Metric retractions and similarity detecting algorithms

Gábor Sági
Alfréd Rényi Institute of Mathematics,
Reáltanoda u. 13-15,
H-1053 Budapest, Hungary
and
Budapest University of
Technology and Economics,
Department of Algebra,
Egry J. u. 1,
H-1111 Budapest, Hungary
Email: sagi@renyi.hu

Karrar Al-Sabti
Budapest University of
Technology and Economics,
Department of Algebra,
Egry J. u. 1,
H-1111 Budapest, Hungary
Email: karrar.al-sabti@edu.bme.hu
and
University of Kufa,
Faculty of Computer Science and Mathematics,
P.O. Box 21, Kufa, Najaf 540011, Iraq,
Email: karrard.alsabti@uokufa.edu.iq

***Abstract*—Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space, let $A \subseteq X$ and let ε be a positive real number. The similarity detecting problem is to find all $a \in A$ for which $\varrho(a, x) \leq \varepsilon$ where $x \in X$ is a given input.**

In this work we study the similarity detecting problem with the additional assumption that $\mathcal{X}$ is an ultrametric space of finite spectrum; these assumptions seem to be natural from the point of view of practical applications. We establish model theoretical results for ultrametric spaces. More concretely, we provide sufficient conditions for the existence of metric retractions for certain ultrametric spaces. Based on these theoretical results, we propose a similarity detecting algorithm for ultrametric spaces. The time complexity of our algorithm will be discussed, as well.

I. Introduction

The similarity detecting problem can be described as follows. Suppose we are given

- a metric space $\mathcal{X} = \langle X, \varrho \rangle$,
- a set $A \subseteq X$ and
- a positive real number $\varepsilon \in \mathbb{R}^+$.

Intuitively, X is a set of instances of an abstract data type, the distance function ϱ measures "how similar" the elements of X are (that is, the distance $\varrho(a, b)$ is smaller, the "similarity" of a and b are larger), A is a "database" and ε is an amount of inaccuracy we tolerate. The problem is to find all elements of A whose distance from a given input x is at most ε.

Often, X may be infinite and A is finite but huge. The crucial point is to find a representation for A for which algorithms can perform efficiently. In these problems, usually the metric space is compact and often, it has an ultrametric distance function. For related investigations we refer to [1], [4], [7] [8], [9] and [10].

As we mentioned in the previous paragraph, the challenge in related investigations is to find suitable representations of metric spaces. Beyond their theoretical interest, this practical aspect also motivates investigations of the model theory of metric spaces: model theoretical results for metric spaces may provide a better understanding for how certain metric spaces can be constructed from their finite subspaces. This structural information may help to design clever representations for metric spaces. In this purely theoretical direction we refer to [3], [5], [11] and [12].

As the main result of the present work, we propose a similarity detecting algorithm for ultrametric spaces. This can be regarded as a continuation initiated in [10]. The differences between [10] and the present work can be summarized as follows: the investigated classes of metric spaces in [10] and in the present work are different and the methods we are utilizing are different. More concretely,

- In [10] we focused our attention to the special case when our space $\mathcal{X}$ is a large (finite) dimensional Euclidean space. By a kind of "dimension reduction" we proposed an algorithm which finds elements of A "similar" to a given input x. In the worst case, the number of steps in

that algorithm was proportional with $|A|$.

• In the present work we will investigate a different class of metric spaces, more concretely ultrametric spaces. In this special case, using metric retractions (see Definition 3.1 below) we propose an algorithm which finds elements of A "similar" to a given input x. Again, the number of required steps in the worst case will be proportional with $|A|$.

The sutructure of this paper is as follows. At the end of this section we are summing up our system of notation. In Section II we sum up the technical preliminaries. Section III is devoted to metric retractions, this provides the theoretical background for our algorithms. Finally, in Section IV we present and analyze the time complexity of our similarity detecting algorithms for ultrametric spaces.

Notation

Our notation is mostly standard, but the following list may help.

