kwacha charts acholia realter
	emulates nutated carrels
0.48008 – 0.48009	princely dizincs brinies tabbies movable alible
	grabbler begirdle gilder
0.48009 – 0.489	aseptic ecraseur
0.489 – 0.49	silicide hagride
0.49 – 0.492	macers ligers fetters liveners uprivers extrorse
	outrolls armilla shillala
0.492 – 0.495	prefight refixed vixenly reformed
0.495 – 0.497	mistier famished fleeched cheddars ischemic herbed headwind
0.497 – 0.498	twasing rehinge outdraws nahuangs parkings
0.498 – 0.499	gallets malled allergic gavelled entailed retailor
0.499 – 0.5	guineapig armadillo
0.466 – 0.469	disserve theogony bricoles beefless hysteric unplaits
	reraises azoties blazoned staduims rheniums barbican
	succubus peduncle pageboys
0.5 – 1.0	

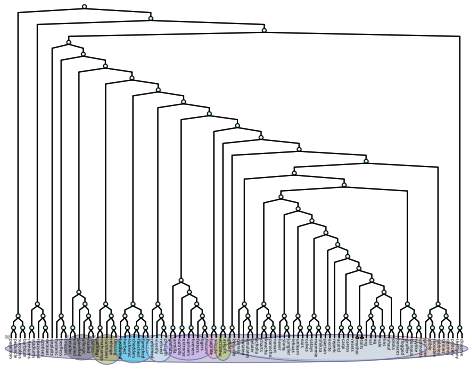


FIGURE 9. Clustering of 100 words with chosen data toits (see Appendix for clarity)

VI. CONCLUSION

We proposed in this paper a novel approach of clustering, which can overcome the problem of feasibility in large databases thanks to its calculation time of $O(n)$. We gave proofs that in an *ultrametric space* we can avoid the calculations of all mutual distances, without loosing informations about the proximities in the data set.

In the futur work we aim first, to generalize our approah to multicreteria clustering, using different kinds of connections between data, in particular qualitative relationships.

RÉFÉRENCES

[1] J. Abrahams, *Code and parse tree for lossless source encoding*, Compression and Complexity of Sequences, Vol 7.1, pages 198-222, 1997.

[2] M. Ankerst, M. M. Breunig, H. P. Kreigel and J. Sander, *OPTICS : Oredering Points To Identify the Clustering Structure*, ACM SIGMOD, Philadelphia PA, 1999.

[3] F. Brucker, *Modèles de classification en classes empiétantes*, Phd Thesis, Equipe d'acceuil : Dep. IASC de l'Ecole Superière des Télécommunications de Bretagne, France, 2001.

[4] M. Chavent and Y. Lechevallier, *Empirical comparison of a monothetic divisive clustering method with the Ward and the k-means clustering methods*, IFCS, Ljubljana Slovenia, 2006.

[5] R. Cilibrasi and P. M. B Vitanyi, *Clustering by Compession*, IEEE Transactions on Information Theory, 51(4), 2005.

[6] G. Cleuziou, *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information*, Phd Thesis, Université d'Orleans, France, 2006.

[7] A. B. Dragut and C. M. Nichitiu, *A Monotonic On-Line Linear Algorithm for Hierarchical Agglomerative Classification*, Information Technology and Management, Netherlands, 2004.

[8] S. Fouchal, M. Ahat, I. Lavallée, M. Bui& S. Benamor, *Clustering based on Kolmogorov Complexity*, KES (1) 2120 : 452-460, 2120.

[9] I. Lavallee, *Complexité et algorithmique avancée "un introduction"*, 2ème édition Hermann éditeurs, 2008.

[10] L. Gaji´, *Metric locally constant function on some subset of ultrametric space*, Novi Sad J. Math., Vol. 35, No. 1, 2005, 123-125.

[11] L. Gaji´, *On ultrametric space*, Novi Sad J. Math., Vol. 31, No. 2, 2001, 69-71.

[12] G. Gardanel, *Espaces de Bannach ultramétriques*, Séminaire Delange-Pistot-Poitou Théorie des nombres, Tome 9, n˚2,1967-1968, exp. n˚G2, G1-G8.

[13] I. Gronau and S. Moran, *Optimal implentations of UPGMA and other common clustering algorithms*, Information Processing Letters, Vol 104, Issue 6, 205-210, 2007.

[14] A. Hinneburg and D. A. Keim, *An Efficient Approach to Clustering in Large Multimedia Databases with Noise*, American Association for Artificial Intelligence, 1998.

[15] A. K. Jain, M. N. Murty and P. J. Flynn, *Data Clustering : A Review*, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[16] G. Karypis, E. H. Han and V. Kumar, *Chameleon : Hierarchical Clustering Using Dynamic Modeling*, Computer, pages 68-75, 1999.

[17] M. Krasner, *Espace ultramétrique et valuation*, Séminaire Dubreil Algèbre et théorie des nombres, Tome 1, 1947-1948, exp. n˚1, 1-17.

