# บทที่ 3
# ขั้นตอนวิธีการประมวลผลข้อมูลเบื้องต้น

## 3.1 ขั้นตอนการเก็บรวบรวมข้อมูล

เนื่องจากรูปแบบของล็อกไฟล์จะแตกต่างกันไปตามโปรแกรมเว็บเซิร์ฟเวอร์ที่ใช้และล็อกไฟล์โดยส่วนมากจะไม่เก็บค่ารหัสประจำตัวผู้ชมซึ่งทำให้เกิดปัญหาในการจัดกลุ่มข้อมูลการเยี่ยมชมของผู้ชม

เพื่อแก้ไขปัญหานี้เราจึงสร้างล็อกไฟล์ที่มีรหัสประจำตัวผู้ชมขึ้นโดยนำสคริปต์แทรกเข้าไปในเว็บเพจ เมื่อผู้ชมเยี่ยมชมเว็บเพจแล้วข้อมูลการเยี่ยมชมของผู้ชมจะถูกส่งกลับไปยังเว็บเซิร์ฟเวอร์โดยสคริปต์ที่แทรกอยู่ในเว็บเพจโดยอัตโนมัติ และข้อมูลดังกล่าวจะถูกบันทึกลงในฐานข้อมูลต่อไป ข้อมูลการเยี่ยมชมที่ได้จากสคริปต์มีส่วนประกอบดังนี้

- ไอพีแอดเดรส
- รหัสประจำตัวผู้ชม
- เวลาที่เยี่ยมชม
- URL ของเว็บเพจที่เยี่ยมชม
- URL ของเว็บเพจที่ใช้อ้างอิงมายังเว็บเพจที่เยี่ยมชม
- และ โปรแกรมเว็บเบราเซอร์ที่ใช้

โดยที่รหัสประจำตัวของผู้ชมแต่ละคนต้องเป็นค่าที่เฉพาะเจาะจงและจะแทรกอยู่ในเครื่องคอมพิวเตอร์ของผู้ชม เมื่อผู้ชมเข้ามาเยี่ยมชมเว็บเพจในเว็บไซต์เป็นครั้งแรกและต่อมาเมื่อผู้ชมเยี่ยมชมเว็บเพจภายในเว็บไซต์อีกสคริปต์ก็จะนำรหัสประจำตัวที่ฝังอยู่ในเครื่องส่งกลับไปยังเว็บเซิร์ฟเวอร์ ซึ่งเราสามารถฝังรหัสประจำตัวผู้ชมในรูปของคุกกี้ได้

## 3.2 ขั้นตอนการกำหนดรหัสประจำตัวผู้ชม

ก่อนจะเริ่มบันทึกค่าต่างๆ เราจะต้องดูว่าผู้ชมรายนี้ได้เริ่มติดต่อหรือกำลังติดต่ออยู่หรือไม่ เพื่อที่จะได้กำหนดค่าอ้างอิงให้กับผู้ชมคนนี้ (ค่าอ้างอิงจะเป็นเสมือนรหัสประจำตัวของเขา) ค่าอ้างอิงดังกล่าวจะกำหนดให้กับตัวแปรชื่อ UID โดยการหาค่ามาจากการสุ่มตัวเลขแล้วนำไปเข้ารหัสแบบ md5() ซึ่งเขียนเป็นคำสั่งสคริปต์ได้ดังนี้

```
$UID= md5(uniqid(rand()));
```

โดยที่ $UID จะได้ค่าอ้างอิงที่มีโอกาสซ้ำกันน้อยมากเพราะค่าที่ได้จะสุ่มมาก่อนจากคำสั่ง rand() ซึ่งจะให้ตัวเลขที่ไม่แน่นอนออกมาในแต่ละครั้งของการสุ่ม จากนั้นก็เอาค่าที่สุ่มได้ไปสร้างหมายเลขขึ้นมาชุดหนึ่ง ด้วยคำสั่ง uniqid() คือการสร้างหมายเลขออกมาโดยกำหนดจากเวลา

ปัจจุบันของเซิร์ฟเวอร์มีหน่วยเป็นไมโครวินาที (microsecond) จากนั้นจึงนำค่าที่ได้ไปเข้ารหัสด้วยคำสั่ง md5() ซึ่งจะทำให้ค่าที่ได้ออกมาแทบไม่มีโอกาสซ้ำกันได้เลย

ตัวอย่างของ $UID ที่ได้จากกระบวนการดังกล่าว คือ

098c03d1a7b6ecceececdc4b3fcde531

6fa7bd2a60b7167334d365ac6dfaa08

โปรดสังเกตว่า ค่าทั้ง 2 กำหนดให้สร้างขึ้นมาในเวลาไล่เลี่ยกันคือเขียนเป็น 2 บรรทัดดังนี้

```
echo md5(uniqid(rand()));
echo md5(uniqid(rand()));
```

