

โครงการทางวิศวกรรม

เรื่อง

ระบบวิเคราะห์การแพร่กระจายข่าวลือบนทวิตเตอร์

The System for Rumor Dispersion Analysis on Twitter

โดย

นายปิยวัฒน์ เลิศวิทยากำจร 5431022721

นางสาวพนิดา นิ่มนวล 5431025621

อาจารย์ที่ปรึกษาโครงการ อ.ดร. พีรพล เวทีกุล ลายมือชื่อ

อาจารย์ที่ปรึกษาโครงการ(ร่วม) ผศ. พิจิตรา สีคาโมไต ลายมือชื่อ

รายงานฉบับนี้เป็นส่วนหนึ่งของวิชาโครงการวิศวกรรมคอมพิวเตอร์

หลักสูตรวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2557

สารบัญ

หัวข้อ	หน้า
ชื่อโครงการ	1
1. ปัญหาและความสำคัญของปัญหา	1
2. วัตถุประสงค์	2
3. เป้าหมาย	2
4. ทฤษฎีที่เกี่ยวข้อง	3
4.1 ทวิตเตอร์ (Twitter)	3
4.2 ทวิตเตอร์ เอพีไอ (Twitter API)	3
4.3 ฐานข้อมูลเชิงเอกสาร (Document-oriented Database)	5
4.4 คลังข้อมูล (Data Warehouse)	6
4.5 การทำเหมืองข้อมูล (Data Mining)	8
4.6 โปรแกรมสำหรับวิเคราะห์ข้อมูลทวิตเตอร์	9
5. งานวิจัยที่เกี่ยวข้อง	11
5.1 Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter	11
5.2 Forecasting with Twitter Data	11
5.3 Twitter Under Crisis: Can we trust what we RT?	12
6. ขอบเขตความต้องการเชิงหน้าที่ของระบบ	12
7. แนวทางในการพัฒนาและเครื่องมือที่เกี่ยวข้อง	13
8. ขั้นตอนการดำเนินงาน	15
8.1 แผนภูมิ Gantt แสดงขั้นตอนการดำเนินงาน	15
8.2 ตารางแสดงขั้นตอนการดำเนินงานในช่วงเดือนต่างๆ	17
9. ประโยชน์ที่คาดว่าจะได้รับ	18
10. รายการอ้างอิง	19

1. ปัญหาและความสำคัญของปัญหา

ในปัจจุบัน สื่อสังคมออนไลน์เป็นช่องทางในการเผยแพร่ข้อมูลข่าวสารจำนวนมากจากหลากหลายที่มาและมีจุดประสงค์การเผยแพร่ข่าวที่แตกต่างกันไป โดยผู้ที่รับข่าวสารเหล่านั้นยากที่จะทราบได้ว่าข่าวใดเป็นข่าวลือ ข่าวใดเป็นข่าวจริง ผู้ที่รับข่าวสารบางกลุ่มเลือกที่จะตรวจสอบที่มาที่ไปจากแหล่งข่าวอื่นๆ ที่เชื่อถือได้ก่อนที่จะปักใจเชื่อ แต่ก็มีผู้รับข่าวสารบางกลุ่มที่เชื่อในข่าวที่ตนเองได้รับ และส่งต่อข่าวนั้นทันทีโดยไม่ผ่านการกลั่นกรองก่อน ซึ่งหากข่าวเหล่านั้นเป็นข่าวลือ และถูกเผยแพร่ออกไป อาจก่อให้เกิดความเข้าใจผิดและส่งผลเสียต่างๆ ตามมามากมาย โดยเฉพาะอย่างยิ่งสำหรับประเด็นที่ละเอียดอ่อนและเป็นที่สนใจของสังคม เช่น ประเด็นทางการเมือง ท่ามกลางวิกฤตการณ์ทางการเมืองของประเทศไทยที่ดำเนินมาเกือบสิบปี ยังมีคนไทยหลายกลุ่มที่ถูกกระตุ้นด้วยข้อมูลการโฆษณาชวนเชื่อของกลุ่มทางการเมือง ทำให้สังคมไทยยังคงอยู่ในวังวนของความขัดแย้งและต่อสู้กันด้วยอารมณ์มากกว่าเหตุผล

ปัญหานี้เป็นที่น่าสนใจสำหรับวงการนิเทศศาสตร์ โดยเฉพาะอย่างยิ่งสาขาวารสารสนเทศ ที่ต้องการหาคำตอบว่าข่าวลือในโลกสังคมออนไลน์มีพฤติกรรมอย่างไร มีช่วงเวลาในการแพร่กระจายสั้นหรือยาวเพียงใด มีใครที่เกี่ยวข้องกับการแพร่กระจายข่าวลือเหล่านี้บ้าง และส่วนหนึ่งในนั้นมีบุคคลที่เป็นสื่อมวลชน, ภาครัฐ, นักการเมือง, นักวิชาการ หรือบุคคลในกลุ่มการเมืองต่างๆ เข้าเกี่ยวข้องด้วยหรือไม่ เพื่อนำไปสู่การวิเคราะห์ความเป็นไปในสังคมโลกแห่งความเป็นจริงอย่างเป็นรูปธรรมและเชื่อถือได้ แต่การศึกษาวิจัยในหัวเรื่องนี้เป็นไปได้ค่อนข้างยากลำบาก เนื่องจากข้อมูลในสังคมออนไลน์มีปริมาณมากและเพิ่มขึ้นอย่างรวดเร็ว ทำให้ต้องใช้ทรัพยากรบุคคลและเวลาจำนวนมากในการวิเคราะห์

โครงการนี้จึงมุ่งเน้นที่จะสร้างระบบวิเคราะห์การแพร่กระจายข่าวลือบนโลกสังคมออนไลน์ ซึ่งจะเป็นเครื่องมือสำคัญที่ช่วยเหลือนักวิจัยทางด้านนิเทศศาสตร์ในการศึกษาคุณลักษณะและพฤติกรรมในแง่มุมต่างๆ ของข่าวลือที่ถูกแพร่กระจายออกไป โดยมีขอบเขตการศึกษาอยู่ที่ข้อมูลบนทวิตเตอร์ซึ่งเป็นสื่อสังคมออนไลน์ที่มีผู้ใช้เป็นอันดับสองของประเทศไทย รองจากเฟซบุ๊ก และมีการนำมาใช้เพื่อกระจายข้อมูลข่าวสารจากทั้งสื่อมวลชน บุคคลสาธารณะ และบุคคลทั่วไป ได้อย่างรวดเร็ว สาเหตุสำคัญที่มุ่งศึกษาไปที่ทวิตเตอร์ เนื่องจากทวิตเตอร์มีลักษณะการส่งข้อมูลเป็นข้อความขนาดสั้น (ไม่เกิน 140 ตัวอักษร) ทำให้ได้ข้อความที่กระชับ แต่ในขณะเดียวกันก็ทำให้เกิดความกำกวมในเนื้อหาได้ง่าย ทำให้มีคนใช้ทวิตเตอร์เป็นเครื่องมือในการเผยแพร่ข่าวลือมากกว่าสื่ออื่น อีกทั้งทวิตเตอร์ยังมีข้อจำกัดทางด้านความเป็นส่วนตัวน้อยกว่าเฟซบุ๊ก ทำให้เหมาะกับการศึกษาในโครงการนี้ โดยระบบที่สร้างขึ้นจะอยู่ในรูปแบบระบบเว็บ (Web Application) ที่สามารถรายงานผลการวิเคราะห์ข้อมูลทวิตเตอร์ตามคำค้นหาของผู้ใช้ อาทิ รายการทวิต (Tweet) ที่เกี่ยวข้อง, อัตราการแพร่กระจายของทวิตที่สนใจ, กลุ่มบุคคลที่เกี่ยวข้องกับทวิตเหล่านั้น และประเภทของอุปกรณ์ที่เกี่ยวข้องใช้เป็นต้น



2. วัตถุประสงค์

เพื่อพัฒนาระบบเว็บสำหรับวิเคราะห์การแพร่กระจายของข้อมูลทวิตเตอร์ ซึ่งจะช่วยลดระยะเวลาและประหยัดค่าใช้จ่ายในการศึกษาพฤติกรรมของชาวสื่อ

3. เป้าหมาย

- 3.1 ระบบสามารถวิเคราะห์ข้อมูลเกี่ยวกับการแพร่กระจายของทวิตบนทวิตเตอร์ได้
- 3.2 ผู้ใช้ทั่วไปสามารถเรียนรู้การใช้งานระบบอย่างเต็มประสิทธิภาพได้อย่างรวดเร็ว
- 3.3 ระบบสามารถแสดงผลลัพธ์ของการวิเคราะห์ได้อย่างชัดเจน และง่ายต่อการเข้าใจ
- 3.4 ระบบมีการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลช่วยในการวิเคราะห์ข้อมูล