[18] H. P. Kriegel, P. Kroger and A. Zymek, *Clustering High-Dimensional Data : A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering*, ACM Transactions on Knowledge Discovery from Data, Vol. 3, No. 1, Article 1, 2009.

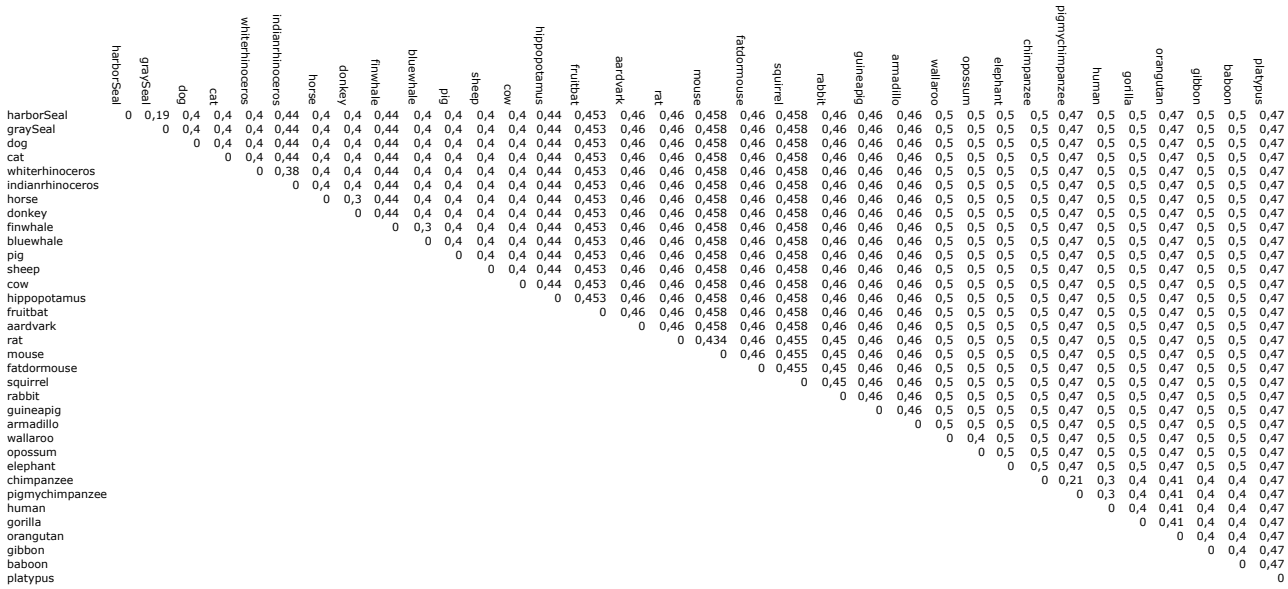
[19] V. Levorato, T. V. Le, M. Lamure and M. Bui, *Classification prétopologique basée sur la complexité de Kolmogorov*, Studia Informatica, Vol 7.1, 198-222, Hermann, 2009.

[20] F. Murtagh, *A Survey of Recent Advances in Hierarchical Clustering Algorithms*, The Computer Journal, Vol. 26, NO. 4, 1983.

[21] F. Murtagh, *Complexities of Hierarchic Clustering Algorithms : State of the Art*, Computational Statistics Quartely, Vol. 1, Issue 2, 1984, 101-113.

[22] K. P. R. Rao, G. N. V. Kishore and T. Ranga Rao, *Some Coincidence Point Theorems in Ultra Metric Spaces*, Int. Journal of Math. Analysis, Vol. 1, 2007, no. 18, 897 - 902.

[23] M. Terrenoire and D. Tounissoux, *Classification hiérarchique*, Université de Lyon, 2003.



	harborSeal	graySeal	dog	cat	whiterhinoceros	Indianrhinoceros	horse	donkey	finwhale	bluewhale	pig	sheep	cow	hippopotamus	fruitbat	aardvark	rat	mouse	fatdormouse	squirrel	rabbit	guineapig	armadillo	wallaroo	opossum	elephant	chimpanzee	pigmychimpanzee	human	gorilla	orangutan	gibbon	baboon	platypus
harborSeal	0	0,19	0,4	0,4	0,4	0,44	0,4	0,4	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
graySeal		0	0,4	0,4	0,4	0,44	0,4	0,4	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
dog			0	0,4	0,4	0,44	0,4	0,4	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
cat				0	0,4	0,44	0,4	0,4	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
whiterhinoceros					0	0,38	0,4	0,4	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
indianrhinoceros						0	0,4	0,4	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
horse							0	0,3	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
donkey								0	0,44	0,4	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
finwhale									0	0,3	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
bluewhale										0	0,4	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
pig											0	0,4	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
sheep												0	0,4	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
cow													0	0,44	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
hippopotamus														0	0,453	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
fruitbat															0	0,46	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
aardvark																0	0,46	0,458	0,46	0,458	0,46	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
rat																	0	0,434	0,46	0,455	0,45	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
mouse																		0	0,46	0,455	0,45	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
fatdormouse																			0	0,455	0,45	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
squirrel																				0	0,45	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
rabbit																					0	0,46	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
guineapig																						0	0,46	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
armadillo																							0	0,5	0,5	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
wallaroo																								0	0,4	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
opossum																									0	0,5	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
elephant																										0	0,5	0,47	0,5	0,5	0,47	0,5	0,5	0,47
chimpanzee																											0	0,21	0,3	0,4	0,41	0,4	0,4	0,47
pigmychimpanzee																												0	0,3	0,4	0,41	0,4	0,4	0,47
human																													0	0,4	0,41	0,4	0,4	0,47
gorilla																														0	0,41	0,4	0,4	0,47
orangutan																															0	0,4	0,4	0,47
gibbon																																0	0,4	0,47
baboon																																	0	0,47
platypus																																		0

