European Journal of Academic Essays 1(7): 59-62, 2014
ISSN: 2183-1904
www.euroessays.org

On the Construction of special metrics involving Levenshtein and Hamming distances

Obeng-Denteh, William[1], Ayekple, Yao Elikem[1] ,Quansah, Amissah Mavis[1], Zigili, David Delali[1]

[1]Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
obengdentehw@yahoo.com

Abstract: Metric spaces are one of the useful types of topological spaces in measurement of length, and distance between points in underlying sets. They are useful for much more abstract and irregular sets than intervals in R^2 or R^3. The ability to measure and compare distances between elements is often crucial and metric spaces provide more structure than general topological spaces. We begin by giving a brief introduction and background to topological spaces in general. We then go ahead to discuss metric spaces with a quick overview on concepts like connectedness, compactness, continuity and completeness and define formally what a metric is and give the basic properties of metric spaces. We state without proof some theorems in topology which we invoke subsequently in the construction of some metric spaces. A large class of metric spaces exists and this work illustrates the construction of some of these metric including the Hamming distance, the Levenshtein distance, the Hausdorff metric and other applications of metric spaces.

Keywords: Hamming distance, Levenshtein distance, Hausdorff, metric spaces.

1. Introduction

A metric space can be thought of as measurement of length and distances between points, area and volumes in an underlying set. A particular important example is the hamming distance and Levenshtein distance [1] which doesn't limit metric spaces to linear structures only. The concept of open and closed sets in Euclidean space extend to metric spaces as do the concepts of convergence of a sequence and continuity of functions.
Technically, a metric space assigns the numeric value to length **and** distances measured between structures that are complete, compact or seperable

Considering the implications and importance of the Hamming distance and the Levenshtein Distance, we believe that seeking to better understand metric spaces and how they are constructed is a worthy field for study. The objective of this paper is to explore the construction of some metric spaces which will lead to further understanding and possibly new insights into some of their uses.

During the 19th century two distinct movements developed that would ultimately produce the sibling specializations of algebraic topology and general topology [2]. The first was characterized by attempts to understand the topological aspects of surface-like objects that arise by combining elementary shapes, such as polygons or polyhedra. One early contributor to combinatorial topology, as this subject was eventually called, was the German mathematician Johann Listing, who published *VorstudienzurTopologie* (1847; "Introductory Studies in Topology") [3], which is often cited as the first print occurrence of the term *topology*

A topological space is a set X together with τ, a collection of subsets of X, satisfying certainaxioms. The collection τ is called a topology on X. The elements of X are usually called *points*, though they can be any mathematical objects [3]. A topological space in which the *points* are functions is called a function space. The sets in τ are called the open sets, and their complements in X are called closed sets. A subset of X may be neither closed nor open, either closed or open, or both. A set that is both closed and open is called a clopen set.

In 1905 a French mathematician Maurice Fréchet initiated the study of metric spaces. According to him a metric on a set X is a function (called the *distance function* or simply distance)

$d : X \times X \rightarrow \mathbf{R}$(where **R** is the set of real numbers). For all x, y, z in X, this function is required to satisfy the following conditions:

1. $d(x, y) \geq 0$ (*non-negativity*, or separation axiom)

Corresponding Author: William Obeng-Denteh, Mathematics Dept., KNUST, Kumasi, obengdentehw@yahoo.com

2. *d*(*x*, *y*) = 0 if and only if *x* = *y* (*identity of indiscernibles*, or coincidence axiom)
3. *d*(*x*, *y*) = *d*(*y*, *x*) (*symmetry*)
4. *d*(*x*, *z*) ≤ *d*(*x*, *y*) + *d*(*y*, *z*) (*subadditivity* / *triangle inequality*).

