TRƯỜNG ĐẠI HỌC VINH
KHOA CÔNG NGHỆ THÔNG TIN

BÀI GIẢNG MÔN HỌC

KHAI PHÁ DỮ LIỆU (Data Mining)
Chương 3: Khai phá luật kết hợp

Giảng viên: TS. Cao Thanh Sơn
Bộ môn các hệ thống thông tin
Email: ctsdhv@gmail.com

2017

Chương 3: Khai phá luật kết hợp

*Based on slides **Data Mining: Concepts and Techniques***
by
Jiawei Han, Micheline Kamber, and Jian Pei, 2011

*and slides **Introduction to Data Mining***
by
Tan, Steinbach, Kumar, 2005

Some illustrative images are downloaded from the Internet.

Nội dung

- Khai phá luật kết hợp
- Một số khái niệm cơ bản
- Phương pháp sinh ứng viên: Apriori
- Phương pháp không sinh ứng viên: FP-Growth

Khai phá luật kết hợp

- Từ một tập hợp các giao dịch, hãy tìm các quy tắc để dự đoán sự xuất hiện của một mục (item) dựa trên sự xuất hiện của các mục khác trong giao dịch

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ví dụ luật kết hợp

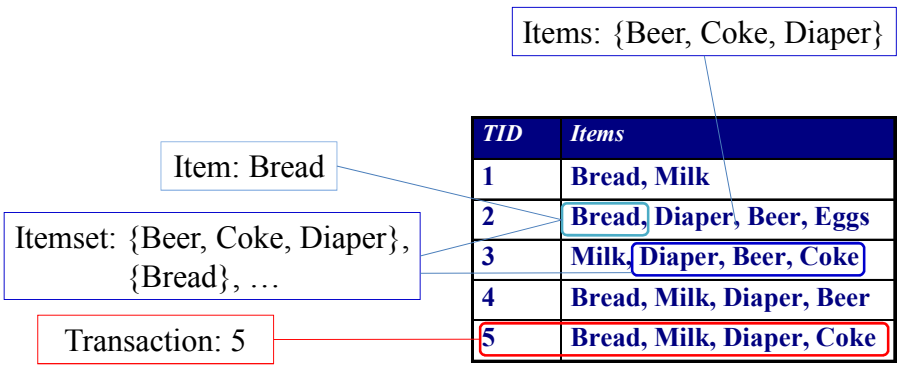
{Diaper} → {Beer},
{Milk, Bread} → {Eggs, Coke},
{Beer, Bread} → {Milk},

→ có nghĩa là cùng xuất hiện!

Một số khái niệm cơ bản

- **item**: phần tử đại diện cho một loại đối tượng dữ liệu (ví dụ: Milk, Bread, …)
- **items** $I = \{i_1, \ldots, i_m\}$: tập gồm các mục (item)
- **itemset** X: là tập gồm một hoặc nhiều mục (items), $X \subseteq I$
- **Transaction Database** D:
  - $D = \{T_1, T_2, \ldots, T_n\}$
  - $T_i \in D$: giao dịch (transaction), $T_i = (TID, X_T)$, $X_T \subseteq I$
  - D: cơ sở dữ liệu giao dịch (transaction database)
  - $|D|$: số giao dịch có trong D
- k-**itemset** $X = \{x_1, x_2, \ldots, x_k\}$: itemset có chứa k mục.

Một số khái niệm cơ bản



Một số khái niệm cơ bản

- Support(X): mức hỗ trợ ứng với tập mục X
  - Mức hỗ trợ tuyệt đối (absolute support hay support count): tần suất xuất hiện của tập thuộc tính X trong CSDL giao dịch D, ký hiệu $count(X)$
  - Mức hỗ trợ tương đối (relative support): tỷ lệ các giao dịch có chứa X trên tổng các giao dịch có trong D.

$$supp(X) = \frac{count(X)}{|D|}$$

Một số khái niệm cơ bản

- Ví dụ
  - X = {Bread, Milk}
  - $count(X)$ = 3
  - $supp(X)$ = 3/5 = 60%

TID	Items
T_1	Bread, Milk
T_2	Bread, Diaper, Beer, Eggs
T_3	Milk, Diaper, Beer, Coke
T_4	Bread, Milk, Diaper, Beer
T_5	Bread, Milk, Diaper, Coke

- Frequent itemset (tập phổ biến):
  - là tập mục có độ hỗ trợ lớn hơn một giá trị ngưỡng *min_sup* nào đó cho trước.

