TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 1**

**NGHIÊN CỨU MỘT SỐ KỸ THUẬT THU GIẢM SỐ CHIỀU DỮ LIỆU**

*Người hướng dẫn*: **Th. VÕ ĐỨC VĨNH**

*Người thực hiện*: **PHAN THANH BÌNH - 51303015**

**TRẦN THANH DUY - 51303038**

Lớp **: 13050302**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2017**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 1**

**NGHIÊN CỨU MỘT SỐ KỸ THUẬT THU GIẢM SỐ CHIỀU DỮ LIỆU**

*Người hướng dẫn*: **Th. VÕ ĐỨC VĨNH**

*Người thực hiện*: **PHAN THANH BÌNH - 51303015**

**TRẦN THANH DUY - 51303038**

Lớp **: 13050302**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2017**

LỜI CẢM ƠN

Chúng em xin gửi tất cả lòng biết ơn và kính trọng của mình đến thầy Võ Đức Vĩnh, người đã hướng dẫn chúng em đồ án trong thời gian qua. Với đề tài nghiên cứu một số kỹ thuật thu giảm số chiều dữ liệu cùng các phương pháp thiết kế thầy đã hướng dẫn sẽ giúp đỡ em trong việc học và làm sau này.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm bài tập lớn của riêng chúng tôi và được sự hướng dẫn của Th. Võ Đức Vĩnh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung bài tập lớn của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*Phan Thanh Bình*

*Trần Thanh Duy*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Ngày nay cùng với sự phát triển của internet thì dữ liệu của ngành công nghệ thông tin ngày càng gia tăng.

Dữ liệu chuỗi thời gian tồn tại trong nhiều ứng dụng thực tế, từ các lãnh vực khoa học kỹ thuật cho đến kinh tế, tài chính. Trong những ứng dụng này, việc tìm kiếm những chuỗi con truy vấn có xuất hiện trong cơ sở dữ liệu chuỗi thời gian là một công việc rất cần thiết. Sự truy tìm dựa vào độ tương tự như vậy là một mô đun căn bản trong nhiều công tác khai phá dữ liệu chuỗi thời gian cao cấp hơn như gom cụm, phân lớp, tìm mô típ, phát hiện mẫu bất thường, khám phá luật kết hợp và trực quan hóa dữ liệu. Mặc dù có nhiều cách tiếp cận khác nhau đã được đề xuất, hầu hết các cách tiếp cận đều dựa trên một tiền đề chung là các phương pháp thu giảm số chiều và các cấu trúc chỉ mục không gian. Bài tổng quan này điểm qua các nghiên cứu mới đây và cho thấy cách mà những phương pháp này hội tụ về một khung thức chung của sự rút trích đặc trưng.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc484395457)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc484395458)

[TÓM TẮT iv](#_Toc484395459)

[MỤC LỤC 1](#_Toc484395460)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 5](#_Toc484395461)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

**DANH MỤC BẢNG**

CHƯƠNG 1 – KHÁI NIỆM

I. Dữ liệu chuỗi thời gian (time series):

1. Khái niệm:

Một chuỗi thời gian là chuỗi vector hoặc số thực phụ thuộc vào thời gian, mỗi giá trị biểu diễn một giá trị đo tại những thời điểm cách đều nhau. Những tập dữ liệu chuỗi thời gian rất lớn xuất hiện trong nhiều lĩnh vực khác nhau như kĩ thuật, kinh tế, khoa học, chính trị…

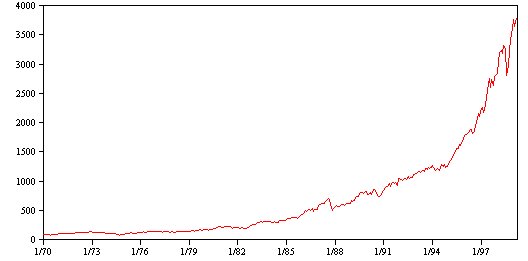
Vậy một chuỗi dữ liệu thời gian là những dữ liệu quan sát liên tục cho một hiện tượng (vật lý, kinh tế, chính trị,…) trong một khoảng thời gian sẽ tạo nên một chuỗi thời gian. Phân tích chuỗi thời gian có mục đích nhận dạng và tập hợp lại các yếu tố, những biến đổi theo thời gian mà nó có ảnh hưởng đến giá trị của biến quan sát.

Bên cạnh đó, dữ liệu chuỗi thời gian cũng có thể phân loại theo dạng đơn điệu (deterministic) hoặc không đơn điệu (stochastic) hay tuyến tính hoặc phi tuyến tính,…

2. Đặc điểm:

Việc nghiên cứu các số liệu, các hành vi trong quá khứ của một chuỗi sẽ giúp chúng ta xác định các mẫu và đưa ra dự báo tốt hơn.

2.1. Xu hướng (Trend):

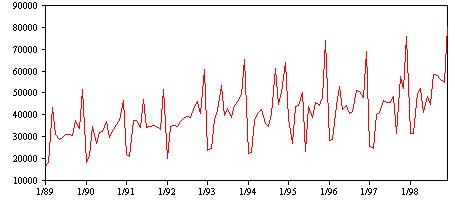


Hình . Trend.

Xu hướng là sự thay đổi dần dần hoặc đi xuống ở cấp độ của chuỗi hoặc xu hướng của các giá trị hàng loạt để tăng hoặc giảm theo thời gian.

Xu hướng có thể là tuyến tính hoặc phi tuyến. Xu hướng tuyến tính là các số dương hoặc âm tăng dần theo cấp độ chuỗi. Xu hướng phi tuyến thường nhân lên với số gia tăng tỉ lệ với các giá trị của chuỗi trước.