4. ทฤษฎีที่เกี่ยวข้อง

4.1 ทวิตเตอร์ (Twitter)^[2]

ทวิตเตอร์ (Twitter) เป็นเว็บไซต์ที่ให้บริการบล็อกสั้น (Micro-Blog) นั่นคือ สามารถเขียนข้อความได้ไม่เกิน 140 ตัวอักษร โดยผู้ใช้สามารถส่งข้อความของตนเองให้เพื่อน ที่ติดตาม ทวิตเตอร์ตนเองอยู่ และสามารถอ่านข้อความของเพื่อนหรือคนที่เราติดตามได้ตามอยู่ได้ ซึ่งทวิตเตอร์ถือว่าเป็นเว็บไซต์ประเภทสื่อสังคมออนไลน์ด้วยเช่นกัน จึงทำให้โดยส่วนใหญ่แล้วผู้ใช้งานจะใช้ทวิตเตอร์เพื่อติดตามข่าวสาร บอกคนอื่นว่าตนเองกำลังทำอะไรหรือมีเหตุการณ์อะไรเกิดขึ้น แบ่งปันเว็บไซต์ที่น่าสนใจที่ได้อ่านมา ถามคำถามที่ต้องการประสบการณ์จากผู้อื่น แลกเปลี่ยนพูดคุยแสดงความคิดเห็น ติดตาม และร่วมรายงานสภาพอากาศ และสภาพการจราจร

คำศัพท์ที่พบบ่อยในทวิตเตอร์^[4]

- | | |
|--------------|---|
| 1. Tweet | การโพสต์ข้อความในทวิตเตอร์ |
| 2. Follow | การติดตามอ่านข้อความในทวิตเตอร์ของคนอื่น |
| 3. Follower | คนที่มาติดตามอ่านทวิตเตอร์ของเรา |
| 4. Following | คนที่เรากำลังติดตามอ่านทวิตเตอร์ของเขา |
| 5. Retweet | การส่งต่อข้อความ หรือการโพสต์ข้อความที่ตนเองชื่นชอบ หรือถูกใจ หรือเห็นว่าเหมาะสมแก่การแนะนำต่อผู้อื่น ซ้ำอีกครั้ง |
| 6. Reply | การตอบกลับข้อความของบุคคลอื่น โดยใช้รูปแบบ @ชื่อทวิตเตอร์ ข้อความที่ต้องการตอบกลับ |

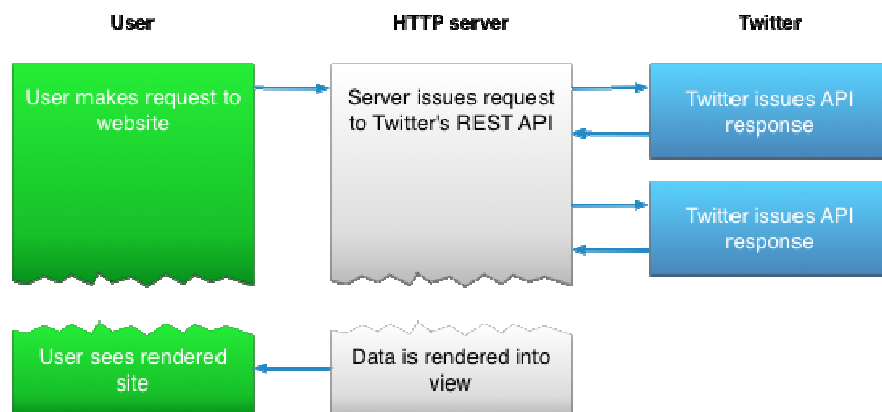
4.2 ทวิตเตอร์ เอพีไอ (Twitter API)^[6]

ทวิตเตอร์ เอพีไอ (Twitter API) แบ่งหลักๆได้ออกเป็น 2 รูปแบบ ได้แก่

4.2.1 REST API

REST API เป็นเอพีไอของทวิตเตอร์ที่ทำให้อ่าน หรือเขียนข้อมูลทวิตเตอร์ได้ โดยมีลักษณะเป็นการขอเป็นครั้งๆไป นั่นคือส่งคำขอไปยังทวิตเตอร์ และทวิตเตอร์ตอบกลับมาก็ถือเป็นการจบการขอ 1 ครั้ง ดังภาพที่ 1 จึงเหมาะกับระบบเว็บที่มีการรับการร้องขอจากผู้ใช้เรื่อยๆ ตัวอย่างการใช้ REST API เช่น การอ่านประวัติของบุคคลต่างๆ การอ่านผู้ติดตามของแต่ละบุคคล การอ่านหน้าหลักของแต่ละบุคคล เป็นต้น โดยผลลัพธ์ที่ได้จะได้อีกกลับมาในรูปแบบของ JSON อย่างไรก็ตาม twitter ได้กำหนดขีดจำกัดของอัตราการขออ่าน หรือเขียนข้อมูลเอาไว้ โดยจำกัดอัตราการขออ่านหรือเขียนข้อมูลโดยพิจารณาเป็นช่วง ช่วงละ 15 นาที ซึ่งโดยส่วนใหญ่ใน 1 ช่วง สามารถร้องขอการอ่านหรือเขียนข้อมูลได้ 15 ครั้ง ยกเว้นบางการร้องขอที่เกี่ยวข้องกับ

การค้นหา สามารถร้องขอได้ 180 ครั้งใน 1 ช่วง เช่น GET statuses / show / :id (การร้องขอทวีตที่มี id ตามที่ต้องการ) GET search/tweets (การร้องขอรายการทวีตที่เกี่ยวข้องกับพารามิเตอร์ที่ส่งไป) เป็นต้น



ภาพที่ 1 รูปแบบการใช้งาน Twitter Rest API

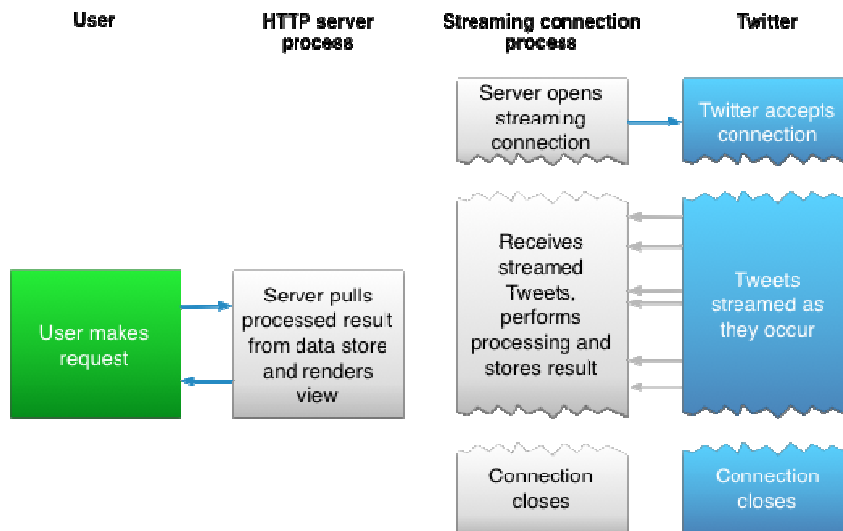
ที่มาภาพ: <https://dev.twitter.com/streaming/overview>

4.2.2 Streaming API

Streaming API เป็นเอพีไอที่ทำให้สามารถได้รับผลลัพธ์จากการร้องขอแบบ REST API ได้อย่างต่อเนื่อง บนการเชื่อมต่อแบบ HTTP จึงทำให้จำเป็นต้องเชื่อมต่อแบบ HTTP ตลอดการนำเข้าหรือร้องขอข้อมูล ดังภาพที่ 2 เช่น การใช้ Streaming API เพื่อการรับทวีต ล่าสุดที่ตรงกับสิ่งที่ระบุไปในพารามิเตอร์

Streaming แบ่งออกเป็น 3 ประเภท ได้แก่

1. Public streams เป็นการรับข้อมูลสาธารณะบนทวีตเตอร์เหมาะสำหรับการติดตามบุคคลหรือเรื่องที่มีความเฉพาะเจาะจง หรือเพื่อการทำเหมืองข้อมูล
2. User streams เป็นการรับข้อมูลหรือการกระทำต่างๆของบุคคลใดบุคคลหนึ่ง
3. Site streams เป็นการรับข้อมูลแบบเดียวกับ User streams แต่เป็นการทำ User streams หลายคน



ภาพที่ 2 รูปแบบการใช้งาน Twitter Streaming API

ที่มาภาพ: <https://dev.twitter.com/streaming/overview>

4.3 ฐานข้อมูลเชิงเอกสาร (Document-oriented Database)