แต่ค่าทั้ง 2 ค่าไม่ได้มีความสัมพันธ์หรือใกล้เคียงกันเลย

**ซอร์ซโค้ด**

```
<? // ตรวจสอบว่ามีรหัสประจำตัวผู้ชม $UID หรือยัง
if (!$UID)
{
// สุ่มค่าแล้วนำค่าที่ได้ไปเข้ารหัสแบบ md5()
// จากนั้นจึงบันทึกเป็นค่าของคุกกี้
// โดยกำหนดให้คุกกี้มีอายุ 1 ชั่วโมง (60x60=3600 วินาที)
$UID=md5(uniqid(rand()));
SetCookie("uid",$UID,time()+3600);
}
else
{
// กำหนดให้คุกกี้ที่มีอยู่เดิมหมดอายุ
SetCookie("uid",$uid,0);
// set cookie กำหนดให้คุกกี้มีอายุ 1 ชั่วโมง (60x60=3600 วินาที)
SetCookie("uid",$uid,time()+3600);
}
?>
```

### 3.3 ขั้นตอนการกำหนดทรานแซคชัน[3]

การกำหนดทรานแซคชัน (transaction identification) คือการสร้างกลุ่มของข้อมูลการเยี่ยมชมเพจของผู้ชมแต่ละคน ดังนั้นงานของการกำหนดทรานแซคชันจึงอาจเป็นการแบ่งทรานแซคชันออกเป็นทรานแซคชันย่อยๆหรือการรวมทรานแซคชันเล็กเข้าด้วยกันเพื่อสร้างทรานแซคชันที่เหมาะสมกับการทำเหมืองข้อมูล

การกำหนดทรานแซคชันจะนำลิสต์ของข้อมูลการเยี่ยมชมและพารามิเตอร์บางอย่างเป็นข้อมูลเข้าและข้อมูลออกเป็นลิสต์ของทรานแซคชัน เมื่อกำหนดให้ L เป็นเซตของล็อกไฟล์แล้วแต่ละระเบียนล็อกไฟล์ l ∈ L จะประกอบไปด้วย รหัสประจำตัวผู้ชม l.uid , URL ของเพจที่ชม l.url และเวลาที่ชม l.time

การกำหนดทรานแซคชันจะพิจารณาระยะเวลาเยี่ยมชมเป็นส่วนประกอบด้วยซึ่งเรามีข้อสันนิษฐานที่ว่าระยะเวลาเยี่ยมชมเว็บเพจมีความสัมพันธ์ต่อการจำแนกว่าการเยี่ยมชมนั้นจะเป็นการเยี่ยมชมเพื่อใช้เป็นทางผ่าน (navigation) หรือเพื่อดูเนื้อหา (content) การเยี่ยมชมแบบใช้เป็นทางผ่านจะมีระยะเวลาสั้นและมีความแปรปวนของระยะเวลาเยี่ยมชมต่ำ ส่วนการเยี่ยมชมแบบดูเนื้อหาจะมีระยะเวลายาวกว่าและมีความแปรปวนของระยะเวลาเยี่ยมชมสูง เมื่อกำหนดเปอร์เซนต์ของการเยี่ยมชมแบบใช้เป็นทางผ่านในล็อกไฟล์มาให้แล้วค่าประมาณของเวลาที่เป็นจุดแบ่งแยก (เรียกว่า เวลาแบ่งแยก) ที่ใช้ในการจำแนกประเภทของการเยี่ยมชมนั้นสามารถคำนวณได้จากค่าประมาณทางสถิติแบบปกติของการกระจายแบบไคสแควร์ของระยะเวลาการเยี่ยมชมทั้งหมด

***นิยาม*** *3.1 กำหนดให้* $uid_t$ *คือ รหัสประจำตัวผู้ชม* $l_i^t.url$ *คือ เว็บเพจที่เยี่ยมชมในการเยี่ยมชมครั้งที่ i* $l_i^t.length$ *คือ ระยะเวลาเยี่ยมชมในครั้งที่ i* $l_i^t.time$ *คือ เวลาที่เริ่มเยี่ยมชมในครั้งที่ i L คือ เซตของล็อกเอ็นทรีทั้งหมด ทรานแซคชันแบบมีระยะเวลาเยี่ยมชมเป็นส่วนประกอบ (t) มีส่วนประกอบ 4 ส่วนดังนี้*

$$t = \langle uid_t, \{(l_1^t.url, l_1^t.time, l_1^t.length), \ldots, (l_m^t.url, l_m^t.time, l_m^t.length)\}\rangle$$

*โดยที่* $l_k^t \in L$ *และ* $l_k^t.uid = uid_t$ *และ* $l_i^t.length = l_{i+1}^t.time - l_i^t.time$

