บทที่ 3

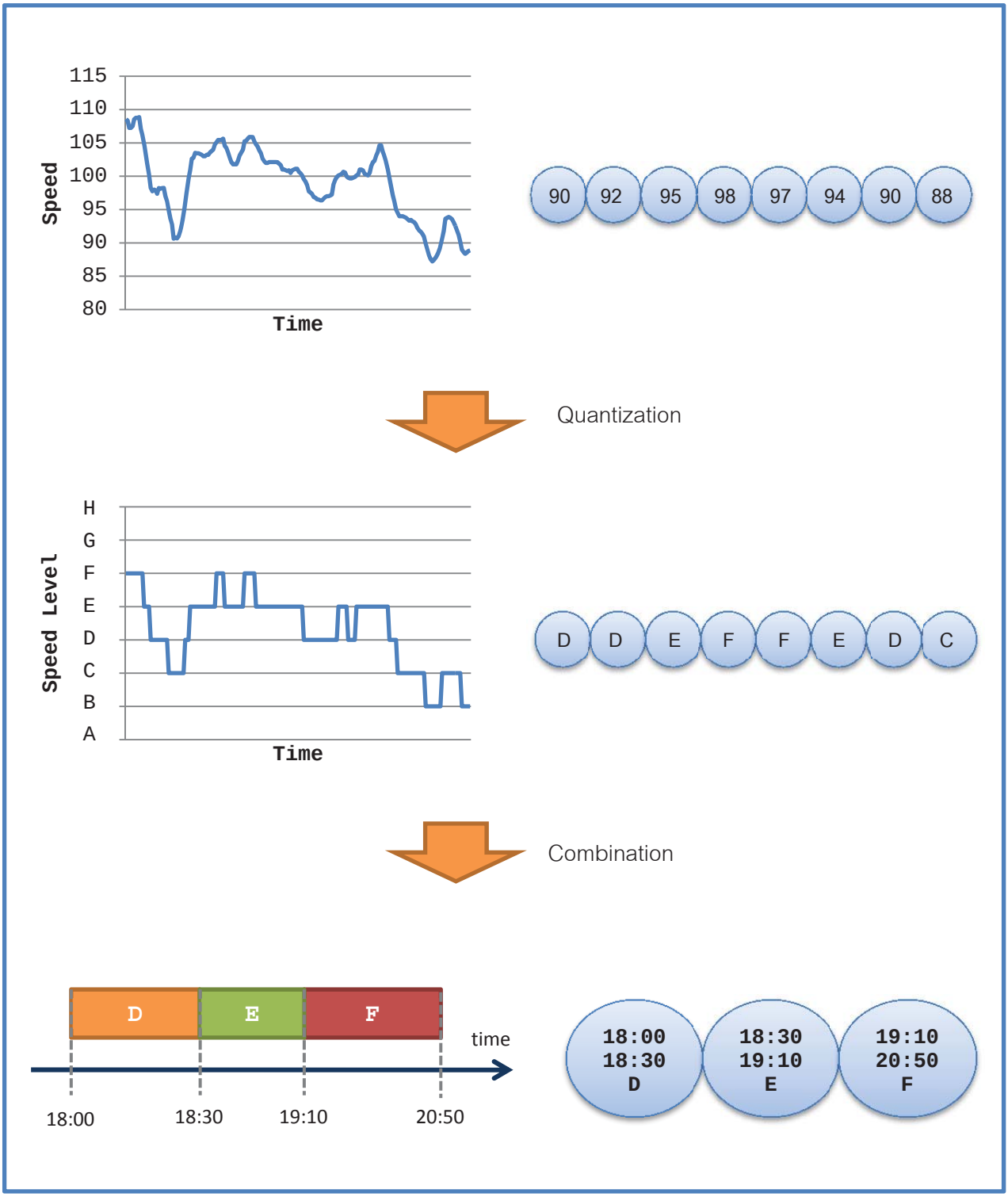
วิธีการที่เสนอ

จุดประสงค์หลักของวิทยานิพนธ์นี้คือการคาดการณ์แนวโน้ม (trend) ข้อมูลความเร็วรถที่จะเกิดขึ้นในอนาคต โดยใช้ขั้นตอนวิธี (algorithm) ของ K-Nearest-Neighbor (KNN) ซึ่งหัวใจสำคัญของขั้นตอนวิธี KNN คือวิธีการวัดความคล้าย (similarity measure) แบบข้อมูลของข้อมูลความเร็วรถที่เกิดขึ้น วิธีการวัดความคล้ายที่เหมาะสมกับลักษณะของข้อมูลความเร็วรถ และมีประสิทธิภาพดี สามารถนำมาใช้กับแบบข้อมูลของข้อมูลความเร็วรถได้ คือ การวัดความคล้ายแบบข้อมูลของอนุกรมเวลา (time series) ที่เรียกว่า Dynamic Time Warping (DTW) แต่ข้อเสียของวิธีการวัดความคล้ายนี้คือใช้เวลาในการคำนวณมาก

วิทยานิพนธ์นี้จึงเสนอให้ทำการแปลงข้อมูลความเร็วรถที่เป็นข้อมูลอนุกรมเวลา (time series) ให้อยู่ในรูปแบบของข้อมูลลำดับ (sequence) โดยนำไปผ่านกระบวนการ quantization เพื่อแปลงให้ข้อมูลแบบมาตราอัตราส่วน (ratio scale) ให้กลายมาเป็นข้อมูลแบบมาตราเรียงลำดับ (ordinal scale) หลังจากนั้นจะทำการรวมระดับความเร็วเดียวกันที่เกิดขึ้นซ้ำกันและอยู่ติดกันเข้าด้วยกัน โดยในหนึ่งหน่วยข้อมูลจะประกอบด้วย เวลาเริ่ม, เวลาสิ้นสุด และ ระดับความเร็ว เพื่อลดจำนวนข้อมูลและเวลาในการคำนวณความคล้าย ดังภาพที่ 7