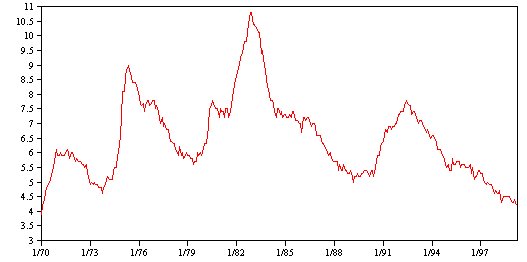
2.2. Chu kỳ liên hợp (Season cycles):



Hình . Season cycles.

Chu kỳ liên hợp là mẫu lặp đi lặp lại, có thể dự đoán được trong giá trị chuỗi. Được gắn với khoảng thời gian của chuỗi. Ví dụ như: dữ liệu hàng tháng, quý, năm,…

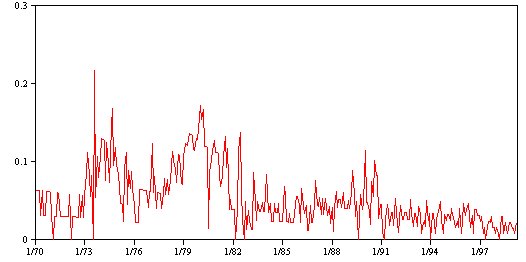
2.3. Chu kỳ không liên hợp (Nonseason cycles):



Hình . Nonseason cycles.

Chu kỳ không liên hợp cũng là một mẫu lặp đi lặp lại nhưng có thể sẽ không dự đoán được trong các giá trị chuỗi. Các mô hình thường không đối xứng và làm cho tính chắc chắn trong dự báo bị giảm đi nhiều. Ví dụ: lịch nghỉ tết âm mỗi năm đều khác nhau…

2.4. Xung và bước (Pulses and Steps):



Hình . Pules and Steps.

Có những thử nghiệm thì thường thay đổi đột ngột về mức độ và thường có hai loại:

+ Sự thay đổi đột ngột tạm thời: Pluses.

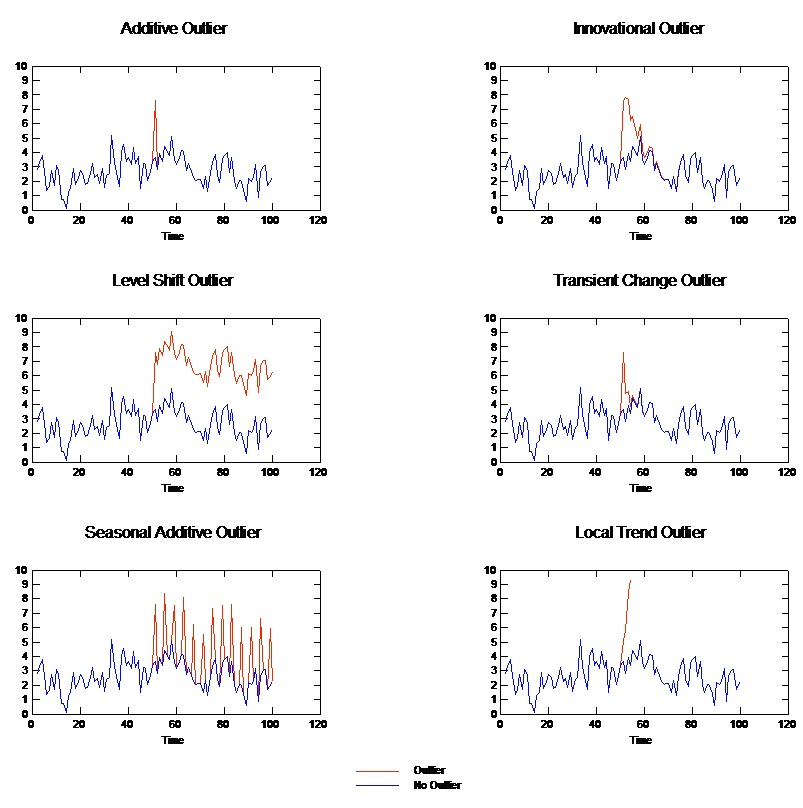
+ Sự thay đổi đột ngột vĩnh viễn: Steps.

Khi các bước hoặc xung được quan sát thấy, điều quan trọng là phải tìm ra một lời giải thích hợp lý. Các mô hình chuỗi thời gian được thiết kế để tính đến sự thay đổi dần dần, không đột ngột. Kết quả là, họ có xu hướng đánh giá thấp xung và bị hủy hoại bởi các bước, dẫn đến mô hình không phù hợp và dự báo không chắc chắn.

2.5. Ngoại lệ (Outliers):

Sự thay đổi về cấp bậc của một chuỗi thời gian mà không thể giải thích được gọi là các ngoại lệ (giá trị ngoài). Những quan sát này không phù hợp với phần còn lại của chuỗi và có thể gây ảnh hưởng đáng kể đến phân tích, và do đó ảnh hưởng đến khả năng dự báo của mô hình chuỗi thời gian.

Hình dưới đây hiển thị một số loại ngoại lệ thường xảy ra trong chuỗi thời gian. Các đường màu xanh đại diện cho một loạt mà không có giá trị ngoài. Các đường màu đỏ cho thấy một mô hình có thể có nếu chuỗi chứa các ngoại lệ. Những ngoại lệ này đều được phân loại là xác định bởi vì chúng ảnh hưởng đến mức trung bình của chuỗi.



Hình . Outliers.

Additive Outlier: Một additive outlier xuất hiện như là một giá trị lớn hay nhỏ đáng ngạc nhiên xảy ra cho một quan sát duy nhất. Các quan sát sau đó không bị ảnh hưởng bởi một additive outlier.