ฐานข้อมูลเชิงเอกสาร (Document-oriented Database) เป็นรูปแบบการเก็บข้อมูลประเภทหนึ่งในกลุ่ม NoSQL¹ ที่ไม่ได้จัดเก็บข้อมูลในลักษณะของตาราง (Table) แบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database) แต่จะเก็บข้อมูลเป็นชุด (Collection) ของเอกสารแทน หากเปรียบเทียบกันแล้ว 1 แถวในตารางของฐานข้อมูลเชิงสัมพันธ์นั้น จะเทียบเท่ากับไฟล์เอกสาร 1 ไฟล์ที่ถูกเก็บเอาไว้ในชุดเอกสารในฐานข้อมูลเชิงเอกสาร

ในวงการคอมพิวเตอร์ยุคปัจจุบัน เริ่มมีการนำฐานข้อมูลเชิงเอกสารเข้ามาใช้แทนฐานข้อมูลเชิงสัมพันธ์มากขึ้น เนื่องจากแนวคิดหลักของฐานข้อมูลเชิงเอกสารที่ไม่มีการจัดเก็บข้อมูลเป็นแถวในตารางนั้น ทำให้ลดข้อจำกัดข้อสำคัญที่ฐานข้อมูลเชิงสัมพันธ์ต้องเผชิญออกไปได้ ซึ่งก็คือความจำเป็นที่ตารางต่างๆจะต้องมีโครงสร้าง (Schema) กำกับที่ชัดเจน ข้อมูลแต่ละแถวจะต้องมีจำนวนคอลัมน์เท่ากันถึงแม้ว่าอาจจะไม่มีข้อมูลในบางคอลัมน์ก็ตาม ผลที่ตามมาคือตารางที่ได้จะมีลักษณะที่เป็นรูโหว่จำนวนมาก ซึ่งเราจะเรียกตารางแบบนี้ว่า Sparse Table ตารางในลักษณะนี้ จะทำให้เราเปลืองเนื้อที่ในการจัดเก็บข้อมูลช่องที่ว่างเปล่าไปโดยใช่เหตุและต้องใช้เวลาในการประมวลผลข้อมูลมาก โดยเฉพาะอย่างยิ่งในเวลาที่ต้อง Join ตารางขนาดใหญ่หลายๆตารางเข้าด้วยกัน^[10]

การจัดเก็บข้อมูลแบบฐานข้อมูลเชิงเอกสารนั้นจะช่วยแก้ปัญหานี้ได้ เพราะฐานข้อมูลจะไม่มีโครงสร้าง เอกสารแต่ละใบจะมีจำนวน Attribute กี่ตัวก็ได้ เท่ากันหรือไม่เท่ากันก็ได้ ทำให้เก็บข้อมูลที่เป็นโครงสร้างลำดับชั้นแบบซับซ้อน (complex-hierarchical) ได้ดี และฐานข้อมูลเชิงเอกสารยังสามารถทำงาน

¹ NoSQL เป็นฐานข้อมูลที่ไม่ได้จัดเก็บอยู่ในรูปของ Relational Database และไม่สามารถใช้ภาษา SQL ในการดำเนินการกับข้อมูลได้ เหมาะสำหรับการจัดการข้อมูลขนาดใหญ่ และมีโครงสร้างไม่ตายตัว

กับภาษาโปรแกรมเชิงวัตถุ (Object-Oriented Programming Language) ได้อย่างคล่องตัวอีกด้วย เนื่องจาก 1 ไฟล์เอกสารก็เปรียบเสมือน 1 Object ที่มี Attribute เป็นข้อมูล Field ต่างๆนั่นเอง

นอกจากนั้น ด้วยความที่ฐานข้อมูลเชิงเอกสารมีลักษณะเป็น NoSQL ทำให้ได้รับข้อดีจากความเป็น NoSQL มาด้วย ^[9] กล่าวคือ เมื่อข้อมูลมีขนาดใหญ่ เราสามารถทำการกระจายข้อมูลไปยัง Server ต่างๆได้อย่างอัตโนมัติ (Auto-Sharding) โดยใช้การกระจายไฟล์ออกไปตามส่วนต่างๆได้เลย ผู้พัฒนา (Developer) ไม่ต้องเขียนโปรแกรมในการกระจายข้อมูลเอง เหมือนฐานข้อมูลเชิงสัมพันธ์ (ซึ่งต้องคำนึงถึงความสัมพันธ์ระหว่างตารางด้วย) ด้วยเหตุนี้ NoSQL จึงรองรับการขยายขนาดของระบบ (Scalable) ได้ดี นอกจากนั้น ยังสามารถทำสำเนาข้อมูล (Replication) ไปยัง Server สำรอง เพื่อรับประกันความพร้อมในการให้บริการ (Availability) ที่สูงขึ้นได้อีกด้วย

ตัวอย่างของ Document-Oriented Database ได้แก่ MongoDB, CouchDB และ Elasticsearch



4.4 คลังข้อมูล (Data Warehouse)

คลังข้อมูล (Data Warehouse) เป็นแหล่งข้อมูลที่รวบรวมข้อมูลจากหลากหลายแหล่งที่แตกต่างกัน ไว้ในที่เดียวกันและรูปแบบเดียวกัน เน้นเรื่องการเอาความรู้ (Information) ที่ผ่านการประมวลผลแล้วออกมา ซึ่งอาจจะออกมาในรูปแบบของรายงานต่างๆ หรือหน้าจาวិเคราะห์ผล เพื่อช่วยในการประกอบการตัดสินใจเรื่องต่างๆ^[3]

การใช้งาน	Operational Database	Data Warehouse
ลักษณะงาน	ใช้เพื่อการทำงานรายวัน (OLTP)	ใช้เพื่อการวิเคราะห์ (OLAP)
ลักษณะข้อมูล	ข้อมูลในแต่ละวัน หรือการทำงานทั่วไป	ข้อมูลในอดีต
การเคลื่อนไหวของข้อมูล	ตลอดเวลา	คงที่จนกว่าจะปรับปรุงใหม่
โครงสร้าง	ลดความซ้ำซ้อน (Normalization)	Dimensional Modeling
Operation	ทำการดำเนินการ(operation)เดิมซ้ำๆ	เปลี่ยนคำถามที่ใช้ไปเรื่อยๆ

ตารางแสดงความแตกต่างระหว่าง Operational Database กับ Data Warehouse

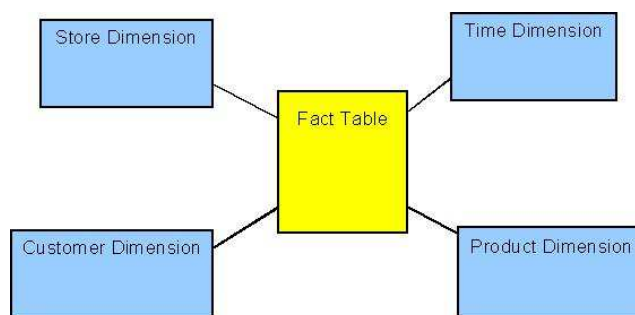
องค์ประกอบของคลังข้อมูล

1. Dimension Table เป็นตารางที่ประกอบด้วยรายละเอียดของข้อมูล
2. Fact Table เป็นตารางที่เก็บค่าที่ต้องการวัดหรือวิเคราะห์ (Measure) ซึ่งเป็นตัวเลข และเก็บ key ที่มาจาก Dimension Table (Dimension Key)

แบบจำลองคลังข้อมูล (Data Warehouse Model)^[5]

โดยทั่วไปสามารถแบ่งลักษณะของแบบจำลองคลังข้อมูล ได้ออกเป็น 2 ประเภท ได้แก่

1. Star schema มีลักษณะคือ มี fact table อยู่ตรงกลางล้อมรอบด้วยหลายๆ dimension table ดังภาพที่ 3

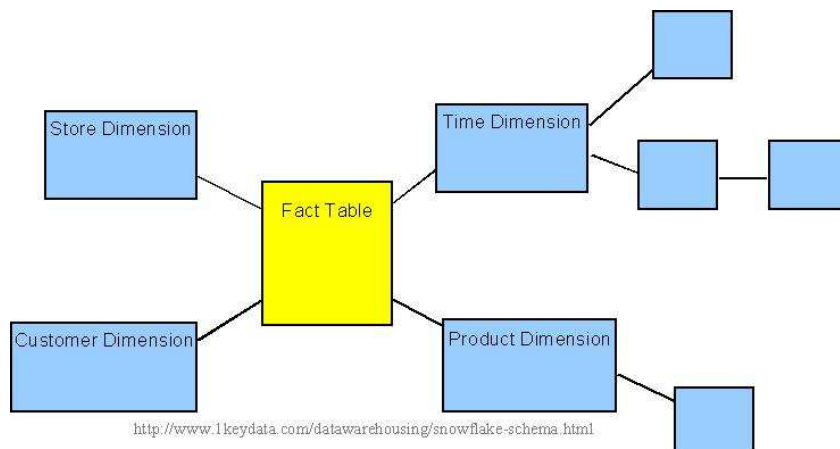


<http://www.1keydata.com/datawarehousing/star-schema.html>

ภาพที่ 3 ตัวอย่างโครงสร้างแบบจำลองคลังข้อมูลแบบ star schema

ที่มาภาพ: <http://www.1keydata.com/datawarehousing/star-schema.html>

2. Snowflake schema มีความแตกต่างจาก Star schema ที่ dimension table จะเก็บข้อมูลในรูป normal form อันเนื่องมาจากอาจเกิดปัญหาด้านการ design ไม่สามารถ implement โดยใช้ star schema ได้ เช่น fact table มีความสัมพันธ์แบบ Many to Many กับ dimension table เป็นต้น ดังภาพที่ 4