ภาพที่ 7

ขั้นตอนการแปลงข้อมูลความเร็วรถเป็นระดับความเร็วรถ และทำการยุบรวมข้อมูลระดับความเร็วรถที่ซ้ำกัน



ข้อมูลลำดับระดับความเร็วรถที่ได้จากการแปลงจากข้อมูลความเร็วรถ จะมีลักษณะที่แตกต่างจากข้อมูลลำดับโดยทั่วไป คือ

1. มีระดับความเร็วเกิดขึ้นซ้ำๆ กัน เพราะข้อมูลความเร็วก่อนการแปลงไม่มีการเปลี่ยนแปลงขึ้นลงอย่างฉับพลัน จะมีการคงที่เป็นระยะเวลาหนึ่ง แล้วจึงจะเริ่มมีการเปลี่ยนแปลง
2. ตัวอักษรแต่ละตัวจะแสดงถึงระดับความเร็วรถ ซึ่งเป็นข้อมูลประเภทมาตราเรียงลำดับ (ordinal scale) เมื่อนำข้อมูลชนิดนี้มาเทียบความแตกต่างกัน จะมีความแตกต่างที่ไม่เท่ากัน ซึ่งต่างจากข้อมูลประเภทมาตรานามบัญญัติ (nominal scale) ที่จะมีความต่างเท่าเทียมกัน
3. การเกิดขึ้นของข้อมูลระดับความเร็วรถมีลักษณะไม่แน่นอน สามารถที่จะเกิดการเปลี่ยนแปลง ก่อน หรือหลังเวลาที่เคยเกิดขึ้นเป็นประจำในอดีต
4. ระยะเวลาในการคงอยู่ของระดับความเร็วรถนั้นไม่แน่นอน อาจจะคงอยู่นานขึ้นหรือเร็วกว่าเดิม เมื่อเทียบจากข้อมูลในอดีตที่เคยเก็บไว้