Innovational Outlier: Innovational Outlier được đặc trưng bởi một tác động ban đầu với các hiệu ứng kéo dài hơn các quan sát tiếp theo. Ảnh hưởng của các bên ngoài có thể tăng theo thời gian.

Level Shift Outlier: Đối với một sự thay đổi mức độ, tất cả các quan sát xuất hiện sau khi di chuyển vượt bậc đến một cấp độ mới.

Transient Change Outlier: Bên ngoài thay đổi như thế nào thì tương tự như các ngoại lệ thay đổi mức độ, nhưng tác động của bên ngoài giảm bớt số mũ theo các quan sát tiếp theo. Cuối cùng, chuỗi trở về mức bình thường.

Seasonal Additive Outlier: Một Seasonal Additive Outlier xuất hiện dưới dạng một giá trị lớn hoặc nhỏ đáng ngạc nhiên xảy ra liên tục trong khoảng thời gian đều đặn.

Local Trend Outlier: Kéo theo chung trong chuỗi một giá trị do một mẫu trong các ngoại lệ sau khi khởi đầu của ngoại tệ ban đầu.

II. Rút trích đặc trưng:

1. Rút trích đặc trưng là gì?

Rút trích đặc trưng là quá trình chuyển đổi tập dữ liệu thô ban đầu thành tập các thuộc tính có thể giúp biểu diễn tập dữ liệu ban đầu tốt hơn, tạo điều kiện để giải quyết các bài toán dễ dàng hơn, giúp tương thích với từng mô hình dự đoán cụ thể, cũng như cải thiện độ chính xác của mô hình dự đoán hiện tại.

2. Tại sao phải rút trích đặc trưng?

Rút trích đặc trưng là nhiệm vụ rất quan trọng giai đoạn tiền xử lý dữ liệu khi triển khai các mô hình khai phá dữ liệu. Vấn đề gặp phải khi các dữ liệu để đưa vào khai phá lại có quá nhiều thuộc tính và nhiều thông tin không cần thiết, gây nhiễu cho việc xây dựng mô hình. Chẳng hạn, một dữ liệu truyền vào gồm hàng trăm thuộc tính dùng để mô tả về khách hàng của một doanh nghiệp được thu thập, tuy nhiên khi xây dựng một Data mining model nào đó chỉ cần khoảng 50 thuộc tính từ hàng trăm thuộc tính đó. Nếu ta sử dụng tất cả các thuộc tính (hàng trăm) của khách hàng để xây dựng mô hình thì ta cần nhiều CPU, nhiều bộ nhớ trong quá trình Training model, thậm chí các thuộc tính không cần thiết đó làm giảm độ chính xác của mô hình và gây khó khăn trong việc phát hiện tri thức.

Các thuộc tính trong dữ liệu có tầm quan trọng, chức năng khác nhau và ảnh hưởng trực tiếp đến mô hình dự đoán, tùy theo từng bài toán mà chúng ta lựa chọn từng thuộc tính khác nhau và cần loại bỏ các thuộc tính không liên quan đến bài toán. Do đó ta cần xác định tốt cấu trúc của các thuộc tính sao cho diễn đạt hiệu quả nhất bản chất của tập dữ liệu.

CHƯƠNG 2 – CÁC PHƯƠNG PHÁP RÚT TRÍCH ĐẶC TRƯNG

I. Các phương pháp thu giảm số chiều dựa vào đặc trưng:

Dữ liệu chuỗi thời gian thường cực kỳ lớn. Tìm kiếm trực tiếp trên những dữ liệu này sẽ rất phức tạp và không hữu hiệu. Để khắc phục vấn đề này, ta nên áp dụng một số phương pháp biến đổi để thu giảm độ lớn của dữ liệu. Những phương pháp biến đổi này thường được gọi là nhũng kỹ thuật thu giảm số chiều (dimensionality reduction). Phương pháp tổng quát để thu giảm số chiều có thể tóm tắt như sau:

1. Thiết lập một độ đo tương tự d.

2. Thiết kế một kỹ thuật thu giảm số chiều để rút trích một đặc trưng có chiều dài k (tức là một đặc trưng gồm k giá trị), với k có thể được xử lý một cách hữu hiệu nhờ một cấu trúc chỉ mục không gian (đa chiều).

3. Cung cấp một độ đo tương tự dk trên một không gian đặc trưng k chiều và chứng tỏ rằng nó tuân thủ điều kiện sau đây: dk(X’, Y’) ≤ d(X, Y). (1)

Điều kiện (1) có nghĩa là hàm khoảng cách tính trên không gian đặc trưng (hay không gian thu giảm) của hai chuỗi thời gian đã được biến đổi X’, Y’ từ hai chuỗi thời gian ban đầu X, Y phải chặn dưới khoảng cách thật giữa chúng trên không gian nguyên thủy. Có ba nhóm phương pháp chính để thu giảm số chiểu là phương pháp biến đổi sang miền tần số, phương pháp xấp xỉ tuyến tính từng đoạn và phương pháp điểm quan trọng.

1. Các phương pháp biến đổi sang miền tần số:

1.1. Phương pháp biến đổi Fourier rời rạc (discrete Fourier tranform – DFT):

Trong phương pháp biến đổi Fourier thì đường dữ liệu ban đầu được biểu diễn bởi các đường sin và cos.

Phương pháp Fourier cho phép so sánh gián tiếp 2 chuỗi X, Y thông qua khoảng cách của hai chuỗi Xf, Yf  đã được biến đổi.

Chuỗi biến đổi *Xf* (*f* = 0, …, *n*-1) của chuỗi *Xt* (*t* = 0,…,*n*-1) được định nghĩa như sau:

*f=*0, 1, …, *n*-1

Xf là các số phức. Hàm tính độ tương tự được đề nghị cho phương pháp này là phương pháp Euclid. Phương pháp lập chỉ mục đề nghị như F-Index, ST-index

**Ưu điểm:**

+ Thu giảm số chiều của dữ liệu gốc.