<http://www.1keydata.com/datawarehousing/snowflake-schema.html>

ภาพที่ 4 ตัวอย่างโครงสร้างแบบจำลองคลังข้อมูลแบบ snowflake

ที่มาภาพ: <http://www.1keydata.com/datawarehousing/snowflake-schema.html>

4.5 การทำเหมืองข้อมูล (Data Mining) ^[11]

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการในการวิเคราะห์ข้อมูลเพื่อที่จะค้นหารูปแบบที่ซ่อนอยู่โดยใช้ระเบียบวิธีแบบอัตโนมัติต่างๆ ซึ่งศาสตร์ทางด้านนี้กำลังได้รับความสนใจมากขึ้น เนื่องจากปัจจุบันความสามารถในการจุข้อมูลมากขึ้น และมีข้อมูลเกิดขึ้นมาใหม่จำนวนมากในทุกๆวัน แต่ความสามารถในการประมวลผลข้อมูลไม่ได้เพิ่มขึ้นทันการเพิ่มของปริมาณข้อมูล ทำให้องค์กรอยู่ในภาวะที่มีข้อมูลมากแต่มีความรู้ น้อย (Data-Rich, Knowledge-Low) กระบวนการทำเหมืองข้อมูลจะช่วยดึงเอาความรู้ออกมาจากข้อมูลที่มีอยู่ เป็นการทำให้ข้อมูลที่มีอยู่มีประโยชน์มากขึ้น

งานทางด้านการทำเหมืองข้อมูลมีหลายลักษณะ อาทิ

- 1) Classification: หาโมเดลที่จะอธิบายว่า Attribute ผลลัพธ์ที่เราสนใจเป็นฟังก์ชันของ Input Attributes ไດบ้าง อย่างไร กล่าวคือวิธีนี้จะหารูปแบบที่ใช้ในการระบุ Class ของแต่ละกรณีตัวอย่างนั่นเอง Algorithm ที่ใช้บ่อยในกลุ่มนี้ ได้แก่ Decision Trees, Neural Network, Naïve Bayes ใช้แก้ปัญหา Churn Analysis, Risk Management, Target Advertisements
- 2) Clustering (Segmentation): แบ่งกลุ่มของ Case ต่างๆตามกลุ่มของ Attribute Case ที่อยู่ในกลุ่มเดียวกัน อาจมี Attribute ที่มีความใกล้เคียงกันมากๆ โดยที่ไม่มี Class Attribute ที่จะเป็นตัวกำหนดทิศทางการทำเหมืองข้อมูล ทุก Attribute จะถูกพิจารณาอย่างเท่าเทียมกัน Algorithm ที่ใช้บ่อยในกลุ่มนี้ อาทิ K-means Clustering, DBSCAN
- 3) Association (Market Basket Analysis): ใช้หลักการวิเคราะห์ว่าสินค้าชนิดใดมักจะอยู่ในตะกร้าเดียวกัน โดยวิเคราะห์จากข้อมูลการซื้อขาย (Sales Transaction) เพื่อดูกลุ่มของสินค้าและปัจจัยการซื้อสินค้าข้ามกลุ่ม มี 2 เป้าหมายหลัก คือ เพื่อดูว่าสินค้าใดที่มักจะพบอยู่ด้วยกันบ่อยๆ และเพื่อที่จะหากฎ (Rules) จากความสัมพันธ์เหล่านั้น
- 4) Regression: หา Pattern เพื่อที่จะตอบค่าที่เป็นตัวเลขบางอย่าง โดยผลลัพธ์เป็นฟังก์ชันของกลุ่มของ Input คล้ายกับ Classification แต่ต่างกันตรงที่ผลลัพธ์ของ Regression จะเป็นค่าที่เป็นตัวเลขเทคนิคที่ใช้ เช่น Linear Regression, Logistic Regression, Regression Tree, Neural Network เช่น การทำนายความเร็วลม จากอุณหภูมิ ความกดอากาศ และความชื้น
- 5) Forecasting: การทำนายค่าบางอย่างในอนาคต โดยมี input เป็นลำดับของค่าต่างๆในช่วงเวลาที่ผ่านมา ใช้หลักของ Machine Learning และสถิติเข้าช่วย เช่น การทำนายยอดการขายไวน์ในเดือนหน้า
- 6) Sequence Analysis: การวิเคราะห์รูปแบบการเรียงลำดับของเหตุการณ์หรือสถานะต่างๆ เช่น การตรวจสอบลำดับการคลิกไปยังหน้าต่างๆบนเว็บไซต์

4.6 โปรแกรมสำหรับวิเคราะห์ข้อมูลทวีตเตอร์

เนื่องจากการวิเคราะห์ข้อมูลทวีตเตอร์เป็นหัวข้อที่บุคคลที่อยู่ในวงการคอมพิวเตอร์ให้ความสนใจเป็นอย่างมากในช่วงหลายปีที่ผ่านมา และถูกนำมาใช้ประโยชน์ในหลายภาคส่วนด้วยกัน ตัวอย่างเช่น ฝ่ายการตลาดใช้การวิเคราะห์อารมณ์ความรู้สึกของทวีต (Sentimental Analysis) เพื่อตรวจสอบว่าลูกค้าพึงพอใจในแคมเปญหรือสินค้าของบริษัทตนหรือไม่, การวิเคราะห์หากกลุ่มลูกค้าที่น่าจะสนใจในผลิตภัณฑ์ของตน โดยดูจากข้อความที่ทวีตเพื่อทำการตลาดให้ตรงกลุ่มเป้าหมาย และการหาผู้มีอิทธิพลในโลกสังคมออนไลน์ เพื่อติดต่อให้ช่วยประชาสัมพันธ์ข่าวสารหรือโฆษณาในหัวเรื่องต่างๆ เป็นต้น ด้วยเหตุนี้ ทำให้มีการผลิตซอฟต์แวร์ออกมาให้บริการทางการวิเคราะห์ข้อมูลทวีตเตอร์อยู่หลายอันด้วยกัน ซึ่งโดยส่วนมากจะอยู่ในรูปแบบของระบบเว็บเพราะสามารถใช้งานและเข้าถึงได้ง่ายเพียงแค่อินเทอร์เน็ต

ตัวอย่างของเว็บไซต์ที่ให้บริการวิเคราะห์ข้อมูลทวีตเตอร์ ได้แก่

1) TweetReach [www.tweetreach.com]

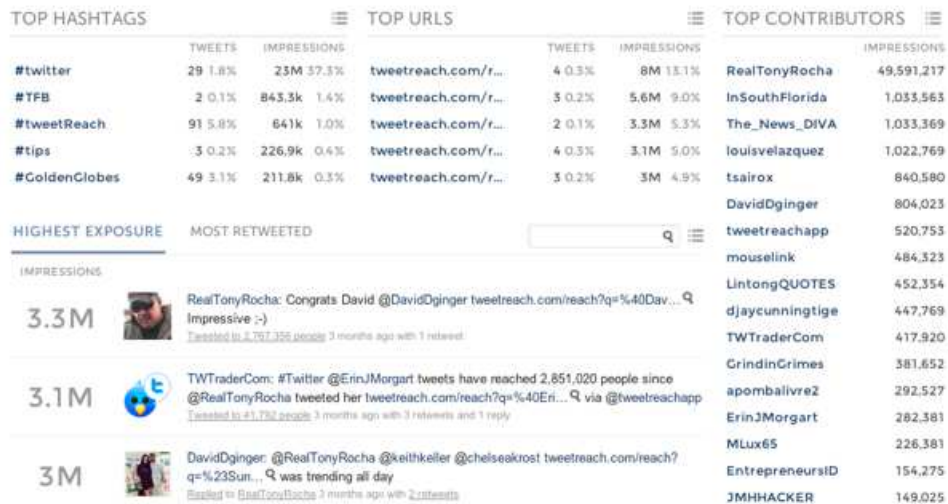
TweetReach ให้บริการค้นหาทวีตตาม username, hashtag, URL หรือคำต่างๆ (สำหรับภาษาไทยยังมีความคลาดเคลื่อนบ้างเล็กน้อย) เมื่อค้นหาข้อมูลเสร็จแล้ว ระบบจะทำการวิเคราะห์ข้อมูลเหล่านั้น และแสดงผลออกมาในแง่มุมต่างๆ อาทิ จำนวนครั้งในการเข้าถึงข้อมูลทวีต, ผู้ใช้ที่มีความสำคัญสูงสุดในด้านต่างๆ, ทวีตที่ถูกทวีตต่อมากที่สุด เป็นต้น โดยข้อมูลที่ได้สามารถ export เป็นไฟล์แบบ .pdf และ .csv ได้