วิธีการวัดความคล้ายที่ได้มีการคิดค้นขึ้นมาก่อนหน้านี้มีข้อจำกัดต่างๆ ดังนี้

1. Edit Distance ให้น้ำหนักความแตกต่างระหว่างข้อมูลที่แตกต่างเท่ากัน สำหรับข้อมูลระดับความเร็วรถเป็นข้อมูลมาตราเรียงลำดับ (ordinal scale) ไม่ใช่ข้อมูลแบบมาตรานามบัญญัติ (nominal scale) จึงไม่สามารถที่จะทำการเปรียบเทียบได้อย่างถูกต้องมากนัก
2. Edit Distance นั้นไม่เหมาะสมในการนำมาวัดความคล้ายของลำดับระดับความเร็ว เนื่องจากการเกิดขึ้นของข้อมูลนั้นไม่แน่นอน และมีเวลาที่เกิดสั้นยาวที่ไม่เท่ากัน เช่น
3. Edit Distance นั้นไม่สามารถที่จะทำการวัดและเปรียบเทียบข้อมูลที่มีหลายมิติ (Domain) พร้อมกันได้ เช่นข้อมูลที่หนึ่งหน่วยข้อมูลประกอบด้วย มิติของเวลาและมิติของระดับความเร็ว Edit Distance ไม่สามารถทำการวัดและเปรียบเทียบความคล้ายได้

4. วิธีการของ Timed String นั้นนำ cost ของ เวลามารวมกับ cost ของ Event โดยที่ cost ของเวลา และ Event นั้น ไม่ได้มีหน่วยเดียวกัน
5. วิธีการของ Timed String นำ cost ของเวลามารวมกับ cost ของ Event โดยทำการถ่วงน้ำหนัก สำหรับสภาพจราจร ไม่สามารถบอกได้อย่างชัดเจนว่า จะทำการกำหนดน้ำหนักของความสำคัญของ เวลา และ Event เป็นค่าเท่าไรจึงจะเหมาะสม

3.1 วิธีการที่เสนอ

จากลักษณะของข้อมูลระดับความเร็วรถ (speed level) ที่ได้หลังจากการทำ quantization แล้ว ในหนึ่งหน่วยของข้อมูลจะประกอบด้วยเวลา และระดับความเร็ว สามารถเขียนให้อยู่ในรูปแบบ (ระดับความเร็ว, เวลา) ตัวอย่างเช่น กำหนดให้ A และ B แทนระดับความเร็วรถที่เกิดขึ้น ณ เวลาต่างๆ โดยที่ A เป็นความเร็วรถที่ช้ากว่า B เราสามารถเขียนแสดงระดับความเร็วรถที่เวลาต่างๆ ได้ดังนี้

```
(A, 18.00), (A, 18.10), (A, 18.20), (B, 18.30) , (B, 18.40)
```

โดยที่ตัวอักษรตัวแรกนั้นจะแสดงถึงระดับความเร็วที่เกิดขึ้น (A ช้ากว่า B) ตัวเลขถัดจากระดับความเร็วคือเวลาที่เกิดขึ้น จะสังเกตได้ว่าระดับความเร็วที่เกิดขึ้นติดและซ้ำกัน สามารถที่จะเขียนรวมให้อยู่ด้วยกันได้ในรูปแบบ (ระดับความเร็ว, เวลาเริ่ม, ระยะเวลา) ดังนี้

```
(A, 18.00, 30), (B, 18.30, 10)
```

โดยที่ตัวอักษรตัวแรกนั้นจะแสดงถึงระดับความเร็วที่เกิดขึ้น (A ช้ากว่า B) ตัวเลขถัดจากระดับความเร็วคือเวลาที่เกิดขึ้น และตัวเลขตัวสุดท้ายคือ ระยะเวลาที่คงอยู่ในระดับความเร็วนี้ในหน่วยนาที วิธีการนี้สามารถที่จะลดขนาดของข้อมูลได้โดยข้อมูลไม่เกิดการสูญหาย

จากลักษณะของข้อมูลระดับความเร็วรถสามารถที่จะนิยาม และจัดให้อยู่ในรูปแบบของลำดับได้ดังนี้ กำหนดให้ $\lambda = \langle l, s, d \rangle$ คือข้อมูลระดับความเร็วรถ โดยประกอบด้วยองค์ประกอบย่อย 3 อย่าง คือ ระดับความเร็วรถ (l) , เวลาเริ่ม (s) และ ระยะเวลา (d) ดังนั้น เซตของข้อมูลระดับความเร็วรถสามารถนิยามได้ดังนี้

$$\Lambda = \{\lambda = \langle l, s, d \rangle \mid l \in \{H, M, L\};\ s, d \in \mathcal{R};\ d > 0\} \qquad (3-1)$$