+ Bảo toàn khoảng cách Euclid sau khi biến đổi dựa vào định lý Parseval và do đó thỏa mãn điều kiện chặn dưới của khoảng cách thực của dữ liệu gốc.

+ Độ phức tạp tính toán O(nlogn) cho phương pháp Fast Fourier Transform (một cải tiến của DFT).

**Nhược điểm:**

+ Chỉ áp dụng được với các loại dữ liệu có năng lượng tập trung ở một vài tần số đầu tiên (được định nghĩa là pink noise, brown noise) trong khi sẽ thiếu sót thông tin quan trọng đối với các loại dữ liệu có năng lượng phân bố đều (white noise) hoặc năng lượng tập trung ở các tần số cuối (blue noise, violet noise).

+ Tính toán phức tạp vì domain thu giảm là các số phức

+ Không tận dụng được các giải thuật và cấu trúc dữ liệu cho dạng ký tự và bit.

+ Khó giải quyết khi các chuỗi thời gian có chiều dài khác nhau.

1. 2. Phương pháp biến đổi Wavelet rời rạc DWT (Discrete Wavelet Transform - DWT):

Phương pháp thu giảm số chiều bằng biến đổi rời rạc Wavelet rời rạc do K. Chan và W. Fu đề nghị năm 1999. Phương pháp DWT cũng giống phương pháp DFT, tuy nhiên đường cơ bản của nó không phải là đường lượng giác sin hay cosin mà là đường haar như trong hình 2. Đường Haar được định nghĩa theo các công thức i ψ j như sau:

Với i = 0, …, 2j – 1.

Ngoài sử dụng đường Haar, phương pháp Wavelet có thể sử dụng các đường cơ bản khác như đường Daubechies, Coiflet, Symmlet…

Phương pháp biến đổi Wavelet rời rạc rất hiệu quả bởi vì nó mã hóa đơn giản và nhanh. Độ phức tạp của việc mã hóa này là tuyến tính. Một ưu điểm của giải thuật Wavelet là nó hỗ trợ nhiều mức phân giải. Tuy nhiên nhược điểm là chiều dài chuỗi dữ liệu ban đầu của nó phải là một số lũy thừa 2 và chiều dài của chuỗi con truy vấn cũng nên là số lũy thừa của 2 thì giải thuật mới thực hiện hiệu quả.

2. Các phương pháp xấp xỉ tuyến tính từng đoạn:

2.1. Phương pháp xấp xỉ tuyến tính từng đoạn (piecewise linear approximation - PLA):

Phương pháp xấp xỉ tuyến tính từng đoạn (piecewise linear approximation) do E. Keogh và cộng sự đề nghị. Trong phương pháp này ta sẽ biểu diễn dữ liệu ban đầu bằng chuỗi các đoạn thẳng tuyến tính. Mỗi đoạn thẳng tuyến tính nối cặp điểm ở hai đầu đoạn thẳng xấp xỉ tốt nhất (best-fit) những điểm có trong phân đoạn chuỗi thời gian đó. Các đoạn thẳng này có thể rời nhau hoặc liên tục. Cách biểu diễn này rất trực quan và nó phù h n thế nữa, việc tìm các chuỗi đoạn thẳng này c thể thực hiện trong thời gian tuyến tính.

Hàm tính khoảng cách được đề xuất là khoảng cách PLA thỏa điều kiện chặn dưới. Cấu trúc chỉ mục là R-tree.

**Ưu điểm:**

+ Xấp xỉ tốt hơn phương pháp APCA vì đoạn thẳng xấp xỉ không cần thiết phải là đoạn thẳng nằm ngang như APCA và do đó lỗi tái tạo của PLA là thấp hơn APCA.

+ Độ phức tạp biến đổi *O*(*n*).

**Nhược điểm:**

+ Tính toán khoảng cách trong cấu trúc chỉ mục khá phức tạp.

2.2. Phương pháp xấp xỉ gộp từng đoạn (piecewise aggregate approximation - PAP):

Phương pháp xấp xỉ gộp từng đoạn (piecewise aggregate approximation) do E. Keogh và cộng sự đề nghị năm 2001. Phương pháp này rất đơn giản, ta tuần tự xấp xỉ k giá trị liền kề nhau thành cùng một giá trị bằn cộng của k điểm đó. Qúa trình cứ tiếp tục như vây từ trái sang phải. Kết quả cuối cùng là đường thẳng có dạng hình bậc thang.

Hỗ trợ các phương pháp tính khoảng cách như: khoảng cách Euclid, DTW, Minkowski…Ngoài ra, một phương pháp đo khoảng cách được đề xuất như sau:

Trong đó, = { | *i* = 1,…,*N*} và = { | *i* = 1,…,*N*} là 2 chuỗi xấp xỉ của 2 chuỗi *X* và *Y*.

Phương pháp này không phụ thuộc vào cấu trúc chỉ mục cụ thể nào, có thể sử dụng F-index, R-tree…

**Ưu điểm:**

+ Thời gian tính toán nhanh và dễ dàng.

+ Hỗ trợ câu truy vấn có chiều dài khác nhau.

+ Cách biểu diễn hỗ trợ nhiều độ đo khoảng cách khác nhau.

+ Hàm khoảng cách được đề nghị có giá trị chặn dưới chặt so với dữ liệu gốc.

**Nhược điểm:**

+ Xây dựng lại chuỗi ban đầu khó và thường sinh lỗi lớn, trong khi các phương pháp trên có công thức để tái tạo lại chuỗi ban đầu với tỉ lệ lỗi nhỏ.