TweetReach มีรูปแบบการให้บริการแบ่งได้ 3 ลักษณะ

1) Free Snapshots - สามารถรายงานผลสถิติข้างต้นเฉพาะ 50 ทวีตล่าสุดเท่านั้น แต่หากต้องการมากกว่านั้น จะมีบริการ Full Snapshots ในราคา \$20 ทวีตสูงสุด 1500 ทวีต ระยะเวลาย้อนหลัง 1-7 วัน เหมาะกับเหตุการณ์สั้นๆ หรือข้อมูลไม่มากนัก

2) Tweetreach Pro - สำหรับการติดตาม Twitter activity แบบ Realtime เช่นการติดตามการพูดคุยกันเกี่ยวกับแบรนด์, สินค้า, ภาคอุตสาหกรรม, การจัดงานอีเวนท์, คู่แข่ง และแคมเปญต่างๆ โดยใช้ tracker มีการแสดงค่าทางสถิติต่างๆ และสามารถดูเปรียบเทียบพร้อมกันหลาย keyword ได้ หากตรวจสอบค่าใดเสร็จแล้ว สามารถ export เป็นไฟล์รายงาน แล้วคืน tracker เพื่อนำไคติดตามใช้ตรวจสอบค่าอื่นๆได้

3) Historical Data - เป็นการวิเคราะห์ข้อมูลทวีตเตอร์ในอดีตที่ผ่านมา ด้วยสถิติในลักษณะเดียวกันกับสองแบบข้างต้น ราคาเริ่มต้น \$199 สำหรับ 1 ไฟล์รายงานที่สนใจ (\$99 สำหรับผู้ใช้ที่สมัคร TweetReach Pro อยู่แล้ว) ซึ่งผู้ใช้สามารถดูข้อมูลเจาะตามช่วงเวลาที่น่าสนใจได้, ทำเป็นไฟล์รายงานแบบ .csv และ .pdf ได้, มีการบอก Contributors ทั้งหมด พร้อมจำนวน follower ของแต่ละคน โดยราคาของข้อมูลขึ้นกับระยะเวลาที่ตรวจสอบ และจำนวนทวีตที่รวบรวมได้



ภาพที่ 5 หน้าแสดงผลการวิเคราะห์ของ TweetReach

2) Keyhole [keyhole.co]

Keyhole เป็นอีกเว็บไซต์หนึ่งที่ให้บริการวิเคราะห์ข้อมูลทวีตเตอร์ เว็บไซต์นี้มีจุดเด่นคือส่วนต่อประสานผู้ใช้ที่สวยงามและใช้งานง่าย มีบริการโดยทั่วไป คือ สามารถแสดงปริมาณผู้โพสต์ตามประเทศต่างๆ ได้, สามารถดาวน์โหลดไฟล์รูปแบบ xls ได้ โดยมีการแบ่งหัวข้อต่างๆชัดเจน สะดวกต่อการใช้งาน และสามารถดาวน์โหลดเฉพาะส่วนได้, แสดง Top posts แบ่งเป็น 3 ชนิด ตาม retweet, klout และ recent, แสดง Top sites ที่ลิงก์ไปถึง, แสดง Top hashtag, Top keyword, แสดง Most influential แบ่งเป็น 3 ประเภท ตาม avg RT, Klout และ Frequency, แสดงรูปภาพ Display ของ Influencer เป็นต้น

นอกเหนือจากบริการที่กล่าวไปข้างต้นแล้ว Keyhole ยังมีบริการพิเศษ อีกสองประเภท ได้แก่ Influencer Identification (หา User ที่มีอิทธิพลซึ่งกำลังพูดคุยในสิ่งคุณที่กำลังสนใจ) และ Newsroom Insights (ค้นหาสิ่งที่ผู้อ่านหรือ Follower สนใจ เพื่อที่จะเขียนข่าวได้ตรงกับความสนใจของผู้อ่าน)

Real-time Tracker: Audi



ภาพที่ 6 หน้าแสดงผลการวิเคราะห์ของ Keyhole

5. งานวิจัยที่เกี่ยวข้อง

5.1 Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter^[8]

งานวิจัยชิ้นนี้มีแนวคิดพื้นฐานว่า รีทวิตสามารถสะท้อนได้ว่าเรื่องใดเป็นที่น่าสนใจในหมู่ผู้ใช้งานทวิตเตอร์ ซึ่งจะนำไปสู่การสร้างโมเดลที่ใช้ในการอธิบายคุณลักษณะของทวิตที่จะถูกรีทวิตต่อไป ในความเป็นจริง ปัจจัยที่จะทำให้ทวิตถูกรีทวิตนั้นแบ่งเป็นสองส่วน ส่วนหนึ่งคือปัจจัยด้านบริบท (Context-based) อาทิ ผู้ใช้ที่โพสต์ (โดยดูจากจำนวน follower, จำนวน followee, จำนวนครั้งที่ถูกรีทวิต เป็นต้น) และเวลาที่โพสต์ อีกส่วนหนึ่งคือปัจจัยด้านเนื้อหา (Content-based) โดยเจาะดูเฉพาะที่ตัวเนื้อหาว่ามีความน่าสนใจเพียงใด ซึ่งงานวิจัยชิ้นนี้จะเน้นการวิเคราะห์ปัจจัยด้านเนื้อหาเท่านั้น

วิธีการดำเนินงานเริ่มต้นจากการนำ Dataset มาแบ่งออกเป็น 2 ส่วน คือ Training Set (75%) และ Testing Set (25%) จากนั้นจึงสกัดเอา Feature ในเชิงเนื้อหาของแต่ละทวิตออกมา เพื่อใช้เป็นปัจจัยในการวิเคราะห์ โดยใช้วิธีการ Logistic Regression ในการสร้างโมเดลจาก Training Set และใช้โมเดลที่ได้ในการวิเคราะห์โอกาสที่ทวิตหนึ่งๆจะถูกรีทวิตนอกจากนั้น Weighting Factor ที่ได้จาก Logistic Regression Model ยังสะท้อนถึงความเป็นปัจจัยส่งเสริมและปัจจัยต่อต้านของการถูกรีทวิตได้อีกด้วย

Feature ที่งานวิจัยนี้ใช้ศึกษา ได้แก่ ความเป็น Direct Message, การปรากฏของ Username-Hashtag-URL, การใช้ ? และ !, การใช้คำในเชิงบวกและลบ, การใช้ Emoticon ในเชิงบวกและลบ, การวิเคราะห์อารมณ์ของทวิต, คำที่ปรากฏในทวิต และหัวข้อของทวิต ซึ่งพบว่า การทำนายจะแม่นยำมากที่สุดเมื่อพิจารณาทุก Feature ประกอบกัน แต่หากพิจารณาแต่ละ Feature แยกจากกัน พบว่าการวิเคราะห์จากคำที่ปรากฏในทวิตและความเป็น Direct Message จะให้ผลที่แม่นยำที่สุด