During the period up to the 1960s, research in the field of general topology flourished and settled many important questions. The notion of dimension and its meaning for general topological spaces was satisfactorily addressed with the introduction of an inductive theory of dimension. Compactness, a property that generalizes closed and bounded subsets of *n*-dimensional. Euclidean space, was successfully extended to topological spaces through a definition involving "covers" of a space by collections of open sets, and many problems involving compactness were solved during this period.

Richard Hamming also introduce *Error detecting and error correcting codes* in 1950 and was called Hamming distance. He introduced it in his fundamental paper on Hamming codes. It is used in telecommunication to count the number of flipped bits in a fixed-length binary word as an estimate of error, and therefore is sometimes called the signal distance.

Hamming weight analysis of bits is used in several disciplines including information theory, coding theory, and cryptography.

For *q*-array strings over an alphabet of size *q* ≥ 2 the Hamming distance is applied in case of orthogonal modulation, while the Lee distance is used for phase modulation. If *q* = 2 or *q* = 3 both distances coincide.

However, for comparing strings of different lengths, or strings where not just substitutions but also insertions or deletions have to be expected, a more sophisticated metric like the Levenshtein distance is more appropriate.

In the light of the above, we seek to construct special metric using the hamming and Levensthein distance for comparing strings where not just substitutions but also insertions or deletions have to be expected.

2. Metric spaces

A metric on a set *X* is a function (called the *distance function* or simply distance)

d :*X* × *X* → R

(whereR is the set of real numbers). For all *x*, *y*, *z* in *X*, this function is required to satisfy the following conditions:

- *d*(*x*, *y*) ≥ 0 (*non-negativity*, or separation axiom)
- *d*(*x*, *y*) = 0 if and only if *x* = *y* (*identity of indiscernibles*, or coincidence axiom)
- *d*(*x*, *y*) = *d*(*y*, *x*) (*symmetry*)
- *d*(*x*, *z*) ≤ *d*(*x*, *y*) + *d*(*y*, *z*) (*subadditivity* / *triangle inequality*).

A metric is called an ultrametric if it satisfies the following stronger version of the *triangle inequality* where points can never fall 'between' other points:

For all *x*, *y*, *z* in *X*, *d*(*x*, *z*) ≤ max (*d*(*x*, *y*), *d*(*y*, *z*))

A metric *d* on *X* is called intrinsic if any two points *x* and *y* in *X* can be joined by a curve with length arbitrarily close to *d*(*x*, *y*).

3. Application and Construction of Metric Spaces

Metric spaces are used in numerous applications involving the storage, manipulation, and presentation of information. Strings of symbols like the letters making up the words you read here, are the basic information units. In any situation where we wish to measure the similarities and differences between information units an appropriate metric can be found to do so.

Error-Correcting Codes

With the incredible amounts of information being transmitted over phone lines, through the internet, or from satellites in space to Earth, it is extremely important to know whether a given message has arrived intact. We expect that there will be some errors in transmission due to electrical surges, cosmic radiation, or a variety of other factors. We want to be able to recognize when this occurs and to correct the faulty message [1].

Each word of length n can be thought of as a vector of length n ,with all entries either 0s or 1s. we write the set of all these possibilities as $V^n = [(a_1, a_2, \cdots, a_n)\ |a_i \in \{0,1\}]$.So V^n is the product of n copies of the set $\{0,1\}$ .We now put a metric on this set.

Definition: the Hamming distance $D_H(x,y)$ between two words of length n is the number of places in which the words differ.

Example given

$x = (0,0,1,1,0,0,0,1,0)$

$y = (0,1,0,1,0,0,1,1,0)$

We find that $x\ and\ y$ differ in second ,third and seventh places,and therefore $D_H(x,y) = 3.$

Definition: A code length n is any subset $C\ of\ V^n$.we elements of C the codewords [1].

If the sender and receiver have agreed on a particular code, then when a word arrives that is not one of the codewords, the receiver knows that at least one error has occurred in transmission.

The Hamming distance between:

- "**toned**" and "**roses**" is 3.
- **1011101** and **1001001** is 2.
- **2173896** and **2233796** is 3.

