

# **Data Mining**

## **Text Mining (Part B)**

**Dr. Jason T.L. Wang, Professor**  
**Department of Computer Science**  
**New Jersey Institute of Technology**

# Where am I?

- Part A presents the tf-idf algorithm for identifying and extracting keywords in text documents.
- Part B presents a text mining algorithm and an example.

# Text Mining: Keyword-Based Association Discovery

Here we represent each document in the database **D** by a set of keywords. Two documents might have different (numbers of) keywords, depending on whether the keywords occur in the documents.

Let  $X, Y$  be keywords.

Support for rule  $X \rightarrow Y$ : The number of documents in the database that contain both  $X$  and  $Y$ .

Confidence for rule  $X \rightarrow Y$ : The percentage of documents in the database containing  $X$  that also contain  $Y$ .

Example association rules:

query: “Find all associations including *gold* and any country”

result:

(gold, copper) → Canada [Support 5, Confidence 0.556]

(gold, silver) → USA [Support 18, Confidence 0.692]

....

Use Apriori algorithm to find these association rules, treating each document as a transaction and each keyword as an item in the transaction.

# Text Mining Example

(support > 1)

Database D

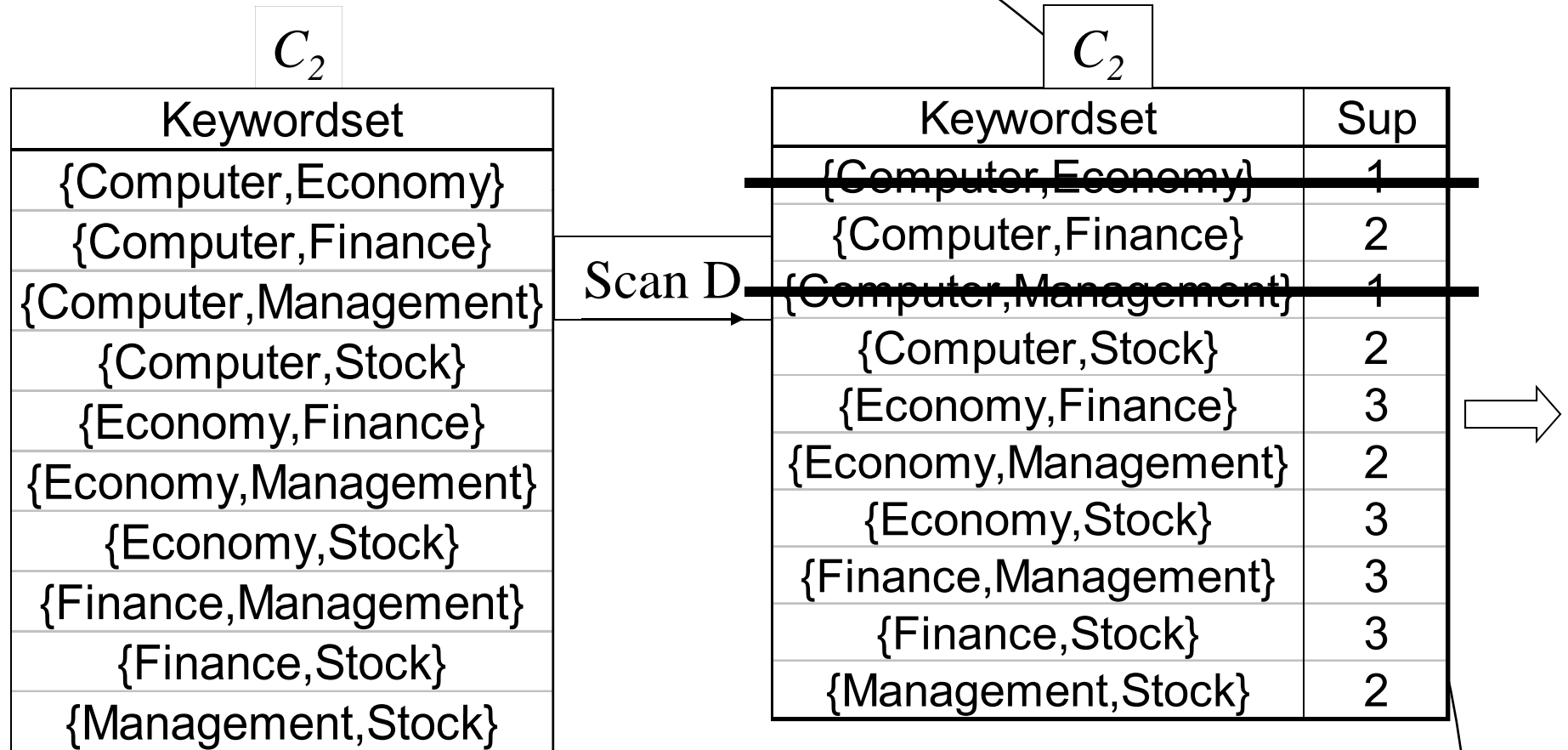
DID	Keyword
D1	Computer, Stock
D2	Economy, Finance, Stock
D3	Computer, Finance, Management
D4	Computer, Economy, Finance, Stock
D5	Economy, Finance, Management
D6	Finance, Management, Stock
D7	Economy, Management, Stock

Scan D

$C_1, L_1$

Keywordset	Sup
{Computer}	3
{Economy}	4
{Finance}	5
{Management}	4
{Stock}	5

# Text Mining Example



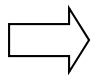
# Text Mining Example

$L_2$

Keywordset	Sup
{Computer, Finance}	2
{Computer, Stock}	2
{Economy, Finance}	3
{Economy, Management}	2
{Economy, Stock}	3
{Finance, Management}	3
{Finance, Stock}	3
{Management, Stock}	2

$C_3$

Keywordset
{Computer, Finance, Stock}
{Economy, Finance, Management}
{Economy, Finance, Stock}
{Economy, Management, Stock}
{Finance, Management, Stock}



# Text Mining Example

$C_3$

Keywordset	Sup
<del>{Computer, Finance, Stock}</del>	<del>1</del>
<del>{Economy, Finance, Management}</del>	<del>1</del>
{Economy, Finance, Stock}	2
<del>{Economy, Management, Stock}</del>	<del>1</del>
<del>{Finance, Management, Stock}</del>	<del>1</del>

$L_3$

Keywordset	Sup
{Economy, Finance, Stock}	2



# **End of Text Mining Module (Part B)**