Throughout $\mathbb{R}$ and $\mathbb{R}^+$ denote the set of real numbers, and the set of positive real numbers, respectively.

Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space, $a \in X$ and let γ be a non-negative real number. As usual, the open γ-ball $B(\gamma, a)$ at a is the set

$$B(\gamma, a) = \{x \in X : \varrho(a, x) < \gamma\}.$$

For a function f, the domain and range of f will be denoted by $dom(f)$ and by $ran(f)$, respectively.

II. PRELIMINARIES

Suppose $\mathcal{X} = \langle X, \varrho \rangle$ is a metric space. $\mathcal{X}$ is defined to be an ultrametric space iff for all $a, b, c \in X$ we have

$$\varrho(a, b) \leq max\{\varrho(a, c), \varrho(c, b)\}.$$

The spectrum of $\mathcal{X}$ is the range of ϱ.

Next we recall a well known method associating a relational structure $\mathcal{A}(\mathcal{X})$ for a metric space $\mathcal{X}$. For each $d \in ran(\varrho)$ we introduce a binary relation R_d as follows:

$$R_d = \{\langle a, b \rangle \in X^2 : \varrho(a, b) \leq d\}$$

and we define

$$\mathcal{A}(\mathcal{X}) := \langle X, R_d \rangle_{d \in ran(\varrho)}.$$

This first order relational structure completely describes $\mathcal{X}$. We will investigate model theoretic properties of $\mathcal{A}(\mathcal{X})$. We assume the reader is familiar with the basics of model theory. We refer to [2] for model theoretic notions not recalled here. Also, as in [8], [11] and [12] by a "model theoretic property of a metric space $\mathcal{X}$" we mean the corresponding property of its associated structure $\mathcal{A}(\mathcal{X})$.

For the reader's convenience we breafly sum up some model theoretical notions and recall a theorem from [12] which will be used in the present work. Let $\mathcal{A}$ be a first order structure, $B \subseteq A, a \in A$ and let Δ be a set of formulas. Then the Δ-type of a in $\mathcal{A}$ over B is defined to be

$$tp^{\mathcal{A}}_{\Delta}(a/B) = \{\varphi(v, \bar{b}) : \varphi \in \Delta, \bar{b} \in B, \mathcal{A} \models \varphi(a, \bar{b})\}.$$

Further, $tp^{\mathcal{A}}_{\Delta}(a/B)$ splits over $C \subseteq B$ iff there are $\bar{b}_0, \bar{b}_1 \in B$ such that

$$tp^{\mathcal{A}}_{\Delta}(\bar{b}_0/C) = tp^{\mathcal{A}}_{\Delta}(\bar{b}_1/C),$$

but $\varphi(v, \bar{b}_0), \neg\varphi(v, \bar{b}_1) \in tp^{\mathcal{A}}_{\Delta}(a/B)$; less formally, $tp^{\mathcal{A}}_{\Delta}(a/B)$ splits over C iff there exist tuples $\bar{b}_0, \bar{b}_1 \in B$ that "look like the same from the point of view of C", but a and some $\varphi \in \Delta$ "distinguish" $\bar{b}_0$ and $\bar{b}_1$.

In addition, B is a splitting base for a iff for all $X \subseteq A - \{a\}$ with $B \subseteq X$, $tp^{\mathcal{A}}(a/X)$ does not split over B. For the reader's convenience we quote here Theorem 3.1 of [12] which, together with its easy consequence Theorem 2.2, will be important in Section III.

Theorem 2.1: (Theorem 3.1 in [12].) Let $\mathcal{X} = \langle X, \varrho \rangle$ be an ultrametric space of finite spectrum, let $Y \subseteq X$ and let $\Delta = \{R_\alpha : \alpha \in ran(\varrho)\}$ be the set of atomic formulas of $\mathcal{A}(\mathcal{X})$. Then for each $a \in X - Y$ there exists a finite $B \subseteq Y$ such that $tp_{\Delta}(a/Y)$ does not split over B. Moreover, $|B| \leq 2 \cdot \binom{|ran(\varrho)|-1}{2}$.

