Klasyfikacja materiałów dźwiękowych w oparciu o algorytm zbiorów domkniętych

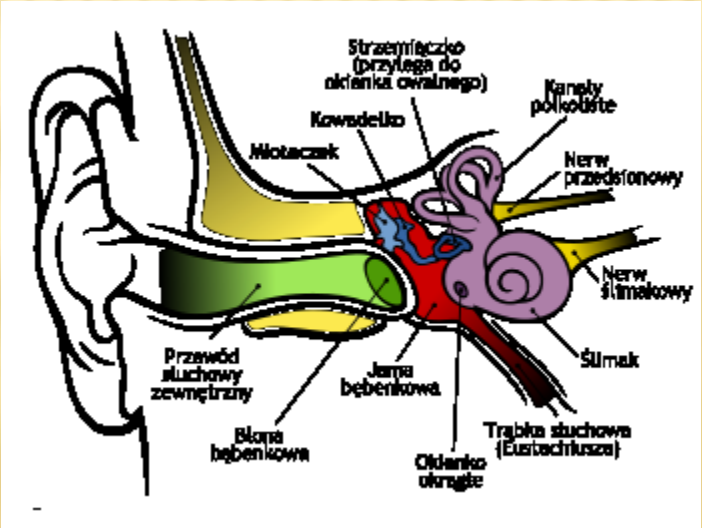
Grzegorz Szulik Instytut Matematyki UJ

PLAN REFERATU

1. Co to jest dźwięk?
2. Początki analizy i syntezy dźwięków
3. Koszyk dźwięków
4. Domknięte zbiory częste
5. Algorytm LCM
6. Przykład zastosowania

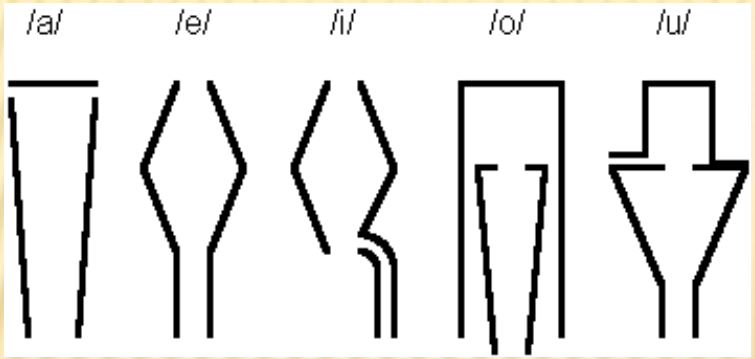
DŹWIĘK

- Wrażenie słuchowe wywołane falą akustyczną rozchodzącą się w ośrodku sprężystym
- Dźwięki słyszalne: 16Hz ÷ 20 kHz
- Głośność, wysokość i barwa dźwięku.



