

สรุป

ประเภทของข้อมูล

- **ชุดข้อมูล** คือ กลุ่มของ data objects ประกอบด้วย Attributes ที่บอกลักษณะ

ตารางที่ 2.1. ชุดข้อมูลนิสิต (Student Information Data Set)

Attributes

ลำดับที่	รหัสนิสิต	ชั้นปี	เกรดเฉลี่ยสะสม
1	1034261	2	2.75
2	1034262	3	3.24
3	1034263	2	3.51
4	1034265	1	2.99
5	1034266	3	3.12

Data objects

Attributes ของ Data objects จำนวน เปลี่ยน แปลง ได้ตลอด
เช่น อุณหภูมิพื้นผิวโลก-ผิวดิน

การแบ่งประเภทของ Attributes โดย Operation ของระบบจำนวน

ตารางที่ 2.2. ชนิดของแอททริบิวต์

ชนิดของแอททริบิวต์	คำอธิบาย	ตัวอย่าง	โอเปอเรชัน
Categorical (เชิงคุณภาพ)	Nominal คำของ Nominal attribute สามารถใช้ในการแยกแยะค่าด้วยโอเปอเรชันได้ ด้วยโอเปอเรชัน Distinctness ($=$, \neq)	รหัสไปรษณีย์ รหัสพนักงาน สีตา เพศ	ฐานนิยม, entropy, contingency correlation, Chi-squared test
	Ordinal ก. binary ด. Discrete จ. มีคุณสมบัติและโอเปอเรชัน Distinctness เช่นเดียวกับ Nominal attributes และ คำของ Ordinal attribute สามารถใช้ในการเรียงลำดับค่าด้วยโอเปอเรชันได้ ด้วยโอเปอเรชัน Order ($<$, $>$)	ความแข็งแรงของแร่ธาตุ, เกรด {A, B+, B, C+, C, D+, D, F},	มัธยฐาน, เปอร์เซนต์ไทล์, rank correlation, run tests, sign tests
Numeric (เชิงปริมาณ)	Interval มีคุณสมบัติและโอเปอเรชัน Distinctness และ Order เช่นเดียวกับ Nominal attributes และ Ordinal attributes นอกจากนี้ ความแตกต่างระหว่าง interval attributes สองค่า คำนวณได้ด้วยโอเปอเรชัน Addition ($+$, $-$) สามารถตีความได้ กล่าวคือ interval attributes จะมีหน่วยของการวัด	อุณหภูมิในหน่วย องศาเซลเซียส หรือ องศาฟาเรนไฮต์, วันที่ตามปฏิทิน	ค่าเฉลี่ย, ส่วนเบี่ยงเบนมาตรฐาน, Pearson's correlation, t-test, F-test
	Ratio ข. Continuous มีคุณสมบัติและโอเปอเรชัน Distinctness, Order, และ Interval เช่นเดียวกับ Nominal attributes, Ordinal attributes, และ Interval attributes นอกจากนี้ อัตราส่วนของ ratio attributes ซึ่งคำนวณได้โดยใช้โอเปอเรชัน Multiplication (\times , $/$) สามารถตีความได้	อุณหภูมิในหน่วยเคลวิน (Kelvin), อายุ, มวล, ความยาว, กระแสไฟฟ้า	ค่าเฉลี่ยเรขาคณิต, ค่าเฉลี่ยฮาร์โมนิก, เปอร์เซ็นต์ความผันแปร

การแบ่งประเภท Attributes. ด้วยจำนวนของค่าที่เป็นไปได้

1. Discrete Attribute ที่มีค่าเป็นไปไม่ได้ไม่จำกัดจำนวนแต่สามารถนับแยกได้ **มีค่าได้กับ Attribute เช่น คนขาว**
2. Binary Attribute มีเพียง 2 ค่า **0/1, จริง/เท็จ, ใช่/ไม่ใช่, ชาย/หญิง**
3. Continuous Attribute มีค่าเป็นจำนวนจริง **มีค่าได้กับ Attributes เช่น ปริมาณ**

ชนิดของ Data sets มี 3 ประเภท

1. ข้อมูล Record data

ข้อมูลที่ใช้ในการคำนวณเชิงข้อมูล มักอยู่ในรูปแบบของ Record data ซึ่งประกอบด้วย Attributes หรือ field ข้อมูลแต่ละค่า

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Soda, Diapers, Milk

(b) Transaction data.