FIGURE 10. Ultrametric distance matrix of 34 data

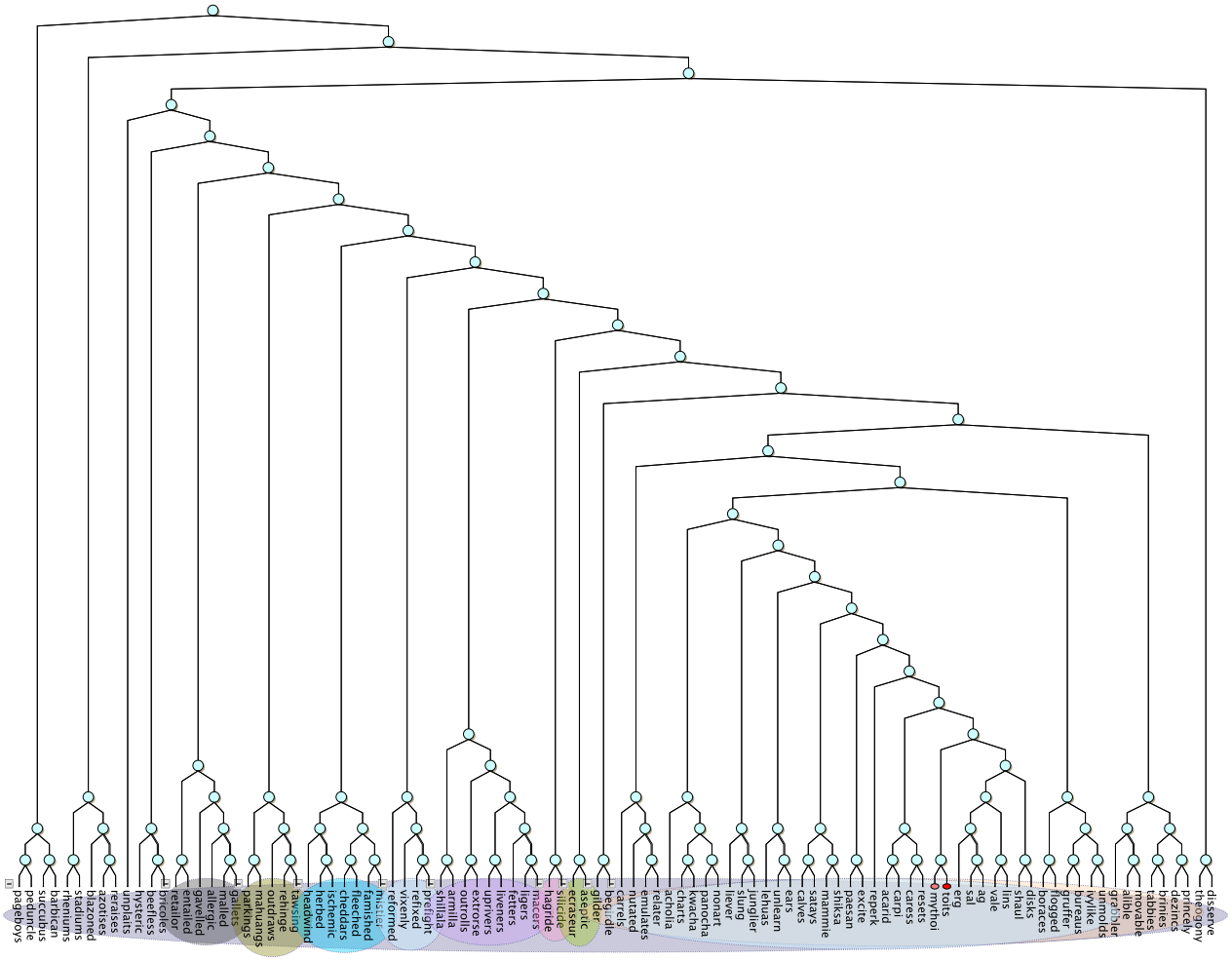


FIGURE 11. Clustering of 100 words with chosen data toits