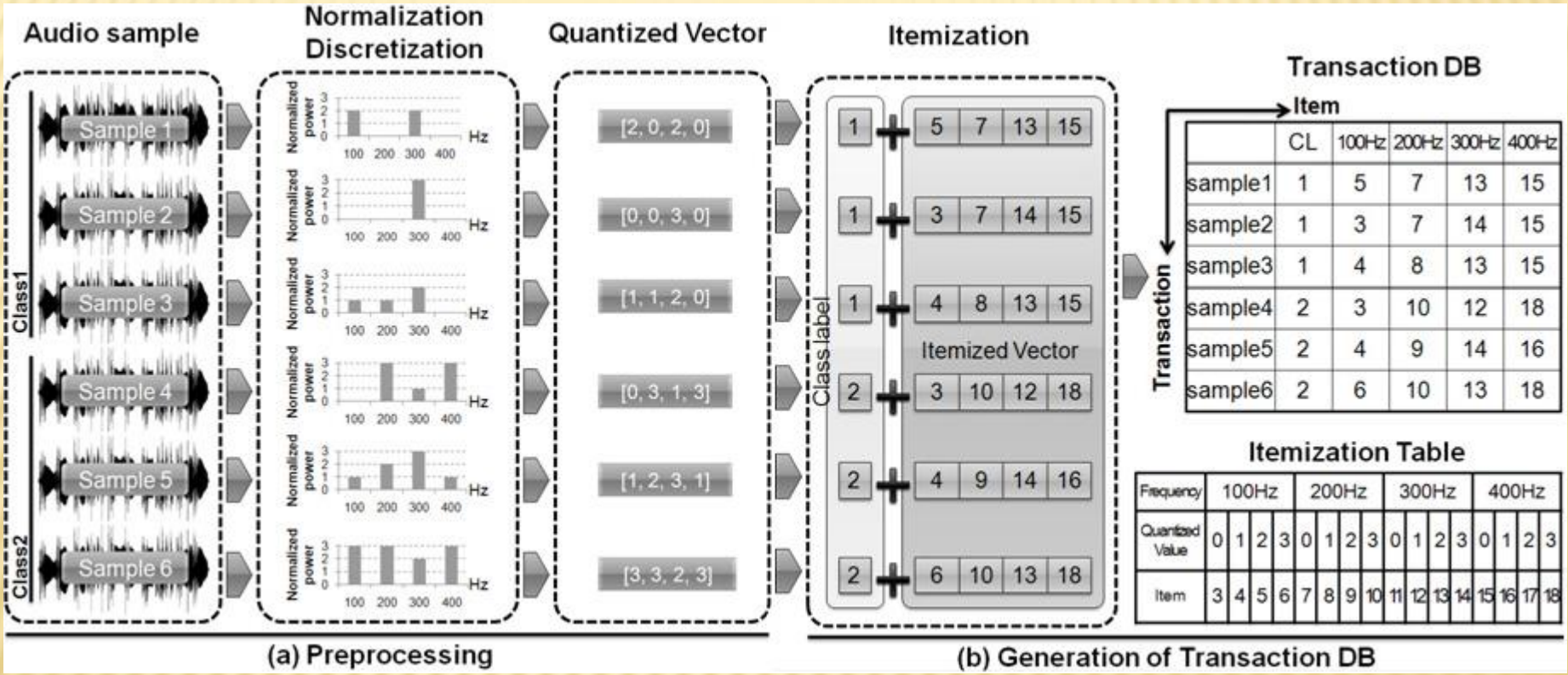
ANALIZA I SYNTEZA DŹWIĘKU

- Rezonatory Christiana Kratzensteina - 1779
- Wolfgang von Kempelen (1734-1804) - maszyna wymawiająca samogłoski



- Charles Wheatstone - 1837
- Alexander Graham Bell (1876 – telefon)

OD FALI DO KOSZYKA DŹWIĘKÓW



	CL	100Hz	200Hz	300Hz	400Hz
sample1	1	5	7	13	15
sample2	1	3	7	14	15
sample3	1	4	8	13	15
sample4	2	3	10	12	18
sample5	2	4	9	14	16
sample6	2	6	10	13	18

Itemization Table

Frequency	100Hz				200Hz				300Hz				400Hz			
Quantized Value	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Item	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

ZBIÓR TRANSAKCJI

- $E = \{1, 2, \ldots, n\}$ – zbiór artykułów
- T – zbiór transakcji
- $t_1, t_2 \ldots \in T$ – transakcje, $t_i \subset E$, $i = 1, 2, \ldots$
- $P \subset E$ nazywamy wzorcem
- transakcję, w której zawiera się wzorzec P nazywamy **wystąpieniem wzorca** P
- Zbiór wystąpień wzorca P oznaczamy przez $T(P)$
- Liczebność zbioru $T(P)$ nazywamy **częstością wzorca** P i oznaczamy przez $frq(P)$

WZORZEC CZĘSTY

- **Minimalnym wsparciem** nazywamy ustaloną liczbę naturalną θ
- Mówimy, że wzorzec *P* jest **częsty**, jeżeli jego częstość nie jest mniejsza niż θ
- Dla dwóch wzorców *P* i *Q* określamy relację równoważności:

$$P \sim Q \Leftrightarrow T(P) = T(Q)$$

- Wzorce maksymalny i minimalny ze względu na relację inkluzji w danej klasie równoważności nazywamy odpowiednio **wzorcem domkniętym** oraz **wzorcem kluczowym**.

DOMKNIĘCIE WZORCA

- Przez *F* oraz *C* oznaczamy odpowiednio zbiory wzorców częstych i domkniętych wzorców częstych.
- **Domknięciem** wzorca *P* nazywamy wzorzec

$$Clo(P) = \bigcap_{T \in T(P)} T$$

- ***i*-prefiksem** wzorca *P* nazywamy wzorzec $P(i) = P \cap \{1, 2, \ldots, i\}$
- Wzorzec *Q* nazywamy domkniętym rozszerzeniem wzorca *P*, gdy $Q = Clo(P \cup \{i\})$

WŁASNOŚCI DOMKNIĘĆ

- $P \subset Q \Rightarrow Clo(P) \subset Clo(Q)$
- $T(P) = T(Q) \Rightarrow Clo(P) = Clo(Q)$
- $Clo(Clo(P)) = Clo(P)$
- $Clo(P)$ jest najmniejszym wzorcem domkniętym zawierającym P
- wzorzec P jest domknięty wtedy i tylko wtedy, gdy $Clo(P) = P$