Một số khái niệm cơ bản

- Luật kết hợp (Association rule)
  - Gọi $X \rightarrow Y$ là một luật kết hợp nếu $X \subseteq I$ và $Y \subseteq I$ và $X \cap Y = \emptyset$
  - Khi X xuất hiện trong D thì sẽ kéo theo sự xuất hiện của Y với một tỷ lệ nào đó.
  - Ví dụ: {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Mức hỗ trợ (*support*) của luật kết hợp $X \rightarrow Y$ trong D là mức hỗ trợ của $X \cup Y$ trong D.

$$supp(X \rightarrow Y) = supp(X \cup Y) = \frac{count(X \cup Y)}{|D|}$$

- Độ tin cậy (*confidence*) của luật kết hợp $X \rightarrow Y$ trong D là tỷ lệ giao dịch có chứa cả X và Y với các giao dịch chỉ chứa X

$$conf(X \rightarrow Y) = \frac{count(X \cup Y)}{count(X)}$$

Một số khái niệm cơ bản

- Ví dụ luật kết hợp
  - $X \rightarrow Y$: {Milk, Diaper} → {Beer}
  - $count(X \rightarrow Y) = count(X \cup Y) = 2$
  - $count(X) = 3$
  - $|D| = 5$
  - $supp(X \rightarrow Y) = \frac{count(X \cup Y)}{|D|} = \frac{2}{5} = 0.4$
  - $conf(X \rightarrow Y) = \frac{count(X \cup Y)}{count(X)} = \frac{2}{3} = 0.67$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Cho một tập giao dịch T, mục đích của luật kết hợp là tìm tất cả các luật (được gọi là luật mạnh, hay luật có giá trị)
  - *support* ≥ *min_sup* threshold (ngưỡng hỗ trợ tối thiểu)
  - *confidence* ≥ *min_conf* threshold (ngưỡng tin cậy tối thiểu)

- Khai phá luật kết hợp

Input: CSDL giao dịch D, Các giá trị ngưỡng *min_sup*, *min_conf*

Output: Tất cả các luật mạnh

Phương pháp khai phá tập phổ biến

- Tìm tập phổ biến (finding frequent itemsets)
  - **Input**: tập các mục *I*, CSDL giao dịch *D*, các giá trị ngưỡng *min_sup*, *min_conf*
  - **Method (naïve algorithm)**:
    - đếm số lần xuất hiện của tất cả các tập con của *I* trong *D*
    - chỉ giữ lại các tập con thoả mãn *min_sup*
  - **Note**: không hiệu quả, có *m* mục cần tính 2^m-1 tập con của *I*.
- Sinh luật kết hợp từ các tập phổ biến (generating association rules from frequent itemsets)
  - **Input**: tập các tập phổ biến
  - **Method**:
    - với mỗi tập phổ biến *X* và $A \subset X$
    - kiểm tra luật $A \rightarrow (X - A)$ có thoả mãn *min_conf* hay không?

Apriori Algorithm [1]

Input:

- *D*, a database of transactions;
- *min_sup*, the minimum support count threshold.

Output: *L*, frequent itemsets in *D*.