+ Không quan tâm đến những điểm đặc biệt khác như điểm giá trị nhỏ nhất, lớn nhất của mỗi đoạn xấp xỉ.

2.3. Phương pháp xấp xỉ hằng số từng đoạn thích nghi (adaptive piecewise constant approximation –APCA):

Phương pháp APCA giống như phương pháp PAA là xấp xỉ dữ liệu ban đầu thành những đoạn thẳng nằm ngang. Tuy nhiên, nó khác với PAA là các đoạn này ở PAA có kích thước bằng nhau, còn ở APCA thì kích thước của các đoạn là khác nhau tùy theo dữ liệu. Những vùng nào trên chuỗi thời gian có biến động nhấp nhô nhiều thì được phân thành những đoạn ngắn, còn những vùng nào ít biến động thì được phân thành những đoạn dài hơn.

Cụ thể, cho chuỗi thời gian *C* = {*c1*,…, *cn*},thì chuỗi biến đổi *C’* có dạng:

*C’* = {< *cv1*, *cr1*>,…, < *cvm*, *crm*>} với cr0 = 0

Trong đó *cvi* là giá trị trung bình cộng của những điểm dữ liệu ở đoạn thứ *i*, tức là

*cvi* = mean(*Ccri+1*,…,*Ccri*)

và cri là điểm bên phải cùng của đoạn thứ i (cho biết chiều dài của đoạn thứ *i*).

Phương pháp APCA hỗ trợ hàm tính khoảng cách Minskowki, Euclid. Ngoài ra còn có hai hàm tính khoảng cách được thiết kế riêng cho phương pháp này, một hàm xấp xỉ khoảng cách Euclid trên dữ liệu gốc nhưng không thỏa mãn chặn dưới (hỗ trợ tốt cho việc tìm kiếm xấp xỉ), một hàm còn lại thỏa mãn chặn dưới nhưng chặt hơn so với dùng các khoảng cách Minskowki hoặc Euclid (hỗ trợ tốt cho việc tìm kiếm chính xác). Cấu trúc chỉ mục được sử dụng là R-tree.

**Ưu điểm:**

+ Xấp xỉ tốt hơn phương pháp PAA.

+ Tỷ lệ lỗi khi tái tạo lại dữ liệu ban đầu nhỏ hơn phương pháp PAA.

**Nhược điểm:**

+ Độ phức tạp biến đổi *O* (*nlogn*).

3. Các phương pháp điểm quan trọng:

Những ưu điểm của phương pháp thu giảm số chiều dựa vào điểm quan trọng là (1) phù hợp với trực giác, (2) các chuỗi thời gian có chiều dài khác nhau có thể so trùng và (3) có thể thu giảm số chiều ở nhiều mức phân giải khác nhau.

3.1. Phương pháp điểm mốc:

Năm 2000, Perng và các cộng sự đã đưa ra một mô hình điểm mốc. Các điểm mốc (landmark) trong một chuỗi thời gian là các điểm có độ quan trọng lớn. Ý chính của mô hình này là sử dụng các điểm mốc để xử lý thay vì làm việc với chuỗi thời gian ban đầu. Tùy theo lãnh vực ứng dụng mà sẽ có những điểm mốc khác nhau, và định nghĩa của các điểm mốc có thể đi từ các khái niệm đơn giản (như các điểm cực đại, cực tiểu địa phương hoặc điểm uốn) đến các cấu trúc phức tạp hơn. Một điểm được gọi là điểm mốc cấp n của một đường cong nếu đạo hàm cấp n của điểm đó bằng 0. Như vậy, các điểm cực đại, cực tiểu địa phương là các điểm mốc bậc 1, còn các điểm uốn là các điểm mốc cấp 2. Càng nhiều loại điểm mốc khác nhau được dùng thì chuỗi thời gian được biểu diễn càng chính xác, tuy nhiên điều này sẽ làm cho cây chỉ mục lớn lên..

Một kỹ thuật làm nhẵn (smoothing) cũng được đưa vào để giúp loại bỏ những điểm mốc không quan trọng, chẳng hạn, một cực trị địa phương biểu diễn sự dao động nhỏ không thể quan trọng như những điểm cực trị toàn cục.

3.2. Phương pháp (perceptually important points - PIP):

Với một chuỗi thời gian T đã được chuẩn hóa, hai điểm PIP đầu tiên được chọn là điểm đầu tiên và điểm cuối cùng của chuỗi T. Điểm PIP thứ ba được chọn là điểm trong T có khoảng cách lớn nhất so với hai điểm PIP đầu tiên. Điểm PIP thứ tư được chọn là điểm trong T có khoảng cách lớn nhất so với hai điểm PIP kế cận đã chọn. Tiến trình xác định các điểm PIP tiếp tục cho đến khi số điểm PIP đạt được yêu cầu. Khoảng cách giữa một điểm trong T với hai điểm PIP kế cận đã chọn là đoạn thẳng đứng (vertical distance) từ điểm cần tính tới đường nối hai điểm PIP kế cận đã chọn.