Theorem 2.2: Let $\mathcal{X} = \langle X, \varrho \rangle$ be an ultrametric space of finite spectrum. Then each $a \in X$ has a splitting base B.

Proof: Apply Theorem 2.1 to $Y = X - \{a\}$. ■

III. METRIC RETRACTIONS

Throughout this section, Δ will denote the set of atomic formulas of the language of the associated structures of the (ultra)metric spaces we are investigating.

Definition 3.1: Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space and let $Y \subseteq X$. A function $f : X \to X$ is defined to be a metric retraction iff $f|_Y$ is the identity function of Y, $ran(f) \subseteq Y$ and

for all $a, b \in X$ either $f(a) = f(b)$ or
$$\varrho(a, b) = \varrho(f(a), f(b)).$$

Further, Y is called a metric retract of X.

Retractions (which are not necessarily metric retractions) of ultrametric spaces has been investigated in [13].

Lemma 3.2: Suppose $\mathcal{X} = \langle X, \varrho \rangle$ is an ultrametric space and $B \subseteq X$ is a splitting base for all $a \in X$. Suppose $x, x', y, y' \in X$ are such that $x \neq y$, $x' \neq y'$,

$$tp_\Delta(x/B) = tp_\Delta(x'/B) \text{ and}$$
$$tp_\Delta(y/B) = tp_\Delta(y'/B).$$

Then

(1) $\varrho(x', y') \leq \varrho(x, y)$;
(2) $\varrho(x', y') = \varrho(x, y)$.

Clearly, (2) implies (1), so (1) seems to be superfluous to state. Indeed, (1) is stated for technical reasons only: as we will see, first we will show (1) and then the proof of (2) will be reduced to (1) by a symmetry argument.

Proof: By assumption, B is a splitting base for y, so we have

$$\varrho(x, y) = \varrho(x', y).$$

Similarly, B is a splitting base for x hence

$$\varrho(x, y) = \varrho(x, y').$$

Combining these observations, we get

$$\varrho(x', y') \leq max\{\varrho(y', x),\ \varrho(x, y),\ \varrho(y, x')\} = \varrho(x, y).$$

Now for (2), we apply the first part twice. In more detail, since the conditions are completely symmetric between x and x' and also between y and y', we can use the previous argument two times (interchanging x with x' and y with y' in the second time) and we get

$$\varrho(x', y') \leq \varrho(x, y) \leq \varrho(x', y'),$$

as desired. ■

Theorem 3.3: As in Lemma 3.2, suppose $\mathcal{X} = \langle X, \varrho \rangle$ is an ultrametric space and $B \subseteq X$ is a spiltting base for all $a \in X$.

Assume further, that $B \subseteq Y \subseteq X$ is such that every Δ-type over B can be realized in Y. Then there exists a metric retraction $f : X \to X$ over Y such that for all $x, y \in X$

$$f(x) = f(y) \text{ implies } tp_\Delta(x/B) = tp_\Delta(y/B).$$

(Note that if B and the range of ϱ are finite, then Y may also be chosen to be finite).

Proof: By our assumptions, for all $x \in X$ there exists $x' \in Y$ such that

$$tp_\Delta(x/B) = tp_\Delta(x'/B).$$

We assume $x' = x$ whenever $x \in Y$. Define $f(x) = x'$ for all $x \in X$. We shall show that this f is a metric retraction over Y. Clearly, $f|_Y$ is the identity function of Y and $ran(f) \subseteq Y$.