ข้อมูลระดับความเร็วรถที่ไม่มีขนาดจะเรียกว่าข้อมูลพร่องระดับความเร็วรถ โดยข้อมูลพิเศษตัวนี้จะมีบทบาทสำคัญสำหรับขั้นตอนวิธี (algorithm) ที่ใช้วัดความคล้าย องค์ประกอบของข้อมูลพร่องระดับความเร็วรถจะคล้ายกับข้อมูลระดับความเร็วรถ แต่จะไม่มีระดับความเร็วรถ

กำหนดให้ $\varepsilon = \langle \phi, \mathrm{s}, \mathrm{d} \rangle$ คือ ข้อมูลพร่องระดับความเร็วรถ ประกอบด้วย เวลาเริ่ม (s) และ ระยะเวลา (d) แต่ระดับความเร็วรถจะไม่มี และจะเขียนแทนด้วยสัญลักษณ์ ϕ ดังนั้นเซตของ ข้อมูลพร่องระดับความเร็วรถ สามารถนิยามได้ดังนี้

$$\Theta = \{\varepsilon = \langle \phi, \mathrm{s}, \mathrm{d} \rangle \mid \mathrm{s}, \mathrm{d} \in \mathcal{R}\} \qquad (3-2)$$

คุณสมบัติของข้อมูลระดับความเร็วรถที่ไม่มีขนาด หรือที่เรียกว่าข้อมูลพร่องระดับความเร็วรถ เมื่อนำไปต่อกับลำดับของข้อมูลระดับความเร็วรถ จะทำให้ลำดับนั้นมีขนาดความยาวเท่าเดิมเช่น

$$\| \lambda_1 \| = \| \varepsilon_0 \lambda_1 \| = \| \lambda_1 \varepsilon_0 \| = \| \varepsilon_0 \lambda_1 \varepsilon_1 \| \qquad (3-3)$$

จากนิยามของข้อมูลระดับความเร็วรถ และ ข้อมูลพร่องระดับความเร็วรถ สามารถที่จะนิยามลำดับข้อมูลความเร็วรถ $\Psi = [\lambda_1, \ldots, \lambda_\mathrm{i}, \ldots, \lambda_\mathrm{n}]$ คือ ลำดับของข้อมูลระดับความเร็วรถที่นำมาเรียงต่อกันตามลำดับเวลา เมื่อ $\lambda_\mathrm{i} \in \{\Lambda \cup \Theta\}$ และ n คือความยาวของ Ψ ซึ่งจะต้องเป็นไปตามเงื่อนไข $l_i \neq l_{i+1}$ และ $\mathrm{s_i} < \mathrm{s_{i+1}}$ สำหรับทุกๆ λ_i และ $\lambda_{\mathrm{i+1}}$

จากนิยามของลำดับข้อมูลความเร็วรถ จะเห็นได้ข้อมูลระดับความเร็วรถที่เกิดขึ้นซ้ำและติดกัน สามารถที่จะรวมให้อยู่ด้วยกันได้โดยไม่มีการสูญหายของข้อมูล และช่วยให้คำนวณความคล้ายได้อย่างรวดเร็วและประหยัดพื้นที่ในการเก็บข้อมูล

วิธีการวัดความคล้ายของลำดับระดับความเร็วรถ

สิ่งสำคัญของวิธีการวัดความคล้ายคือ operation ที่ใช้ในการแปลง (transform) ลำดับของข้อมูลระดับความเร็วรถให้เหมือนกัน สำหรับ operation ที่จะใช้มี 3 operations คือ matching, substitution และ deletion กำหนดให้ $X = [x_1, \dots, x_i, \dots, x_m]$ และ $Y = [y_1, \dots, y_j, \dots, y_n]$ คือลำดับของข้อมูลระดับความเร็วรถ โดยมีความยาว m และ n ตามลำดับ

1. Delete operation

Operation นี้จะทำการลบข้อมูลระดับความเร็วรถออกจาก X หรือ Y และทำการต่อข้อมูลที่ขาดเข้าด้วยกัน ดังนั้น cost ที่ได้จาก operation นี้ก็คือ cost ที่เกิดจากการลบระดับความเร็วและระยะเวลาออกจากลำดับข้อมูล