4. Các phương pháp rời rạc hóa:

Trong phương pháp rời rạc hóa (discretization) thì từ dữ liệu ban đầu, ta sẽ chia thành những đoạn dữ liệu nhỏ hơn, rồi sau đó, tương ứng với mỗi đoạn nhỏ này ta sẽ mã hóa chúng bởi những đặc trưng của đoạn và tập hợp những đặc trưng của những đoạn nhỏ này sẽ làm thành một tràng ký hiệu biểu diễn cho dữ liệu ban đầu. Có một số phương pháp rời rạc hóa đã được đề xuất như SAX, ESAX, và iSAX. Lợi ích quan trọng của việc rời rạc hóa dữ liệu chuỗi thời gian là điều này cho phép sử dụng những cấu trúc dữ liệu và giải thuật vốn có trong lãnh vực xử lý chuỗi ký tự như cây hậu tố, kỹ thuật băm, mô hình xích Markov,v.v…

4.1. Phương pháp xấp xỉ gộp ký hiệu hóa (symbolic aggregate approximation – SAX):

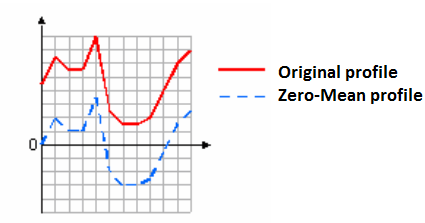
Phương pháp này dựa trên phương pháp thu giảm số chiều PAA và giả sử dữ liệu thu giảm số chiều đã đươc chuẩn hóa. SAX là quá trình ánh xạ biểu diễn PAA của chuỗi thời gian thành một chuỗi ký tự rời rạc. Gọi a là kích thước của bộ ký hiệu mà được dùng để rời rạc hóa chuỗi thời gian. Để ký hiệu hóa chuỗi thời gian chúng ta phải tìm thấy các trị (điểm ngắt).

Điểm ngắt (breakpoint): Các điểm ngắt là danh sách các số có thứ tự *B* = *β1*,…,*βα-1* sao cho diện tích dưới cung Gauss *N*(*0,1*) từ *βi* đến *βi+1* bằng 1/*α* (trong đó *β0* và *βα* lần lượt là -∞ và ∞).

Để hiểu rõ hơn về phương pháp này chúng ta cần phải thực hiện các bước sau:

\* Chuẩn hóa dữ liệu:

Ta chuẩn hóa dữ liệu về trung bình zero và độ lệch chuẩn 1, vì không có ý nghĩa khi so sánh các chuỗi thời gian có độ dịch chuyển (offset) và biên độ (amplitude) khác nhau.



Hình 6. Phương pháp chuẩn hóa trung bình zero.

\* Ký hiệu hóa chuỗi thời gian:

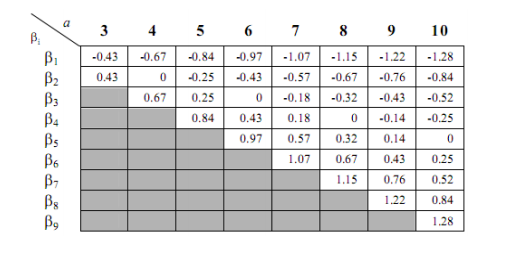
Sau khi chuyển đổi dữ liệu chuỗi thời gian về dạng PAA, ta sẽ tiếp tục chuyển đổi để rời rạc hóa dữ liệu. Dữ liệu chuỗi thời gian sau khi chuẩn hóa sẽ có sự phân bố theo hàm Gauss.

Tận dụng tính chất này, phương pháp SAX rời rạc hóa dữ liệu chuỗi thời gian thành các ký tự khác nhau. Gọi α là kích thước của tập ký tự dùng để rời rạc hóa dữ liệu.

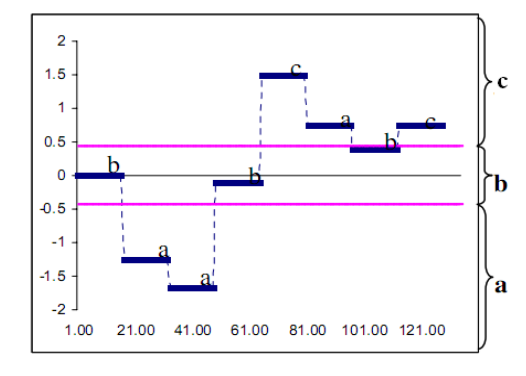
Khi đã xác định được tập các điểm ngắt, chuỗi thời gian sẽ được rời rạc hóa theo cách sau đây. Tất cả các giá trị PAA nhỏ hơn giá trị điểm ngắt nhỏ nhất ánh xạ thành ký tự “a”, tất cả giá trị PAA lớn hơn hoặc bằng giá trị điểm ngắt nhỏ nhất và nhỏ hơn giá trị điểm ngắt nhỏ thứ hai ánh xạ thành ký tự “b”, cứ tiếp tục như vậy cho đến hết chuỗi. Hình 6 minh họa ý tưởng này. Chúng ta gọi chuỗi các ký tự này (theo thứ tự rời rạc hóa) là từ ( word).

Từ (word): Một chuỗi con *T* chiều dài *n* có thể được biểu diễn thành một từ *C* = *c1c2*…*cw* như sau. Giả sử alphai là ký tự thứ *i* trong tập ký tự, tức là *alpha1* = ***a*** và *alpha2* = ***b***. Sự ánh xạ dạng biểu diễn PAA P của ***T*** thành từ ***C*** như sau:

*ci* = alphaj, nếu và chỉ nếu



Hình 7. Bảng tra tìm các giá trị β theo từng giá trị α tương ứng.



Hình 8. Một chuỗi thòi gian được biến đổi PAA rồi mã hóa thành các ký hiệu SAX. Chuỗi thời gian được mã hóa thành baabccbc

Như vậy, ta đã có dạng biểu diễn SAX của dữ liệu chuỗi thời gian.