Next assume, $x, y \in X$ are arbitrary; we shall show

$$(*) \qquad f(x) = f(y) \text{ or } \varrho(x, y) = \varrho(f(x), f(y)).$$

If $f(x) = f(y)$ then $(*)$ holds obviously. If $(x) \neq f(y)$ then the conditions of Lemma 3.2 are satisfied for x, y and for $x' = f(x),\ y' = f(y)$. It follows from Lemma 3.2(2), that

$$\varrho(x, y) = \varrho(x', y') = \varrho(f(x), f(y)),$$

that is, $(*)$ holds, as desired. ■

Lemma 3.4: Suppose $\mathcal{X} = \langle X, \varrho \rangle$ is an ultrametric space, $Y \subseteq X$, $f : X \to Y$ is a metric retraction over Y and $\varepsilon \in \mathbb{R}^+$ such that for all $x \in X$ there exists $x' \in X$ such that $f(x) \neq f(x')$ and $\varrho(x, x') \leq \varepsilon$. Then for all $x, y \in X$

$$f(x) = f(y) \text{ implies } \varrho(x, y) \leq \varepsilon.$$

Proof: Let $x, y \in X$ be such that $f(x) = f(y)$. By assumption, there exists x' such that $\varrho(x, x') \leq \varepsilon$ and $f(x) \neq f(x')$. Observe, that $f(y) \neq f(x')$ hence $\varrho(x', y) = \varrho(f(x'), f(y))$ Now

$$\varrho(x, y) \leq max\{\varrho(x, x'), \varrho(x', y)\} \leq$$
$$max\{\varepsilon, \varrho(f(x'), f(y))\} \leq$$
$$max\{\varepsilon, \varrho(f(x'), f(x))\} = max\{\varepsilon, \varrho(x', x)\} = \varepsilon,$$

as desired. ■

Lemma 3.5: Suppose $\mathcal{X} = \langle X, \varrho \rangle$ is an ultrametric space of finite spectrum. Let $\varepsilon_0 \in \mathbb{R}^+$ be the smallest element of $ran(\varrho) - \{0\}$ and suppose $\varepsilon > \varepsilon_0$ is such that for all $x \in X$ there exists $x' \in X$ with $\varrho(x, x') = \varepsilon$. Then there exists $Y \subseteq X$ and $f : X \to Y$ such that

(1) f is a metric retraction over Y;
(2) $f(x) = f(y)$ implies $\varrho(x, y) \leq \varepsilon$.

Proof: Let $\{a_i : i < \kappa\} \subseteq X$ be such that

$$\bigcup_{i<\kappa} B(\varepsilon_0, a_i) = X$$

(where $B(\varepsilon_0, a_i)$ is the ball with origin a_i and radius ε_0). By Theorem 2.2, for all i there exists a splitting base for a_i in $X - \{a_i\}$ such that

$$|B_i| \leq 2 \cdot (\tbinom{|ran(\varrho)|-1}{2}).$$

Let

$$B = \bigcup_{i<\kappa} B_i \bigcup \{a_i : i < \kappa\}.$$

Then B is a splitting base for all $a \in X$ because of the following. Assume $u, v \in X - \{a\}$ are such that for all $b \in B$ we have

$$\varrho(u, b) = \varrho(v, b).$$

We shall show

$$(*) \qquad \varrho(a, u) = \varrho(a, v).$$

By construction, there exists $i < \kappa$ such that $\varrho(a, a_i) \leq \varepsilon_0$. We proceed by a case distinction.
Case 1: $u = a_i$. Since $a_i \in B$, we get

$$0 = \varrho(u, a_i) = \varrho(v, a_i),$$

that is, $v = a_i$, therefore $u = v$, as well. Hence $(*)$ holds, as desired.
Case 2: $v = a_i$. This case can be treated similarly to the previous case.
Case 3: $u, v \in X - \{a_i\}$. By construction, B_i is a splitting base for a_i. Hence $\varrho(a_i, u) = \varrho(a_i, v)$. If $\varrho(a_i, u) > \varepsilon_0$, then $(*)$ follows from Lemma 2.7 of [12]. Finally, suppose

$$\varrho(a_i, u) = \varrho(a_i, v) = \varepsilon_0.$$

Then

$$\varrho(a, u) \leq max\{\varrho(a, a_i), \varrho(a_i, u)\} = \varepsilon_0.$$

But $a \neq u$, so $\varrho(a, u) = \varepsilon_0$. One can show similarly, that $\varrho(a, v) = \varepsilon_0$, as well. Therefore $(*)$ holds in this last case, too.