Method:

```
(1)   L1 = find_frequent_1-itemsets(D);
(2)   for (k = 2; Lk−1 ≠ φ; k++) {
(3)       Ck = apriori_gen(Lk−1);
(4)       for each transaction t ∈ D { // scan D for counts
(5)           Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)           for each candidate c ∈ Ct
(7)               c.count++;
(8)       }
(9)       Lk = {c ∈ Ck | c.count ≥ min_sup}
(10)  }
(11)  return L = ∪k Lk;
```

Apriori Algorithm [1]

```
procedure apriori_gen(L_{k-1}:frequent (k-1)-itemsets)
(1)   for each itemset l_1 ∈ L_{k-1}
(2)     for each itemset l_2 ∈ L_{k-1}
(3)       if (l_1[1] = l_2[1]) ∧ (l_1[2] = l_2[2])
              ∧... ∧ (l_1[k-2] = l_2[k-2]) ∧ (l_1[k-1] < l_2[k-1]) then {
(4)         c = l_1 ⋈ l_2; // join step: generate candidates
(5)         if has_infrequent_subset(c, L_{k-1}) then
(6)           delete c; // prune step: remove unfruitful candidate
(7)         else add c to C_k;
(8)       }
(9)   return C_k;

procedure has_infrequent_subset(c: candidate k-itemset;
          L_{k-1}: frequent (k-1)-itemsets); // use prior knowledge
(1)   for each (k-1)-subset s of c
(2)     if s ∉ L_{k-1} then
(3)       return TRUE;
(4)   return FALSE;
```

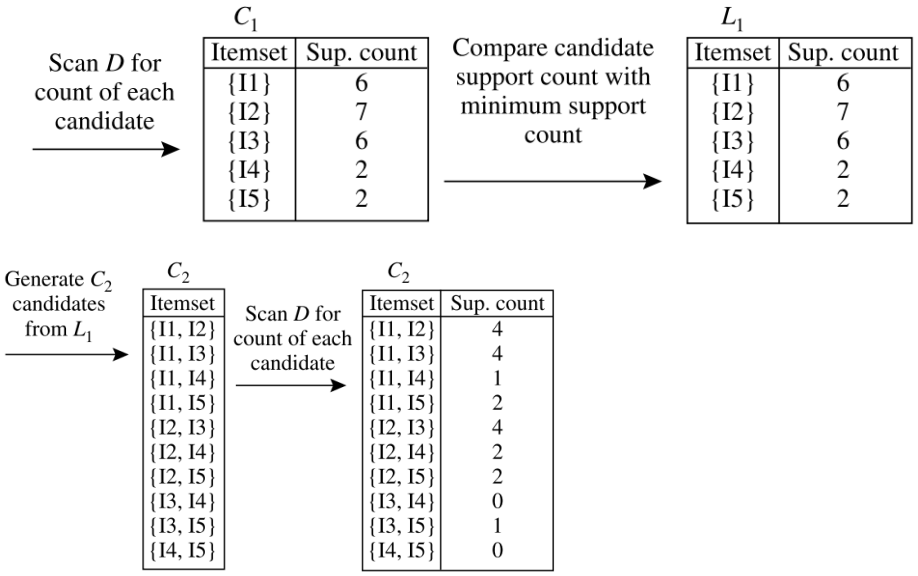

Ví dụ áp dụng giải thuật Apriori

- Cho CSDL giao dịch *D* như sau [1]:

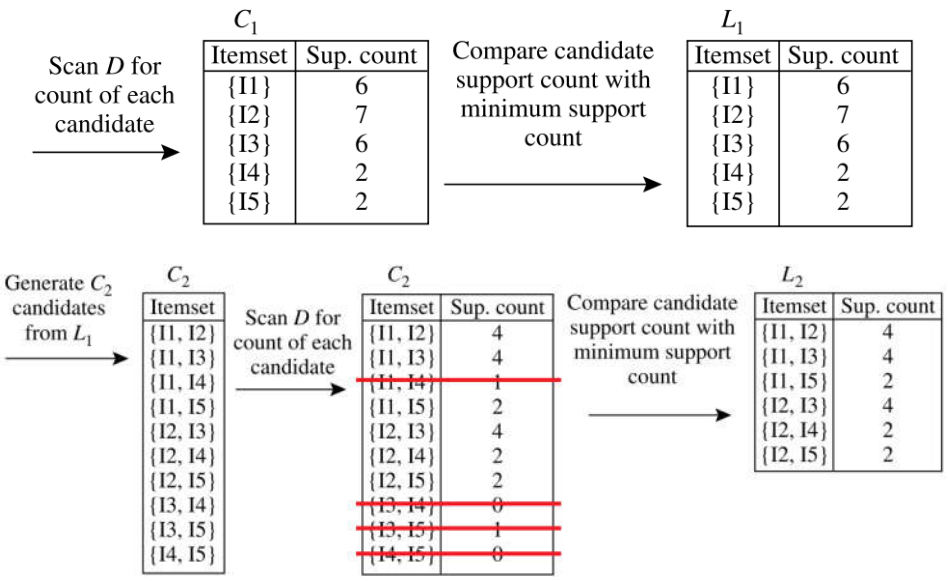
TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