5.2 Forecasting with Twitter Data^[7]

งานวิจัยชิ้นนี้พยายามที่จะหาคำตอบว่าค่าบ่งชี้อารมณ์ (public sentiment indicator) ที่ได้จากทวิตในแต่ละวันสามารถนำไปปรับปรุงการพยากรณ์ตัวบ่งชี้ต่างๆทางด้านสังคม เศรษฐกิจ หรือ การค้าขายได้หรือไม่ โดยทำการเก็บข้อมูลและประมวลผลข้อมูลทวิตเตอร์ตั้งแต่เดือนมีนาคม 2011 จนถึงธันวาคม 2013 ในสอง domain หลักที่สนใจคือ ข้อมูลทวิตเตอร์ที่เกี่ยวข้องกับด้านตลาดหุ้นและด้านรายได้จากการขายบัตรเข้าชมภาพยนตร์สำหรับทั้งสองกลุ่มนี้ ทางทีมวิจัยได้สร้างและประเมินผลโมเดลต่างๆที่ใช้ในการทำนาย ซึ่งมีทั้งโมเดลที่ใช้และไม่ใช้ Attribute ที่เกี่ยวข้องกับทวิตเตอร์เข้ามาร่วมด้วย โดยอาศัยเทคนิค Summary tree ซึ่งเป็นเทคนิคในกลุ่ม Decision tree แบบหนึ่งในการทำเหมืองข้อมูลขนาดใหญ่ที่รวบรวมได้ และนำผลลัพธ์ที่ได้มาช่วยปรับปรุงความสามารถในการทำนายของอีกโมเดลหนึ่ง ซึ่งจากผลลัพธ์สามารถสรุปได้ว่า สำหรับข้อมูลด้านการเงินและตลาดหุ้น โมเดลในกลุ่ม nonlinear สามารถจัดการกับความไม่แน่นอนและใช้ประโยชน์จากข้อมูลทวิตเตอร์ได้ดี ในขณะที่โมเดลในกลุ่ม linear ไม่สามารถทำงานได้ ส่วนข้อมูลการซื้อขายบัตรเข้าชมภาพยนตร์พบว่าเทคนิค Support Vector Machine สามารถทำงานได้ดีที่สุด

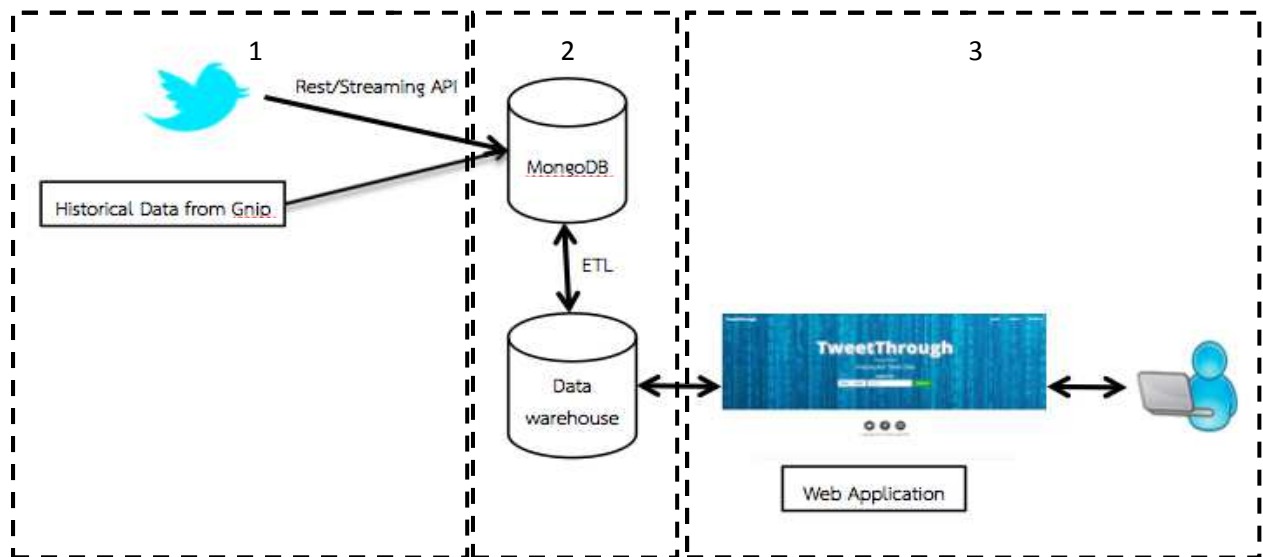
5.3 Twitter Under Crisis: Can we trust what we RT? ^[1]

การวิจัยนี้ศึกษาพฤติกรรมของผู้ใช้ทวิตเตอร์ในขณะเกิดภาวะผิดปกติ โดยสนใจกิจกรรมบนทวิตเตอร์ในช่วงการเกิดแผ่นดินไหวในชิลี ปี 2010 ผู้วิจัยได้ทดลองปล่อยข่าวจริง และข่าวลือในช่วงการเกิดแผ่นดินไหว แล้วติดตามลักษณะการกระจายตัวของข่าวลือ และข่าวจริงบนทวิตเตอร์ เพื่อนำมาเปรียบเทียบกัน และเพื่อดูความน่าเชื่อถือของทวิตเตอร์ ในฐานะแหล่งข่าวหนึ่งในภาวะผิดปกติ เช่น เช่น การวิเคราะห์จำนวนทวิต จำนวนการรีทวีต อัตราการแพร่กระจายเมื่อเวลาผ่านไป การสิ้นสุดการแพร่กระจาย เป็นต้น ซึ่งผู้วิจัยพบว่าการแพร่กระจายของข้อความที่เป็นข่าวลือมีพฤติกรรมแตกต่างจากการแพร่กระจายข่าวทั่วไป เพราะข่าวลือมักจะมีวาทกรรมคลุมเครือ ชวนให้ตั้งคำถามมากกว่าข่าวทั่วไป ดังนั้น จึงมีความเป็นไปได้ในการตรวจจับข่าวลือโดยใช้การวิเคราะห์หลายๆรูปแบบร่วมกัน

6. ขอบเขตความต้องการเชิงหน้าที่ของระบบ

- 1) ระบบสามารถค้นหา และแสดงผลการวิเคราะห์ข้อมูลตามการค้นหาโดยใช้คำสำคัญ (keyword), ข้อความทวิต หรือชื่อผู้ใช้ (User) ได้ ภายใต้ขอบเขตข้อมูลที่นำเข้ามา
- 2) ระบบสามารถวิเคราะห์ และแสดงผลการวิเคราะห์ต่างๆได้ ดังนี้
 - 2.1) ค่าสถิติเบื้องต้น ได้แก่
 - จำนวนทวิตที่เกี่ยวข้องทั้งหมด
 - จำนวนทวิตแบ่งตามกิจกรรมต่างๆ นั่นคือ ทวิต รีทวีต หรือตอบกลับ
 - จำนวนครั้งที่ทวิตที่สนใจปรากฏบนหน้า feed ของผู้ใช้งานทั้งหมด
 - จำนวนผู้ใช้ทวิตเตอร์ทั้งหมดที่เกี่ยวข้องกับการเผยแพร่ข้อความ (ในฐานะผู้ทวิต, ผู้รีทวีต หรือผู้ทวิตตอบกลับ)
 - 2.2) การวิเคราะห์อัตราเร็วของการเผยแพร่ผ่านทางกราฟแสดงอัตราการทวิตข้อความตามเวลาต่างๆ โดยผู้ใช้งานสามารถปรับความละเอียดของเวลาได้ เช่น เดือน สัปดาห์ วัน เป็นต้น และสามารถเลือกแสดงเป็นช่วงเวลาตามที่ต้องการได้
 - 2.3) การแสดงรายชื่อผู้ใช้ทวิตเตอร์ทั้งหมดที่เกี่ยวข้องกับการเผยแพร่ข้อความ พร้อมทั้งระบุผู้ใช้ที่มีอิทธิพลสูงสุดในการเผยแพร่ข้อมูล (ผู้ใช้ที่มีผู้ติดตามสูงสุด และผู้ใช้ที่ถูกรีทวีตข้อความสูงสุด)
 - 2.4) การแสดงกิจกรรมของกลุ่มผู้ใช้ที่สนใจ (Interesting Account) ที่เกี่ยวข้องกับการวิเคราะห์
 - 2.5) การแสดงรายการของทวิต พร้อมทั้งอุปกรณ์ที่ใช้ โดยแสดงเรียงลำดับแบบต่างๆ ได้แก่
 - เรียงลำดับตามเวลาการทวิต
 - เรียงลำดับตามจำนวนครั้งการถูกรีทวีต
 - เรียงลำดับตามจำนวน Follower ของผู้ทวิต
 - 2.6) การแสดงสัดส่วนของอุปกรณ์ที่ใช้ทวิตโดยสามารถเลือกแสดงตามกิจกรรมต่างๆได้
- 3) ระบบสามารถเชื่อมโยงบัญชีผู้ใช้ของผู้เผยแพร่กลับไปยังหน้า profile ของบัญชีผู้ใช้นั้นๆได้