$1 \le i \le m-1$ *และ 2 นาที* $\le l_i^t.length \le$ *30 นาที*

จากนิยาม 3.1 เราสามารถอธิบายได้ว่าแต่ละทรานแซคชันสร้างขึ้นมาจากข้อมูลการเยี่ยมชมของผู้ชมคนเดียวกันโดยพิจารณาจากค่ารหัสประจำตัวผู้ชมและระยะเวลาของการเยี่ยมชมในแต่ละครั้งถูกประมาณจากค่าผลต่างระหว่างเวลาของการเยี่ยมชมครั้งถัดไปกับเวลาเยี่ยมชมในครั้งนั้น แต่ในกรณีของการเยี่ยมชมครั้งสุดท้ายของทรานแซคชันจะไม่มีค่าเวลาการเยี่ยมชมในครั้ง

ถัดไปเพื่อนำมาใช้ประมาณค่าระยะเวลาเยี่ยมชม ดังนั้นเราจึงกำหนดให้การเยี่ยมชมในครั้งสุดท้ายทั้งหมดให้เป็นการเยี่ยมชมเพื่อดูเนื้อหา และจะไม่นำมาคำนวณหาเวลาแบ่งแยก ซึ่งการกำหนดแบบนี้จะก่อให้เกิดความผิดพลาดถ้าการเยี่ยมชมสุดท้ายเป็นการเยี่ยมชมแบบใช้เป็นทางผ่านเพื่อใช้ออกจากเว็บไซต์ แต่ก็มีความเป็นไปได้ที่ผู้ชมจะเยี่ยมชมเว็บไซต์มากกว่าหนึ่งครั้งซึ่งในกรณีนี้ระยะเวลาของการเยี่ยมชมครั้งสุดท้ายของการเยี่ยมชมรอบแรกอาจถูกคำนวณได้เป็นระยะเวลานานนับชั่วโมง ซึ่งโดยปกติแล้วมีความเป็นไปได้ต่ำมากที่ผู้ชมจะเยี่ยมชมเว็บไซต์เป็นระยะเวลานับชั่วโมง เพื่อป้องกันข้อผิดพลาดเราจะกำหนดให้ระยะเวลาของการเยี่ยมชมในแต่ละครั้งมีค่าไม่เกินครึ่งชั่วโมง ตัวอย่างเช่นการรับโทรศัพท์หรือพักเที่ยงจะทำให้เกิดการจำแนกที่ผิดพลาดโดยจะจำแนกการเยี่ยมชมแบบใช้เป็นทางผ่านเป็นการเยี่ยมชมเพื่อดูเนื้อหา ซึ่งมีความเป็นไปได้ต่ำที่ความผิดพลาดในลักษณะนี่จะเกิดขึ้นบนเพจเดียวกัน แต่ค่าขีดแบ่ง (threshold) คือค่าซัพพอร์ตขั้นต่ำ (minimum support) ของอัลกอริทึมการทำเหมืองข้อมูลจะขจัดความผิดพลาดเหล่านี้ออกไป

เมื่อได้ค่าเวลาแบ่งแยกแล้วจะเปรียบเทียบกับค่าระยะเวลาของการเยี่ยมชมในแต่ละครั้งกับเวลาแบ่งแยกเพื่อวิเคราะห์และกำหนดว่าควรเป็นทรานแซคชันประเภทใด

ถ้า C เป็นเวลาแบ่งแยกแล้ว เงื่อนไขของทรานแซคชันแบบใช้เป็นทางผ่าน คือ

$$1 \leq i \leq (m-1) \ : \ l_k^t.\text{length} \leq C$$

$$\text{และ } i = m \ : \ l_k^t.\text{length} > C$$

ซึ่งจะเพิ่มเข้าไปในนิยามที่ 3.1 และเงื่อนไขของทรานแซคชันแบบดูเนื้อหา คือ

$$1 \leq i \leq m \ : \ l_k^t.\text{length} > C$$

ซึ่งจะเพิ่มเข้าไปในนิยามที่ 3.1 อีกเช่นกัน

ค่าพารามิเตอร์ตัวหนึ่งซึ่งวิธีการกำหนดทรานแซคชันแบบนี้ต้องการคือค่าประมาณเป็นเปอร์เซนต์ของการเยี่ยมชมแบบใช้เป็นทางผ่านซึ่งหาได้จากโครงสร้างและเนื้อหาของเว็บไซต์หรือประสพการณ์ของนักวิเคราะห์ข้อมูล แต่ในงานวิจัยนี้จะกำหนดให้ C มีค่าเท่ากับ 2 นาที