Let $Y \subseteq X$ be such that every Δ-type over B can be realized in Y. Then, by Theorem 3.3 there exists a metric retraction $f : X \to Y$ such that for all $x, y \in X$

$$f(x) = f(y) \text{ implies } tp_\Delta(x/B) = tp_\Delta(y/B).$$

But for all $x \in X$ there exists $i < \kappa$ such that $\varrho(x, a_i) \leq \varepsilon_0$. As $a_i \in B$, it follows, that

$$f(x) = f(y) \text{ implies } \varrho(y, a_i) = \varrho(x, a_i) \leq \varepsilon_0,$$

hence

$$f(x) = f(y) \text{ also implies } \varrho(x, y) \leq \varepsilon_0.$$

In particular, if $\varrho(x, x') = \varepsilon$, then $f(x) \neq f(x')$. Therefore, (2) follows from Lemma 3.4. ■

IV. Similarity detecting based on metric retractions

Based on the sections above, We describe a similarity detecting algorithm for ultrametric spaces of finite spectrum. Our algorithm consists of two parts, the Initializing and the Searching Part. The Initializing Part executes the necessary steps we need to obtain a relatively fast Searching Part. The Initializing Part should be executed once at the beginning. We assume that the Searching Part will be used much more frequently.

Explanations and comments will be provided right after describing the initialization and searching parts of our algorithm.

Throughout this section we fix a finite ultrametric space $\mathcal{X} = \langle X, \varrho \rangle$ and a number $\varepsilon > min(ran(\varrho) - \{0\})$. Further, we assume that for all $x \in X$ there exists $x' \in X$ with $\varrho(x, x') = \varepsilon$.

Our algorithms can be described as follows.

Initializing Part.
Input: A subset $A \subseteq X$.
(1) Find a metric retraction $f : X \to Y$ for some Y as in Lemma 3.5;
(2) Fix an enumeration $Z = \{y_i : i \leq n\}$ of $\{f(x) : x \in A\}$.
(3) for each $y \in Z$ compute the list

$$l_y = \{x \in A : f(x) = y\}.$$

Searching Part.
Input: $x \in X$.
(1) Let $u := f(x)$, let $i := 1$.
(2) While $i \leq n$ do
(3) If $\varrho(u, y_i) \leq \varepsilon$ then add l_{y_i} to the output;
(4) let $i := i + 1$;
(5) end while.

Remarks on the Initializing Part. The proof of Theorem 3.1. in [12] and the proof of Theorem 3.3 above are constructive, hence in Step 1 one can compute f and Y which satisfy the conclusion of Lemma 3.5.

The number of required elementary steps for

Step 1 is proportional with $|X|^2$, but this should be executed only once.

We note that in Step 2, Z is a subset of $Y = ran(f)$, this is the reason for denoting the elements of Z by y_i. The number of required elementary steps for Step 2 is proportional with $|A|$, as we may assume that the input A is given in some data structure, for example in a list: $A = \{a_i : i \leq |A|\}$.

The number of required elementary steps for Step 3 is proportional with $|A|$. If the input A is given in a list $A = \{a_i : i \leq |A|\}$, then going through in this list, in the i^{th} elementary step we add a_i to the list $l_{f(a_i)}$.

If the database A changes in time, then it is enough to correct the lists containing the modified elements. This cost is negligible (proportional with the number of modified points in A).

Remarks on the Searching Part. First we note that the Searching Part always provides sound answers because of the following. Assume $x \in X$ is the input, $u := f(x)$ and $a \in A$ is a part of the output. Then $\varrho(u, y_i) \leq \varepsilon$ for some $i \leq n$ and a belongs the list l_{y_i}. In addition, $f(a) = y_i$ holds, as well.