Definition: Let C be a code of length n .define the minimum distance of the code C to be least hamming distance between two codewords in the codes.

Example

Consider the code length 6 given by

C= $\{(0,0,1,0,0,0),(1,0,0,1,1,1),(1,1,1,0,1,1),(0,1,0,0,1,0)\}$

a.The hamming distance between (0,0,1,0,0,0) and (1,0,0,1,1,1) is 5

b. The hamming distance between (0,0,1,0,0,0) and (1,1,1,0,1,1) is 4

c.The hamming distance between (0,0,1,0,0,0) and (0,1,0,0,1,0) is 3

d. The hamming distance between (1,0,0,1,1,1) and (1,1,1,0,1,1) is 3. [4]

e.The hamming distance between (1,0,0,1,1,1) and (0,1,0,0,1,0) is 4

f. The hamming distance between (1,1,1,0,1,1) and (0,1,0,0,1,0) is 3

Min hamming distance of $\{a,b,c,d,e,f\}$ is 3

Therefore the hamming distance 0f the above set C is 3.

Lenvenshtein distance

Definition: the levenshtein distance between sequences $x\ and\ y$ is given by

$$D_L(x,y) = min\{i_s + d_s + r_s\}$$

Where the minimum is taken over all sequences S that turn $x\ into\ y$

Example

DNA Sequences

DNA is a long thin molecule made up of millions of atom.within its structure lies the code that determines our genetic makeup and composed of nucleotides. DNA molecule consists of two chain wound together to form the familiar double helix. nucleotides in dna chain pair but do so with neighbours on opposite chain, the nucleotides are in four types namely; adinine(A) ,cytosine(C), guanine(G) and thymine(T) .the sequence of nucleotide on one chain determines the sequence on the opposite chain and we can represent part or all of a dna molecules with a sequence of letters A,C,G and T [1].

One of the most important problems in dna is how to compare distinct dna sequences .how different is one sequence of dna from another? This is a measure of the evolutionary distance between the two sequences, Measuring the distance between the two sequence is a function of these differenes provides insight into the nature of the evolutionary history of each species.

During the course of evolution ,dna sequence differences arise in a variety of ways .one of them is the nucleotide substitution ,the apparent replacement of a letter in one dna sequence relative to the original sequence. Another commonly occurring changes is the insertion or deletion of nucleotides realized as the insertion or deletion of letters corresponding to dna sequence [1].

Let $x = AGTTCGAATCC$ and $y = AGCTCAGGAATC$

Then we can get from $x\ to\ y$ by the following process:

x : AGTTCGAATCC

Replace T : AGCTCGAATCC

Insert A : AGCTCAGAATCC

Insert G : AGCTCAGGAATC

We can check, by examining all possibilities with three or fewer operations, that the fewest number of insertions,deletions and replacements to get us from $x = AGTTCGAATCC$ and $y = AGCTCAGGAATC$ is four, therefore $D_L(x,y) = 4$.

4. Conclusion

The work has been able to come out with construction of a lot of examples using the Hamming distance, the Levenshtein distance, the Hausdorff metric and other applications of metric spaces.

5. Recommendation

It is recommended that students could take up a research in into Hamming distance and the Levenshtein distance in other applications in the pure and applied mathematics and the sciences.

References

[1] Adams, C. and Franzosa, R. Introduction to Topology; Pure and Applied, (2008)

[2] "Topology." *Encyclopaedia Britannica. Encyclopaedia Britannica Online*. Encyclopædia Britannica Inc., 2014. Available:

<http://www.britannica.com/EBchecked/topic/599686/topology>. "Vorstudien zur Topologie."

[3] Hazewinkel, Michiel, ed. (2001), "Topological space", Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4

[4] Quansah, A. M. & Zigli, D.D. On the Construction of Special Metrics, BSc. Thesis, Department of Mathematics, KNUST, Kumasi, 2014