\* Hàm tính độ tương tự:

+ Hàm tính độ tương tự trên dữ liệu gốc được tính theo hàm Euclid:

+ Sau khi chuyển đổi hai chuỗi con *Q* và *C* sang dạng PAA, ta có hàm tính khoảng cách trong không gian PAA thỏa điều kiện chặn dưới hàm Euclid:

+ Hàm tính độ tương tự (thỏa điều kiện chặn dưới) trên không gian các ký tự:

+ Hàm dist () ở công thức trên có thể được tính bằng cách tra Bảng 1 sau:

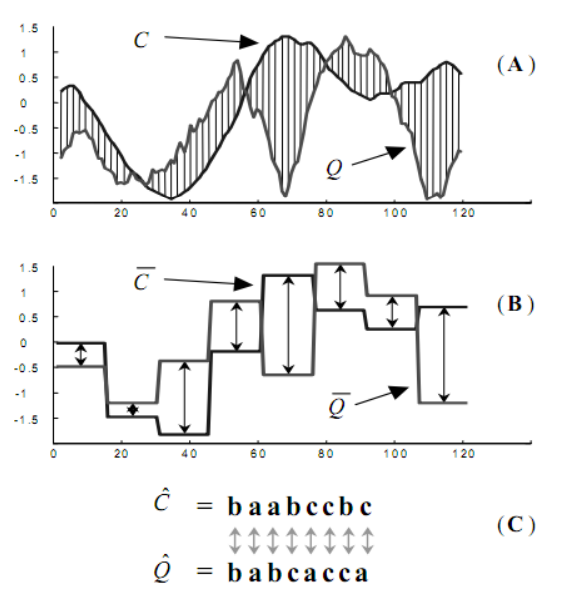
Ví dụ, dist (*a*,*b*) = 0 và dist(*a*, *c*) = 0.67

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | a | b | c | d |
| a | 0 | 0 | 0.67 | 1.34 |
| b | 0 | 0 | 0 | 0.67 |
| c | 0.67 | 0 | 0 | 0 |
| d | 1.34 | 0.67 | 0 | 0 |

Bảng . Bảng tra cứu giá trị hàm dist giữa hai ký tự trường hợp *α* = 4.

Giá trị ô cell(*r, c*) ở hàng *r*, cột *c* trong bảng tra cứu trên có thể được tính nhờ vào công thức sau:

Với giá trị α cho trước, bảng tra cứu chỉ được tính một lần và lưu trữ lại để tra cứu khoảng cách một cách nhanh chóng.



Hình 9. 6 Minh họa 3 dạng biểu diễn chuỗi thời gian và các hàm tính khoảng cách tương ứng. A) Chuỗi thời gian ban đầu và hàm tính khoảng cách Euclid. B) Dạng biểu diễn PAA của chuỗi thời gian và hàm tính khoảng cách DR. C) Dạng biểu diễn SAX và hàm tính khoảng cách MINDIST.

Trong nhiều trường hợp ta chỉ cần quan tâm đặc trưng của chuỗi thời gian thì phương pháp mã hóa SAX này rất thích hợp và hiện tại đang được ứng dụng nhiều. Ngoài ra, trong phương pháp mã hóa SAX, ta có thể sử dụng những cấu trúc dữ liệu và giải thuật có sẵn về xử lý chuỗi ký tự như trong lĩnh vực xử lý dòng ký tự và xử lý trình tự sinh học. Tuy nhiên, nhược điểm chính của SAX là cách định nghĩa những đặc trưng cũng như phương pháp này không hỗ trợ tốt việc tính khoảng cách Euclid và dữ liệu chuỗi thời gian được giả định là phải thỏa phân bố xác xuất Gauss.

4.2. Phương pháp Extended SAX (ESAX):

Năm 2006, Lkhagva và các cộng sự đã đề xuất một phương pháp rời rạc hóa là sự mở rộng của SAX, gọi là ESAX (Extended SAX) để đem lại một cách rời rạc hóa hữu hiệu hơn SAX khi áp dụng vào dữ liệu chuỗi thời gian trong lãnh vực tài chính. Do SAX dựa vào cách biểu diễn PAA mà trong đó sự thu giảm số chiều là sử dụng các giá trị trung bình của các chuỗi con được phân đoạn với độ dài bằng nhau, nên có khả năng bị mất đi một số mẫu quan trọng trong dữ liệu chuỗi thời gian tài chính. Để khắc phục nhược điểm này, Lkhagva và các cộng sự vẫn dựa vào PAA nhưng sau khi đạt được giá trị trung bình của mỗi phân đoạn, họ đưa thêm vào giá trị nhỏ nhất và lớn nhất của mỗi phân đoạn. Như vậy mỗi phân đoạn được diễn tả bằng một bộ ba , và sau đó bộ ba này sẽ được mã hóa bằng 3 ký hiệu thay vì chỉ một ký hiệu như trong phương pháp rời rạc hóa SAX.

4.3. Phương pháp IndexableSAX (iSAX):

Phương pháp rời rạc hóa iSAX (indexable symbolic aggregate approximation) được Shieh và Keogh đưa ra năm 2008. Phương pháp này là sự mở rộng của phương pháp SAX mà thay vì mã hóa mỗi phân đoạn chuỗi thời gian thành một ký hiệu như là chữ hay số nguyên, iSAX mã hóa nó thành số nhị phân, thí dụ “00”, “01”,”10”,”11”.

Với cách này iSAX có thể mã hóa chuỗi thời gian thành chuỗi bit và nhờ đó tiết kiệm được rất nhiều chỗ bộ nhớ lưu trữ. Ngoài ra, iSAX còn hỗ trợ khả năng biểu diễn đa mức phân giải (multi-resolution) dựa trên khả năng đa mức phân giải của các mức rời rạc hóa. Bằng việc kết hợp với một cấu trúc chỉ mục cây phân cấp (hierarchical tree) với phương pháp iSAX, ta có thể truy vấn nhanh trên cơ sở dữ liệu chuỗi thời gian với kích thước lớn lên đến hàng terabyte.