7. แนวทางในการพัฒนาและเครื่องมือที่เกี่ยวข้อง



ภาพที่ 7 สถาปัตยกรรมระบบ

ระบบสามารถแบ่งออกเป็นส่วนใหญ่ๆได้ 3 ส่วน ได้แก่



1. การนำเข้าข้อมูล ที่มาของข้อมูลประกอบด้วย 2 ส่วน ได้แก่
 - 1.1. ข้อมูลจากผู้ให้บริการข้อมูลทางทวิตเตอร์ นั่นคือ Gnip สำหรับข้อมูลในอดีตซึ่งไม่สามารถใช้ทวิตเตอร์ เอฟไอ ในการเก็บข้อมูลได้ จึงจำเป็นต้องซื้อข้อมูลจากผู้ให้บริการ
 - 1.2. ข้อมูลที่ได้จากการ stream โดยใช้ Streaming API จากทวิตเตอร์ เพื่อเก็บข้อมูลปัจจุบันถึงอนาคตเกี่ยวกับเรื่องราวที่สนใจหรือคาดว่าจะเป็นอย่างลวง
2. การจัดเก็บข้อมูล แบ่งออกเป็น 2 ส่วน ได้แก่
 - 2.1. NoSql ทางกลุ่มได้เลือกใช้ mongoDB ในการเก็บข้อมูลที่ได้จากการ stream และผู้ให้บริการข้อมูล เนื่องจากผลลัพธ์ที่ได้จากสองแหล่งนี้มีรูปแบบเป็น JSON และมี field ไม่เท่ากันในแต่ละแถวข้อมูล ซึ่งเหมาะสม mongoDB ซึ่งมีลักษณะเป็นฐานข้อมูลเชิงเอกสาร
 - 2.2. Data Warehouse เนื่องจากการวิเคราะห์ข้อมูลและออกรายงานต่างๆ เหมาะสมกับการกระทำงาน Data Warehouse ซึ่งเป็น structure database ดังนั้น ในส่วนนี้จึงเป็น database ส่วนที่ใช้เพื่อการวิเคราะห์ต่างๆ โดยใช้ pentaho integration tools เพื่อทำการ extract transform load (ETL) มาจาก mongoDB
3. การแสดงผลหรือหน้าจอติดต่อกับผู้ใช้ ส่วนติดต่อกับผู้ใช้มีลักษณะเป็นระบบเว็บ ซึ่งแสดงผลการวิเคราะห์ต่างๆ ตามคำค้นหาที่ผู้ใช้กรอกเข้ามาเพื่อทำการวิเคราะห์ผล ดังนี้
 - 3.1. ค่าสถิติทั่วไป ได้แก่
 - 3.1.1. posts จำนวนข้อความที่เกี่ยวข้องกับคำค้นหา นั่นคือมีคำค้นหานั้นๆปรากฏอยู่

- 3.1.2. จำนวนบัญชีผู้ใช้ที่มีส่วนเกี่ยวข้องกับการเผยแพร่ข้อมูลที่เกี่ยวข้องกับคำค้นหา
- 3.1.3. impression ผลรวมของผู้ติดตามของแต่ละบัญชีบุคคล
- 3.1.4. แผนภูมิวงกลมแสดงอัตราส่วนของข้อความ ว่าเป็นทวีต รีทวิต หรือการตอบกลับอย่างละเอียด
- 3.1.5. ข้อความที่มีคำค้นหาที่ถูกรีทวิตมากที่สุด 3 อันดับแรก
- 3.1.6. แผนภูมิวงกลมแสดงอัตราส่วนของโปรแกรม (application) ชนิดต่างๆที่บุคคลที่เกี่ยวข้องใช้ในการทวีต รีทวิต หรือตอบกลับข้อความ
- 3.2. อัตราการแพร่กระจาย (Speed and Life Cycle) แสดงกราฟจำนวนกิจกรรมต่างๆ (ทวีต รีทวิต หรือตอบกลับ) ในช่วงเวลาต่างๆ
- 3.3. บุคคล (Contributor or User) เป็นข้อมูลเกี่ยวกับผู้ที่เกี่ยวข้องกับการเผยแพร่ข้อมูลที่มีคำค้นหา
 - 3.3.1. บุคคลที่มี follower มากที่สุด
 - 3.3.2. บุคคลที่โพสต์ข้อความที่ถูกรีทวิต มากที่สุด
 - 3.3.3. รายการบุคคลที่เกี่ยวข้องกับการเผยแพร่คำค้นหานี้ โดยแสดงถึงจำนวนครั้งในแต่ละกิจกรรม (ทวีต รีทวิต หรือตอบกลับ)
 - 3.3.4. รายการบุคคลที่เกี่ยวข้องกับการเผยแพร่คำค้นหานี้ ที่เป็นที่น่าสนใจของสายนิเทศศาสตร์ โดยมีการแบ่งเป็นกลุ่มๆตามที่ทางนิเทศศาสตร์ต้องการวิเคราะห์ โดยแสดงถึงจำนวนครั้งในแต่ละกิจกรรม (ทวีต รีทวิต หรือตอบกลับ)
 - 3.3.5. กราฟแสดงการแพร่กระจายข้อมูลโดยแสดงให้เห็นภาพการกระจายข้อมูล ตามกลุ่มคนต่างๆที่ทางนิเทศศาสตร์สนใจ เมื่อเวลาเปลี่ยนไป
- 3.4. Timeline แสดงรายละเอียดข้อความต่างๆที่เกี่ยวข้องกับคำค้นหาโดยเรียงตามเวลา

8. ขั้นตอนการดำเนินงาน

8.1 แผนภูมิ Gantt แสดงขั้นตอนการดำเนินงาน

Task Name	Start Date	End Date	Q3			Q4			Q1			Q2		
			Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
วางแผน	07/01/14	10/31/14				วางแผน								
ศึกษาที่มาและความสำคัญของโครงการ	07/01/14	07/31/14	ศึกษาที่มาและความสำคัญของโครงการ											
รวบรวมความต้องการของผู้ใช้	07/04/14	08/31/14	รวบรวมความต้องการของผู้ใช้											
ศึกษาเครื่องมือที่เกี่ยวข้องกับการพัฒนาระบบเว็บ	07/21/14	10/05/14	ศึกษาเครื่องมือที่เกี่ยวข้องกับการพัฒนาระบบเว็บ											
กำหนดวัตถุประสงค์และเป้าหมาย	08/01/14	09/21/14	กำหนดวัตถุประสงค์และเป้าหมาย											
กำหนดขอบเขตของโครงการ	09/01/14	09/07/14	กำหนดขอบเขตของโครงการ											
ศึกษาแหล่งขายข้อมูลทวิตเตอร์	08/04/14	10/05/14	ศึกษาแหล่งขายข้อมูลทวิตเตอร์											
ศึกษาโปรแกรมวิเคราะห์ข้อมูลทวิตเตอร์	09/08/14	09/21/14	ศึกษาโปรแกรมวิเคราะห์ข้อมูลทวิตเตอร์											
ประชุมหาแนวทางการซื้อข้อมูล	10/06/14	10/06/14	ประชุมหาแนวทางการซื้อข้อมูล											
ศึกษางานวิจัยที่เกี่ยวข้อง	09/15/14	10/31/14	ศึกษางานวิจัยที่เกี่ยวข้อง											
ออกแบบระบบ	09/08/14	11/09/14				ออกแบบระบบ								
ออกแบบสถาปัตยกรรมของระบบ	09/08/14	11/09/14	ออกแบบสถาปัตยกรรมของระบบ											
ออกแบบส่วนต่อประสานกับผู้ใช้	09/08/14	09/15/14	ออกแบบส่วนต่อประสานกับผู้ใช้											
ประชุมนำเสนอส่วนต่อประสานกับผู้ใช้ ครั้งที่ 1	09/16/14	09/16/14	ประชุมนำเสนอส่วนต่อประสานกับผู้ใช้ ครั้งที่ 1											
ปรับแก้ส่วนต่อประสานงานกับผู้ใช้	09/08/14	10/15/14	ปรับแก้ส่วนต่อประสานงานกับผู้ใช้											
ประชุมนำเสนอส่วนต่อประสานกับผู้ใช้ ครั้งที่ 2	10/16/14	10/16/14	ประชุมนำเสนอส่วนต่อประสานกับผู้ใช้ ครั้งที่ 2											
ออกแบบโครงสร้างคลังข้อมูล	09/08/14	09/22/14	ออกแบบโครงสร้างคลังข้อมูล											
พัฒนาระบบ	10/13/14	01/24/15				พัฒนาระบบ								
ติดตั้งระบบจัดเก็บข้อมูลทวิตเตอร์โดยใช้ Streaming API	10/13/14	11/11/14	ติดตั้งระบบจัดเก็บข้อมูลทวิตเตอร์โดยใช้ Streaming API											
ซื้อและนำเข้าข้อมูลเข้าสู่ระบบ	11/24/14	12/01/14	ซื้อและนำเข้าข้อมูลเข้าสู่ระบบ											
ทำการ ETL ข้อมูลโดยใช้ pentaho	11/18/14	12/08/14	ทำการ ETL ข้อมูลโดยใช้ pentaho											
พัฒนาด้านแบบของระบบ	10/17/14	11/22/14	พัฒนาด้านแบบของระบบ											
ประชุมนำเสนอต้นแบบของระบบ ครั้งที่ 1	11/23/14	11/23/14	ประชุมนำเสนอต้นแบบของระบบ ครั้งที่ 1											
ปรับแก้ต้นแบบของระบบ	11/24/14	11/30/14	ปรับแก้ต้นแบบของระบบ											
ประชุมนำเสนอต้นแบบของระบบ ครั้งที่ 2	12/01/14	12/01/14	ประชุมนำเสนอต้นแบบของระบบ ครั้งที่ 2											
พัฒนาระบบจริง	12/02/14	01/24/15	พัฒนาระบบจริง											