2. Substitute operation

Operation นี้จะทำการแทนที่ระดับความเร็วรถให้เป็นระดับเดียวกัน และทำให้ขนาดของระยะเวลานั้นเท่ากัน cost ที่เกิดขึ้นจะเกิดจากการแทนที่ระดับความเร็วรถและการตัดระยะเวลาที่เกินออก ถ้าทำการแทนที่ความเร็วรถด้วยระดับที่ใกล้เคียงกัน cost ที่เกิดขึ้นจะน้อยกว่าการแทนที่ด้วยระดับความเร็วรถที่ห่างกัน

3. Match operation

Operation นี้จะเกิดขึ้นในกรณีที่ ระดับความเร็วของข้อมูลเป็นระดับเดียวกัน และ cost ของ operation จะเกิดจากความต่างของระยะเวลา ถ้าหากระยะเวลาเท่ากันความต่างก็จะมีค่าเท่ากับ 0

จาก operation ที่ใช้ในการจัดเรียง (align) ลำดับของข้อมูลระดับความเร็วรถสามารถที่จะนิยามการวัดความคล้ายของลำดับระดับความเร็วรถได้ดังนี้ กำหนดให้ $\eta = \{del, sub, match\}$ เป็นเซตของ operation ที่ใช้ในการจัดเรียง และ $Y = [\upsilon_1, \dots, \upsilon_i, \dots, \upsilon_n]$ เป็นลำดับของ operation ที่ใช้ในการจัดเรียงลำดับของข้อมูลระดับความเร็วรถ จาก X เป็น Y ผลรวมของ cost ที่มาจาก operation คือ $C(Y) = \sum_1^n \xi(\upsilon_i)$ เมื่อ $\xi(\upsilon_i)$ คือ cost ของแต่ละ operation ในลำดับของ Y ดังนั้น cost ที่ใช้แปลงจาก X เป็น Y ที่มีค่าต่ำที่สุดคือ

$$S(X,Y) = \left\{ C(Y) \;\middle|\; \begin{array}{l} Y \text{ คือลำดับของ operation} \\ \text{ที่ใช้ในการแปลงจาก } X \text{ เป็น } Y \end{array} \right\} \qquad (3-4)$$

จากนิยามและคำจำกัดความของวิธีการวัดความคล้ายของลำดับของข้อมูลระดับความเร็วรถ สามารถที่จะกำหนดเงื่อนไขของ recurrent relation ที่ใช้ในกระบวนการหาความคล้ายของ X และ Y โดยใช้ Dynamic programming ได้ดังนี้

$$\begin{aligned} d(0,0) &= \sigma(x_1, y_1) \\ d(i,0) &= d(i-1,0) + \delta(x_i \to \varepsilon_0) \\ d(0,j) &= d(0,j-1) + \delta(\omega_0 \to y_j) \\ d(i,j) &= \min \begin{cases} d(i-1,0) + \delta(x_i \to \varepsilon_j), \\ d(0,j-1) + \delta(\omega_i \to y_j), \\ d(i-1,j-1) + \delta(x_i, y_j) \end{cases} \end{aligned} \qquad (3-5)$$

เมื่อ

$$\sigma(x_1, y_1) = \begin{cases} \dfrac{start(x_1) - start(y_1)}{\Omega(Y)} & \text{เมื่อ } y_1 \text{ เกิดก่อน } x_1 \\ \dfrac{start(y_1) - start(x_1)}{\Omega(X)} & \text{เมื่อ } x_1 \text{ เกิดก่อน } y_1 \end{cases} \qquad (3-6)$$

และ

$$\delta(x_i, y_j) = \sqrt{u(x_i, y_j)^2 + v(x_i, y_j)^2} \qquad (3-7)$$

ในช่อง $d(0,0)$ จะเก็บ cost ของการปรับเวลาเริ่มของทั้งสองลำดับให้เท่ากัน ถ้า X เกิดขึ้นก่อน Y ระยะเวลาที่ทำการปรับจะถูกทำการ normalized ด้วยระยะเวลาทั้งหมดของ $\Omega(X)$ ถ้า Y เกิดขึ้นก่อน X ระยะเวลาที่ทำการปรับจะถูกทำการ normalized ด้วยระยะเวลาทั้งหมดของ $\Omega(Y)$