Case 1: $\varrho(u, y_i) = 0$, that is, $u = y_i$. As $u = f(x)$ and $y_i = f(a)$, in this case $f(x) = f(a)$, hence by Lemma 3.5(2) we have $\varrho(x, a) \leq \varepsilon$, so in this case a is a sound output.

Case 2: $\varrho(u, y_i) > 0$. As in the previous case, $u = f(x)$ and $y_i = f(a)$. In particular, $f(x) \neq f(a)$. Since f is a metric retraction,

$$\varrho(a, x) = \varrho(f(a), f(x)) = \varrho(y_i, u) \leq \varepsilon,$$

therefore a is a sound output, as well. As each element of A is contained in some list l_{y_i}, it also follows that the output contains all $a \in A$ for which $\varrho(x, a) \leq \varepsilon$.

The number of required steps is proportional with $|Z|$, which in general, may be as large, as $|A|$, but clearly, $|Z| \leq |A|$. The precise number of required steps strongly depends on the structure of A, hence, at that level of generality we cannot improve further the estimation for the time complexity.

ACKNOWLEDGMENT

This work has been supported by Hungarian National Foundation for Scientific Research grant K129211.

REFERENCES

[1] M. BRESSAN, N. CESA-BIANCHI, S. LATTANZI, AND A. PAUDICE, *Exact Recovery of Clusters in Finite Metric Spaces Using Oracle Queries,* Proceedings of Thirty Fourth Conference on Learning Theory, PMLR 134:775-803 (2021).

[2] C. C. CHANG, H. J. KEISLER, *Model Theory,* North–Holland, Amsterdam (1973).

[3] CONANT, G., *Neostability in countable homogeneous metric spaces,* Ann. Pure Appl. Logic, 168, 1442-1471 (2017).

[4] Faloutsos, C. and Lin, K., *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*, ACM., Vol. 24, No. 2, 163–174, (1995).

[5] M. ETEDADIALIABADI, S. GAO, AND L. M. FRANÇOIS AND J. MELLERAY, *Dense locally finite subgroups of automorphism groups of ultraextensive spaces,* Adv. Math. 391 (2021).

[6] Hjaltason, G.R. and Samet, H., *Contractive embedding methods for similarity searching in metric spaces*, Technical report, Computer Science Department, Center for Automation Research, Institute for Advanced Computer Studies, University of Maryland, (2000).

[7] K. T. HUBER, V. MOULTON AND A. SPILLNER, *Optimal realizations and the block decomposition of a finite metric space,* Discrete Applied Mathematics, Volume 302, pp.103-113 (2021).

[8] G. SÁGI, *Almost injective Mappings of Totally Bounded Metric Spaces into Finite Dimensional Euclidean Spaces,* Advances in Pure Mathematics, 9. pp. 555-566 (2019).

[9] G. Sági , and D. Nyiri, *On embeddings of finitw mwtric spaces*, in: the Proceedings of the 13th International Scientific Conference on Informatics (editors: V. Novitzka, S. Korečko and A. Szakál), IEEE, (2015).

[10] G. Sági and K. Al-Sabti, *Totally bounded metric spaces and similarity detecting algorithms, ,* in: the Proceedings of the 15th International Scientific Conference on Informatics (editors: W. Steingartner, S. Korečko and A. Szakál), pp. 338–342, IEEE, (2019).

[11] G. Sági and K. Al-Sabti, *Totally Bounded Metric Spaces, Their Model Theoretic Stability and Similarity Detecting Algorithms,* IPSI Transactions on Internet Research, Vol. 16, No. 2, pp. 24-30 (2020).

[12] G. Sági and K. Al-Sabti, *On some model theoretic properties of totally bounded ultrametric spaces*, Accepted for publication in *Mathematics*, (2022).

[13] G. Sági and K. Al-Sabti, *On the endomorphism monoid of certain ultrametric speces*, Submitted (2022).