Task Name	Start Date	End Date	Q3			Q4			Q1			Q2		
			Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
 ทดสอบระบบ	01/19/15	02/22/15												
ทดสอบระบบโดยผู้พัฒนา	01/26/15	02/07/15												
จัดทำเอกสารและคู่มือการใช้งาน	01/19/15	02/01/15												
ทดสอบระบบโดยผู้ใช้	01/25/15	02/07/15												
ปรับแก้ระบบเพิ่มเติม	02/09/15	02/22/15												
 เขียนบทความวิชาการ	02/23/15	04/02/15												
ศึกษางานวิจัยที่เกี่ยวข้อง	02/23/15	03/08/15												
สรุปหัวข้อและเค้าโครงของงานวิจัย	03/09/15	03/15/15												
ออกแบบ ทดลอง และสรุปผล	03/16/15	03/26/15												
เขียนบทความวิชาการ	03/09/15	04/02/15												

8.2 ตารางแสดงขั้นตอนการดำเนินงานในช่วงเดือนต่างๆ

เดือน	งานที่ทำ
กรกฎาคม	เริ่มต้นศึกษารายละเอียดโครงการ
	รวบรวมความต้องการจาก ผศ. พิจิตรา
	กำหนดวัตถุประสงค์ และเป้าหมายของโครงการ
	ศึกษาเครื่องมือที่เกี่ยวข้องกับการพัฒนาเว็บแอปพลิเคชัน ได้แก่ ทวิตเตอร์เบื้องต้น และ Twitter API
	ค้นหาแหล่งขายข้อมูลทวิตเตอร์เพื่อนำเสนอต่อ ผศ.พิจิตรา
สิงหาคม 2557	นำเสนอแหล่งขายข้อมูลทวิตเตอร์ต่อ ผศ.พิจิตรา และเก็บความต้องการเพิ่มเติม
	ศึกษาโปรแกรมที่มีลักษณะคล้ายคลึงกัน
	ศึกษา MongoDB และทดลอง stream ข้อมูลจากทวิตเตอร์
	ติดต่อแหล่งขายข้อมูลทวิตเตอร์เพื่อสืบราคา และนำเสนอต่อ ผศ.พิจิตรา
กันยายน 2557	กำหนดขอบเขตของระบบให้ชัดเจน
	ออกแบบสถาปัตยกรรมระบบ
	จัดทำหน้าเว็บ (Mock-up)
	นำเสนอส่วนต่อประสานกับผู้ใช้(Mock-up) ต่อ ผศ. พิจิตรา ครั้งที่ 1 เพื่อทำการปรับแก้ต่อไป
	ศึกษาเครื่องมือที่จำเป็นต้องใช้เพิ่มเติม <ul style="list-style-type: none"> - Laravel Frameworks เพิ่มเติม เช่น package เพิ่มเติม - Pentaho
	ออกแบบโครงสร้างคลังข้อมูล
ตุลาคม 2557	แก้ไขโครงสร้างและหน้าเว็บไซต์ (Mock-up) เพื่อนำเสนอ ผศ.พิจิตรา
	นำเสนอหน้าเว็บ (Mock-up) ต่อ ผศ.พิจิตรา ครั้งที่ 2
	ติดตั้งระบบจัดเก็บข้อมูลทวิตเตอร์โดยใช้ streaming API
	ปรับแก้โครงสร้างคลังข้อมูล
พฤศจิกายน 2557	ซื้อข้อมูลและนำเข้าข้อมูลที่ได้มายังคลังฐานข้อมูล
	ทำการ ETL ข้อมูล โดยใช้ pentaho
	พัฒนาต้นแบบระบบ (Prototype)
	นำเสนอต้นแบบระบบต่อ ผศ. พิจิตรา เพื่อทำการปรับแก้ต่อไป
ธันวาคม 2557 – มกราคม 2558	พัฒนาระบบให้สมบูรณ์
มกราคม 2558	ทดสอบความถูกต้องของระบบโดยผู้พัฒนา และปรับแก้ไขหากพบข้อผิดพลาดเพิ่มเติม
	นำเสนอระบบทั้งหมดต่อ ผศ.พิจิตรา และปรับแก้ระบบเพิ่มเติม
	จัดทำเอกสาร และคู่มือการใช้งานของระบบที่สร้างขึ้น

เดือน	งานที่ทำ
กุมภาพันธ์ 2558	ทดสอบความถูกต้องของระบบโดยผู้ใช้ และปรับแก้ไขหากพบข้อผิดพลาดเพิ่มเติม
มีนาคม 2558	ศึกษางานวิจัยที่เกี่ยวข้อง
	สรุปหัวข้อและเค้าโครงของงานวิจัยที่จะทำเป็นบทความวิชาการ
	ดำเนินการออกแบบ ทดลอง สรุปผล และเขียนบทความวิชาการ
เมษายน 2558	เขียนบทความวิชาการ

9. ประโยชน์ที่คาดว่าจะได้รับ

9.1 ได้ระบบเว็บที่สามารถช่วยเหลือนักวิจัยทางด้านนิเทศศาสตร์ในการศึกษา ลักษณะ และ พฤติกรรมของชาวไลบรารีสังคมออนไลน์ในแง่ต่างๆ ได้สะดวกและรวดเร็วมากขึ้น

9.2 ระบบสามารถวิเคราะห์ข้อมูลจากทวีตเตอร์ออกมาแสดงผลในมุมมองใหม่ เช่น กราฟแสดงการ กระจายข้อมูลในกลุ่มคนต่างๆตามเวลาที่เปลี่ยนแปลงไป เป็นต้น

9.3 ผู้พัฒนาระบบเว็บได้ประสบการณ์การพัฒนาระบบเว็บจริง รวมถึงความรู้ต่างๆที่เกี่ยวข้อง เช่น Twitter API, คลังข้อมูล, การออกแบบหน้าจอแสดงการวิเคราะห์ผล และการติดตั้งระบบ เป็นต้น

10. รายการอ้างอิง

- [1] Marcelo Mendoza, Barbara Poblete, Carlos Castillo: "**Twitter Under Crisis: Can we trust what we RT?**". In SOMA 2010: KDD Workshop on Social Media Analytics, Washington, DC. July 2010.
- [2] issuu. **คู่มือการใช้ Twitter เบื้องต้น**. [online]. [cited 27 Oct, 2014].
Available from: http://issuu.com/sudjing/docs/twitter_tanasat/4
- [3] 1keydata. **Data Warehousing**. [online]. [cited 30 Oct, 2014].
Available from: <http://www.1keydata.com/datawarehousing/datawarehouse.html>
- [4] THAI BLACKBERRY CLUB. **ทำความรู้จักกับ Twitter เบื้องต้น**. [online]. [cited 30 Oct, 2014].
Available from: <http://thaibbclub.com/forums/twitter-t27722.html>
- [5] คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี. **การออกแบบระบบ Data Warehousing**. [online]. [cited 30 Oct, 2014].
Available from: <http://www.bus.rmutt.ac.th/~suwat/Data%20Warehousing.pdf>
- [6] Twitter. **Documentation**. [online]. [cited 30 Oct, 2014].
Available from: <https://dev.twitter.com/overview/documentation>
- [7] MARTA ARIAS, ARGIMIRO ARRATIA, and RAMON XURIGUERA: "**Forecasting with Twitter Data**". ACM Transactions on Intelligent Systems and Technology, Vol. 5, No. 1, Article 8, Publication date: December 2013.
- [8] Nasir Naveed et al. "**Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter**". ACM Web Sci Conference 2011. June 14-17, 2011
- [9] ThaiMongo. **ทำความรู้จัก NoSQL คืออะไร** [online]. [cited 28 Oct, 2014].
Available from: <http://www.thaimongo.com/บทความ-mongodb/37-ทำความรู้จัก-nosql-คืออะไร.html>
- [10] crowdoutsourcing. **มาใช้ NOSQL กันเถอะ** [online]. [cited 28 Oct, 2014].
Available from: <http://no-sql.blogspot.com/2010/02/nosql-introduction-to-nosql.html>
- [11] Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat. **Data Mining with Microsoft SQL Server 2008**. Wiley. October, 2008.