ฟังก์ชั่น $\delta(x,y)$ คือฟังก์ชั่นคำนวณความคล้ายของข้อมูลระดับความเร็วรถ โดยประกอบด้วยความแตกต่างของระดับความเร็วรถ ซึ่งเขียนแทนด้วย $u(a,b)$ และความแตกต่างของระยะเวลา ซึ่งเขียนแทนด้วย $v(a,b)$ และใช้วิธีการหาระยะของข้อมูลใน 2 มิติของ Euclidian มาทำการรวมค่าทั้งสองที่อยู่ต่างมิติ (domain) กันเข้าด้วยกัน โดยที่ค่าทั้งสองนี้จะเป็นค่าที่ถูกทำการ normalized เรียบร้อยแล้ว

1. ความต่างของระดับความเร็ว

ข้อมูลระดับความเร็วเป็นข้อมูลแบบมาตราเรียงลำดับ (ordinal scale) ซึ่งแปลงมาจากความเร็วของรถที่แล่นอยู่ ถ้าหากเปรียบเทียบระดับความติดขัดเดียวกัน ค่าความแตกต่างของระดับความติดขัดจะมีค่าเท่ากับศูนย์ และสำหรับการเปรียบเทียบระดับที่ต่างกัน ก็จะได้ค่าความต่างกันตามระดับชั้นที่ห่างกัน โดยจะทำการ normalize ค่าความต่างของระดับความเร็ว ด้วยการนำมาหารด้วยจำนวนของระดับทั้งหมด

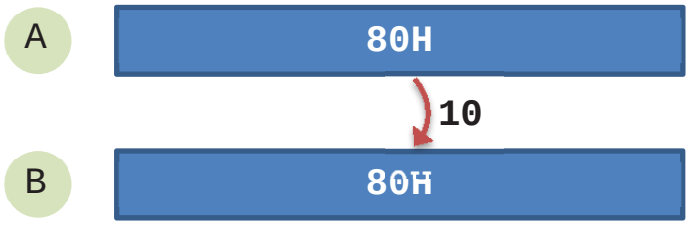
2. ความแตกต่างของระยะเวลา

สูตรของการหาความแตกต่างของระยะเวลากำหนดไว้ดังนี้

$$v(x_i, y_j) = \begin{cases} \dfrac{\text{duration}(x_i) - \text{duration}(y_j)}{\Omega(X)} & \text{เมื่อ } \text{duration}(x_i) > \text{duration}(y_j) \\ \dfrac{\text{duration}(y_j) - \text{duration}(x_i)}{\Omega(Y)} & \text{เมื่อ } \text{duration}(y_j) > \text{duration}(x_i) \end{cases} \quad (3-8)$$

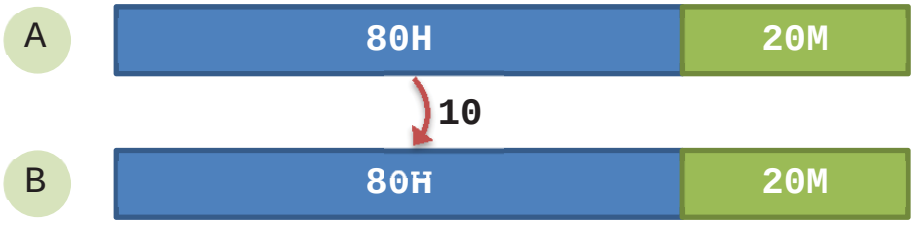
โดยค่าความแตกต่างของเวลานี้จะถูก normalized ด้วยระยะเวลาทั้งหมดของ X เมื่อ $\text{duration}(x_i) > \text{duration}(y_j)$ หรือระยะเวลาทั้งหมดของ Y เมื่อ $\text{duration}(y_j) > \text{duration}(x_i)$

3.2 พิสูจน์ (proof)