Projection of Color	Projection of Flavor	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	beer	soda	bread	milk	diapers	popcorn	lunchmeat
Document 1	3	0	5	0	2	6	0
Document 2	0	7	0	2	1	0	0
Document 3	0	1	0	0	1	2	2

(d) Document-term matrix.

2. ข้อมูลกราฟ

กราฟเป็นโครงสร้างข้อมูลสำหรับแสดงความสัมพันธ์ระหว่าง Data object เรามักใช้กับกราฟ 2 กรณี

1. แสดงความสัมพันธ์ระหว่าง Data object

Useful Links:

- Bibliography
- Other Useful Web sites
 - ACM SIGKDD
 - KDDmag
 - The Data Mine

Knowledge Discovery and Data Mining Bibliography
(Get updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Srikant, "Advances in Knowledge Discovery and Data Mining", AAAI Press/The MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

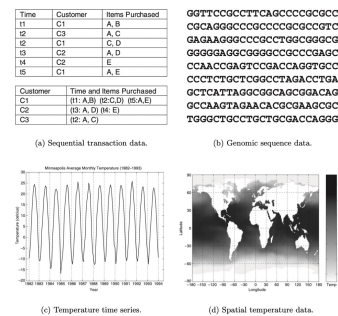
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Mathews, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for Knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

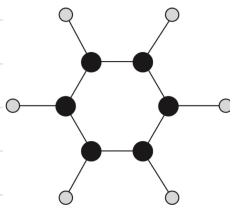
3. ข้อมูลแบบมีลำดับ

ข้อมูลบางชนิดประกอบด้วย Attributes ที่มีค่าตามลำดับ เช่น เวลาอันต่อเนื่องกัน



(a) Linked web pages.

2. Data object มีโครงสร้างแบบกราฟ



(b) Benzene molecule.

คุณภาพข้อมูล

- การได้มาซึ่งคุณภาพข้อมูล เทคนิคการจัดการกับคุณภาพของข้อมูล แบ่งได้ 2 กลุ่ม
1. เทคนิคสำหรับ การตรวจจับและการแก้ไขปัญหามาตรฐานข้อมูล
 2. การใช้เทคนิคการกำหนดค่าข้อมูลที่เหมาะสมต่อข้อมูลคุณภาพต่ำ

การวัดผลกระทบทางธุรกิจ และ การเก็บ ข้อมูล

ปัญหาคุณภาพข้อมูลเกิดได้จากผลกระทบของมนุษย์ ซึ่งจำกัดของอุปกรณ์ เช่น ข้อมูลรบกวน, ข้อมูลที่ซ้ำ, การนำข้อมูล, การที่ข้อมูล และ การไม่แม่นยำ

- การวัดผลกระทบทางธุรกิจ
- การวัดผลกระทบจากการเก็บข้อมูล

ปัญหาคุณภาพที่เกี่ยวกับองค์ประกอบข้อมูล

ข้อมูลที่มีคุณภาพสูง คือข้อมูลที่นำมาใช้ในการประยุกต์ใช้งาน ปัญหาคุณภาพของข้อมูลของการได้มาซึ่งข้อมูลที่ไปของข้อมูล ทำให้ไปใช้งาน มี 3 อย่าง

1. Timeliness เวลาในการเก็บ
2. Relevance ข้อมูลทุกชั้นที่จำเป็นต่อการวิเคราะห์
3. Knowledge about the Data คุณภาพเอกสาร

PDCA Present

- ขั้น 8 คือแผนการ พัฒนาแผนใหม่ ผู้จัดการ หรือหัวหน้า
- ขั้น 2 คือแผนการ ปฏิบัติการใหม่ ผู้แก้ไขต่อไป