### 3.4 ขั้นตอนการวิเคราะห์ข้อมูลทางสถิติ

เมื่อได้ล็อกไฟล์ของการเยี่ยมชมไซต์ในหนึ่งวันจากในฐานข้อมูลแล้วเราก็จะนำมาวิเคราะห์เพื่อหาสถิติต่างๆที่เกี่ยวข้องกับข้อมูลของการชมไซต์โดยใช้คำสั่งภาษา SQL

ข้อมูลสถิติที่ทำการวิเคราะห์มีดังต่อไปนี้

- สถิติแยกตามค่าไอพีแอดเดรส (IP address)

  แสดงจำนวนผู้ชมแยกตามค่าไอพีแอดเดรส โดยแสดงจำนวนผู้ชมจริงและจำนวนเปอร์เซนต์ของแต่ละไอพีแอดเดรส

- สถิติแยกตามเว็บเพจในไซต์ (URL)
  เป็นสถิติที่แสดงจำนวนผู้ชมแยกตามเว็บเพจ โดยแสดงจำนวนครั้งของการเยี่ยมชมทั้งหมดและจำนวนผู้ชมจริงที่เข้ามาดูเว็บเพจแต่ละเพจในไซต์
- สถิติแยกตามแหล่งอ้างอิง (referrer)
  เป็นสถิติที่แสดงเกี่ยวกับแหล่งอ้างอิงต่างๆที่ช่วยนำผู้ชมเข้ามายังไซต์ เช่น จากเว็บไซต์อื่น, เซิร์ชเอ็นจิน, ไฟล์ในเครื่องผู้ชม, เว็บเพจอื่นในไซต์ และอื่นๆ โดยแสดงจำนวนผู้ชมจริง, จำนวนเปอร์เซนต์แยกตามแหล่งอ้างอิง
- สถิติแยกตามโปรแกรมเว็บเบราเซอร์ (browser)
  เป็นสถิติที่ได้จากข้อมูลของระบบปฏิบัติการและโปรแกรมเว็บเบราเซอร์แยกตามยี่ห้อและเวอร์ชัน โดยแสดงจำนวนผู้ชมจริงและจำนวนเปอร์เซนต์แยกตามโปรแกรมเว็บเบราเซอร์
- สถิติแยกตามชั่วโมงเข้าชม (visitors per hour)
  เป็นสถิติที่แสดงจำนวนผู้ชมแยกตามแต่ละชั่วโมงของวัน โดยเริ่มตั้งแต่เวลา 00.00-00.59 น. ไปจนถึง 23.00–23.59 น. โดยแสดงจำนวนผู้ชมจริงและจำนวนเปอร์เซนต์ของแต่ละชั่วโมง
- สถิติแยกตามวันเข้าชม (visitor)
  เป็นสถิติที่แสดงจำนวนครั้งที่เข้าชมไซต์และจำนวนผู้ชมจริงในแต่ละวัน
- สถิติแยกตามจำนวนเว็บเพจที่เยี่ยมชม (number of visited pages)
  เป็นสถิติที่แสดงว่าผู้ชมคลิกท่องไปในไซต์จำนวนกี่เว็บเพจในการมาเยือนแต่ละครั้ง ถ้าสถิติชี้ว่ามีผู้ชมจำนวนมากที่ท่องเว็บเพจเพียงไม่กี่เพจ แสดงว่าการจัดระเบียบหรือการวางลิงก์ในไซต์อาจมีปัญหาหรือไม่เหมาะสม
- สถิติแยกตามระยะเวลาที่ท่องในไซต์ (Time spent on site)
  แสดงระยะเวลาที่ผู้ชมใช้ในการชมไซต์ โดยแสดงจำนวนผู้ชมจริงแยกตามระยะเวลาที่ชม

## 3.5 สรุป

ผู้วิจัยได้เสนอวิธีสร้างล็อกไฟล์การเยี่ยมชมซึ่งมีรหัสประจำตัวผู้ชมขึ้นโดยนำสคริปต์ที่ทำหน้าที่เก็บข้อมูลการเยี่ยมชมไปฝังไว้ในเว็บเพจทุกเพจในไซต์โดยที่รหัสประจำตัวผู้ชมจะถูกฝังอยู่ในเครื่องคอมพิวเตอร์ของผู้ชมในรูปของคุกกี้ และต่อมาจะนำล็อกไฟล์การเยี่ยมชมในแต่ละวันที่ได้ส่งเข้าไปให้กับขั้นตอนการกำหนดทรานแซคชันซึ่งจะพิจารณาระยะเวลาเยี่ยมชมเป็นส่วนประกอบด้วยโดยแบ่งทรานแซคชันของผู้ชมแต่ละคนตามรหัสประจำตัวผู้ชมและระยะเวลาของการเยี่ยมชมในแต่ละครั้งจะมีค่าอยู่ในช่วง 2-30 นาที