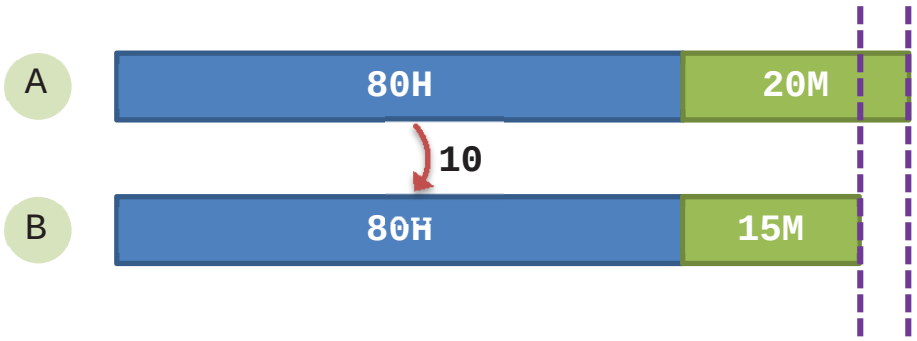


ถ้าหากมีลำดับข้อมูลลำดับระดับความเร็วรถ A และ B โดยที่ A และ B เป็นลำดับข้อมูลความเร็วที่มีระดับความเร็ว H ขนาด 80 นาทีเท่ากัน โดยเกิดจากการ edit ด้วย operation โดยสมมุติว่า cost ที่เกิดขึ้นจากการ edit นี้เท่ากับ 10

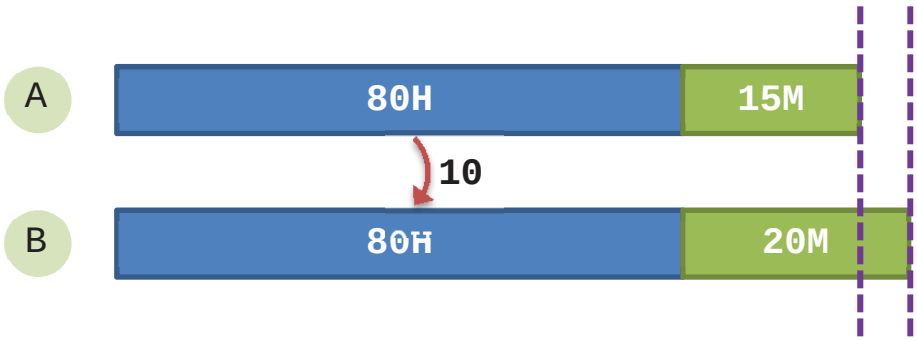
1. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ A และ B ค่า cost ที่ทำให้ A และ B เหมือนกันก็ยังคงเป็น 10 เท่าเดิม



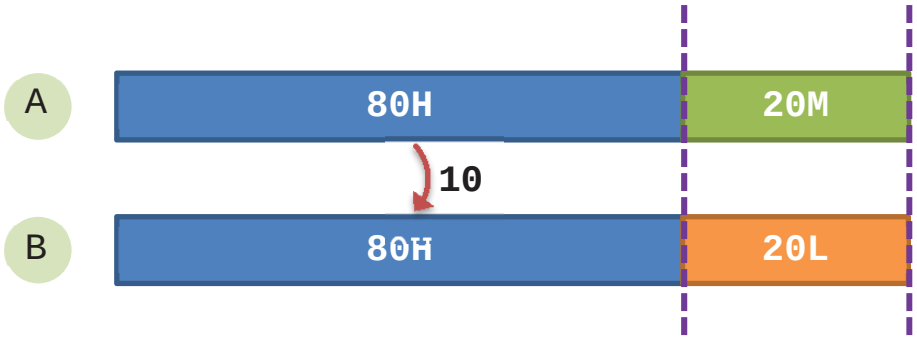
2. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ A และ ข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 15 นาที ไปทำการต่อท้ายของ B ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดเวลาที่มีขนาดไม่เท่ากันออก