- Giả sử *min_sup* = 2
- Để thuận tiện, giả sử rằng các mục trong một giao dịch hay trong tập mục được sắp xếp theo thứ tự từ điển.

Ví dụ áp dụng giải thuật Apriori (2)



Ví dụ áp dụng giải thuật Apriori (3)



Ví dụ áp dụng giải thuật Apriori (4)

Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$\bowtie \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$

$= \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\},$
$\{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}.$

Ví dụ áp dụng giải thuật Apriori (5)

Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$\bowtie \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$

$= \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\},$
$\{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}.$

Generate C_3 candidates from L_2 →

C_3 Itemset
{I1, I2, I3}
{I1, I2, I5}

Scan D for count of each candidate →

C_3 Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Compare candidate support count with minimum support count →

L_3 Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

The algorithm uses $L_3 \bowtie L_3$ to generate a candidate set of 4-itemsets, C_4. The join results in {{I1, I2, I3, I5}},

$C_4 = \phi$, and the algorithm terminates

Sinh luật kết hợp

- Với mỗi tập phổ biến *W* tìm được, sinh ra mọi tập con thực sự *X* (khác rỗng) của nó
- Với mỗi tập phổ biến *W* và một tập con thực sự *X* khác rỗng, sinh luật X → (*W* - *X*) thoả mãn *min_conf*

$$conf\big(X \rightarrow (W - X)\big) = \frac{count(W)}{count(X)} \geq min\_conf$$

Sinh luật kết hợp – Ví dụ [1]

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

- $\{I1, I2\} \Rightarrow I5, \quad conf = 2/4 = 50\%$
- $\{I1, I5\} \Rightarrow I2, \quad conf = 2/2 = 100\%$
- $\{I2, I5\} \Rightarrow I1, \quad conf = 2/2 = 100\%$
- $I1 \Rightarrow \{I2, I5\}, \quad conf = 2/6 = 33\%$
- $I2 \Rightarrow \{I1, I5\}, \quad conf = 2/7 = 29\%$
- $I5 \Rightarrow \{I1, I2\}, \quad conf = 2/2 = 100\%$

- *W* = {I1, I2, I5}
- Các luật kết hợp nào có thể sinh từ *W*?
- Tập con thực sự không rỗng của *W*:

  {I1, I2}, {I1, I5}, {I2, I5}, {I1}, {I2}, and {I5}

- Ngưỡng *min_conf* = 70%,
- Output?

Thuật toán Apriori

- Bài tập
  - *min_sup* = 2
  - *min_conf* = 60%

TID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Giải thuật FP-Growth

- Tìm tập phổ biến với giải thuật FP-Growth
  - Phát hiện các tập phổ biến không cần tạo các ứng viên
  - Đọc tài liệu [1, 2, 3]

Tài liệu tham khảo

[1] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*, 3rd ed, Morgan-Kaufmann Publishers, 2012.

[2] Nguyễn Hà Nam, Nguyễn Trí Thành, Hà Quang Thuỵ, *Giáo trình khai phá dữ liệu*, NXB Đại học Quốc gia Hà Nội, 2013.

[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005

[4] WEKA, *www.cs.waikato.ac.nz/ml/weka/*