CHƯƠNG 3 – CƠ SỞ LÝ THUYẾT

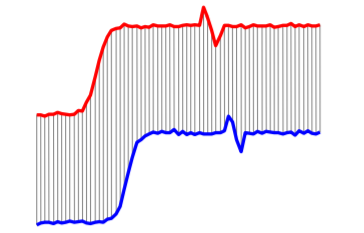
Đối với bài toán tìm kiếm tương tự trên dữ liệu chuỗi thời gian thì dữ liệu được biểu diễn thành những dãy số thực, thí dụ T = t1,…tn. Cho hai chuỗi thời gian X = x1, x2,…,xn và Y = y1,y2,…,yn. Ta cần phải tính độ tương tự SIM(X, Y) của hai chuỗi thời gian này.

Đã có nhiều độ đo tương tự đã được sử dụng. Việc chọn một độ đo tương tự là tùy thuộc rất nhiều vào miền ứng dụng và trong nhiều trường hợp thì một độ đo thuộc chuẩn Lp đơn giản như độ đo Euclid là đủ tốt để dùng. Tuy nhiên trong nhiều trường hợp thì độ đo Euclid tỏ ra quá thô và cứng nhắc vì không thích nghi được với những phép biến đổi như tịnh tiến (shifting), co giãn biên độ (scaling) hay xoắn trục thời gian (time warping).

I. Độ đo Euclid:

Trong [toán học](https://vi.wikipedia.org/wiki/To%C3%A1n_h%E1%BB%8Dc), khoảng cách Euclid là [khoảng cách](https://vi.wikipedia.org/wiki/Kho%E1%BA%A3ng_c%C3%A1ch) giữa hai điểm được tính bằng [công thức Pytago](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%8Bnh_l%C3%BD_Pytago).

Cho hai chuỗi thời gian A = a1…an và B = b1…bn độ đo khoảng cách Euclid giữa hai chuỗi thời gian này được cho bởi công thức:



Hình 10. Tính khoảng cách theo Euclid

Độ đo khoảng cách Euclid có ưu điểm là dễ hiểu, dễ tính toán, dễ mở rộng cho nhiều bài toán khác nhau. Tuy nhiên nó cũng có một số nhược điểm là do phương pháp này tính toán dựa trên giá trị nên đối với các trường hợp tính chất của hai mẫu là giống nhau nhưng giá trị khác nhau (có  đường căn bản khác nhau hay có biên độ dao động khác nhau) thì khoảng cách hai mẫu sẽ rất khác nhau.

II. Độ đo xoắn thời gian động (Dynamic time warping - DTW):

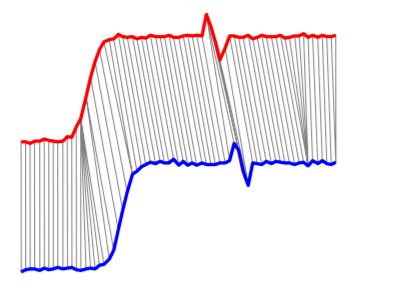
Phương pháp xoắn thời gian động (Dynamic Time Warping - DTW) được đề xuất bởi Bernt và Clifford, 1994.

Dynamic Time Warping là một kỹ thuật tìm ra sự liên kết tối ưu giữa hai chuỗi thời gian nếu một chuỗi thời gian có thể "cong" không tuyến tính bằng cách kéo dài hoặc co lại nó dọc theo trục thời gian của nó.

Sự cong giữa hai chuỗi thời gian này có thể được sử dụng để tìm các vùng tương ứng giữa hai chuỗi thời gian hoặc để xác định sự tương tự giữa hai chuỗi thời gian.

DTW tạo điều kiện cho việc khám phá các mô hình linh hoạt từ chuỗi thời gian.

DTW được sử dụng trong quá trình ghi lại giọng nói để xác định xem hai dạng sóng mô tả cùng một cụm từ nói hay không.



Hình . tính khoảng cách theo DWT.

Công thức để tính khoảng cách theo DWT được phát biểu như sau:

Cho hai dữ liệu chuỗi thời gian A và B:

+ Biên độ của A biểu diễn ở ma trận sẽ có công thức:

+ Biên độ của B biểu diễn ở ma trận sẽ có công thức:

+ Biên độ giá trị của A và B:

Để hiểu rõ hơn về công thức trên chúng ta sẽ có một ví dụ như sau:

Với hai chuỗi thời gian:

A = 1, 3, 4, 9, 8, 2, 1, 5, 7, 3.

B = 1, 6, 2, 3, 0, 9, 4, 3, 6, 3.

Áp dụng công thức tính cho vị trí A:

A(0,0): |1 - 1| + D(0 – 1, 0) = 0

A(1,0): |3 - 1| + D(1 – 1, 0) = 2

A(2,0): |4 - 1| + D(2 – 1, 0) = 5

.

.

.

A(9,0): |3 - 1| + D(9 – 1, 0) = 33

Áp dụng công thức tính cho vị trí B:

B(0,0): |1 - 1| + D(0, 0 – 1) = 0

B(0,1): |1 – 1| + D(0, 1 – 1) = 5

B(0,2): |1 – 6| + D(0, 2 – 1) = 6

.

.

.

B(0,9): |1 – 3| + D(0, 9 – 1) = 29

Vị trí biên độ A và B:

A(1,1): |3 – 6| + min(0, 5, 2) = 3

A(2,1): |4 – 6| + min(2, 3, 5) = 4

Như vậy theo như 3 công thức trên chúng ta sẽ điền được hết dữ liệu vào ma trận và có kết quả như sau:

**TÀI LIỆU THAM KHẢO**

1. Time-Series Data Mining - PHILIPPE ESLING and CARLOS AGON, Institut de Recherche et Coordination Acoustique/Musique (IRCAM).
2. AN OVERVIEW OF SIMILARITY SEARCH IN TIME SERIES DATA - Dương Tuấn Anh.
3. Time Series Model - www.ibm.com.