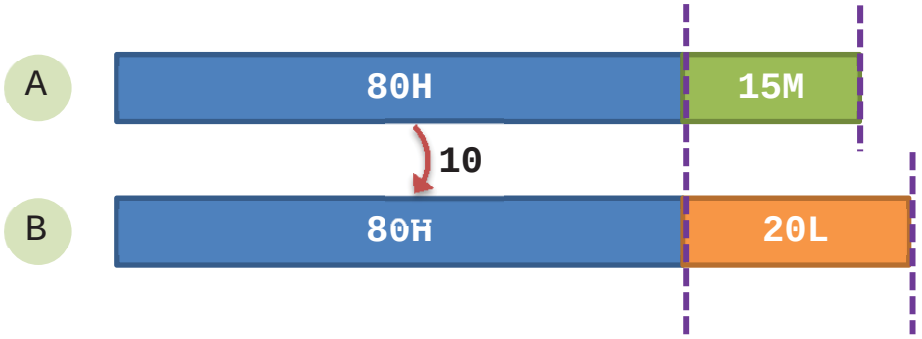
3. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 15 นาที ไปทำการต่อท้ายของ A และ ข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ B ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดเวลาที่มีขนาดไม่เท่ากันออก



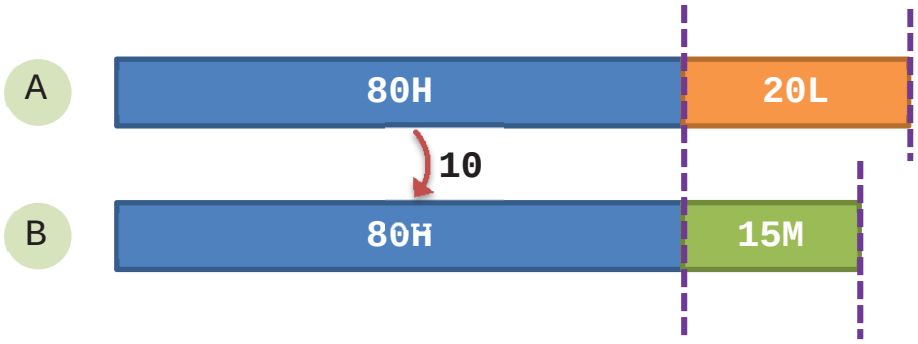
4. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ A และ ข้อมูลระดับความเร็วรถที่มีระดับความเร็ว L ขนาด 20 นาที ไปทำการต่อท้ายของ B ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดข้อมูลที่เพิ่มมาทั้งสองอันนี้ออกไป



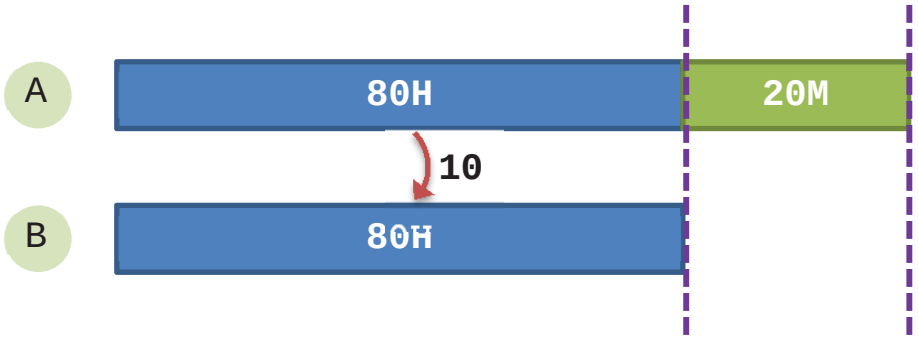
5. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 15 นาที ไปทำการต่อท้ายของ A และ ข้อมูลระดับความเร็วรถที่มีระดับความเร็ว L ขนาด 20 นาที ไปทำการต่อท้ายของ B ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดข้อมูลที่เพิ่มมาทั้งสองอันนี้ออกไป



6. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ A และ ข้อมูลระดับความเร็วรถที่มีระดับความเร็ว L ขนาด 15 นาที ไปทำการต่อท้ายของ B ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดข้อมูลที่เพิ่มมาทั้งสองอันนี้ออกไป



7. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ A ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดข้อมูลที่เพิ่มมานี้ออกไป



8. ถ้าหากนำข้อมูลระดับความเร็วรถที่มีระดับความเร็ว M ขนาด 20 นาที ไปทำการต่อท้ายของ B ค่า cost ที่ทำให้ A และ B เหมือนกันจะต้องถูกเพิ่มด้วย cost ของการตัดข้อมูลที่เพิ่มมานี้ออกไป