ALGORYTMY WYSZUKIWANIA WZORCÓW CZĘSTYCH

- AIS – Agrawal, Imieliński, Swami – 1993
- SETM – Houtsma, Swami – 1995
- APRIORI – Agrawal, Srikant – 1994
- DHP – Park, Chen, Yu – 1995
- Partition – Savasere, Omieciński, Navathe – 1995
- ECLAT – Zaki, Parthasarathy, Ogihara, Li – 1997
- FP-GROWTH – Han, Pei, Yin – 2000
- TREE-PROJECTION – Agarwal, Aggarwal, Prasad – 2000
- PASCAL – Bastide, Taouil, Pasquier, Stumme, Lakhal – 2000
- H-MINE – Pei, Lu, Nishio, Tang, Yang – 2001
- RELIM – Borgelt - 2005

ALGORYTMY WYSZUKIWANIA MAKSYMALNYCH WZORCÓW CZĘSTYCH

- MAX_MINER – Bayardo – 1998
- DepthProject – Agrawal, Aggarwal, Prasad – 2000
- MAFIA – Burdick, Calimlim, Gehrke – 2001
- GenMax – Gouda, Zaki – 2001

ALGORYTMY WYSZUKIWANIA DOMKNIĘTYCH WZORCÓW CZĘSTYCH

- CLOSE – Pasquier, Bastide, Taouil, Lakhal – 1999
- CLOSET – Pei, Han, Mao – 2000
- CHARM – Zaki, Hsiao - 2002

WYLICZANIE WZORCÓW DOMKNIĘTYCH

- proste wyliczanie, klasyfikacja, wyszukiwanie maksymalnego:
  $\rightarrow$ czas obliczeń: $O(|F|^2)$
  $\rightarrow$ potrzebna pamięć: $O(|F|)$
- zwykle $|F|$ jest dużo większe niż $|C|$
- pomysł $\rightarrow$ operować tylko na wzorcach domkniętych

ROZMNAŻANIE WZORCÓW DOMKNIĘTYCH

Lemat 1:

Niech P oraz Q będą wzorcami takimi, że:

$\rightarrow T(P) = T(Q)$

$\rightarrow P \subset Q$

Wówczas dla dowolnego $i \notin P$

$$T(P \cup \{i\}) = T(Q \cup \{i\})$$

Dowód:

$T(P \cup \{i\}) = T(P) \cap T(\{i\}) = T(Q) \cap T(\{i\}) = T(Q \cup \{i\})$

ROZMNAŻANIE WZORCÓW DOMKNIĘTYCH

Lemat 2:

Każdy domknięty wzorzec $P \neq \perp$ jest domkniętym rozszerzeniem innego domkniętego wzorca.

Dowód:

Niech Q będzie wzorcem powstałym przez usuwanie elementów z P do momentu zmiany jego częstości i niech element i będzie ostatnim usuniętym. Wówczas $Clo(Q \cup \{i\}) = P$.

Taki wzorzec musi istnieć ponieważ $P \neq \perp$.

Skoro $T(Q) \neq T(Q \cup \{i\})$, to $i \notin Clo(Q)$.

Wobec tego $Clo(Q \cup \{i\}) = Clo(Clo(Q) \cup \{i\})$.

Czyli P jest domkniętym rozszerzeniem $Q \cup \{i\}$.

ALGORYTM WZORCÓW DOMKNIĘTYCH

1. $D = \{\perp\}$
2. $D' = \{Clo(P \cup \{i\}) \mid P \in D, i \in E \setminus P\}$
3. if $D' = \emptyset$ then output D; stop
4. $D = D \cup D'$; go to 2.

$\rightarrow$ problem: potrzebujemy sporo pamięci na przechowywanie D...

ROZSZERZENIE ZACHOWUJĄCE PREFIKS

- Niech P będzie wzorcem domkniętym.
- **Rdzeniem** wzorca P (ozn. $core\_i(P)$) będziemy nazywać najmniejszą liczbę i taką, że
$$T(P(i)) = T(P)$$
- Niech ponadto $core\_i(\perp) = 0$.

ROZSZERZENIE ZACHOWUJĄCE PREFIKS

- Wzorzec Q nazywamy **rozszerzeniem domkniętym zachowującym prefiks** (*ppc-extension*) wzorca P, gdy:

I. $Q = Clo(P \cup \{i\})$ dla pewnego $i \notin P$

II. $i \notin P$ oraz $i > core\_i(P)$

III. $P(i\text{-}1) = Q(i\text{-}1)$

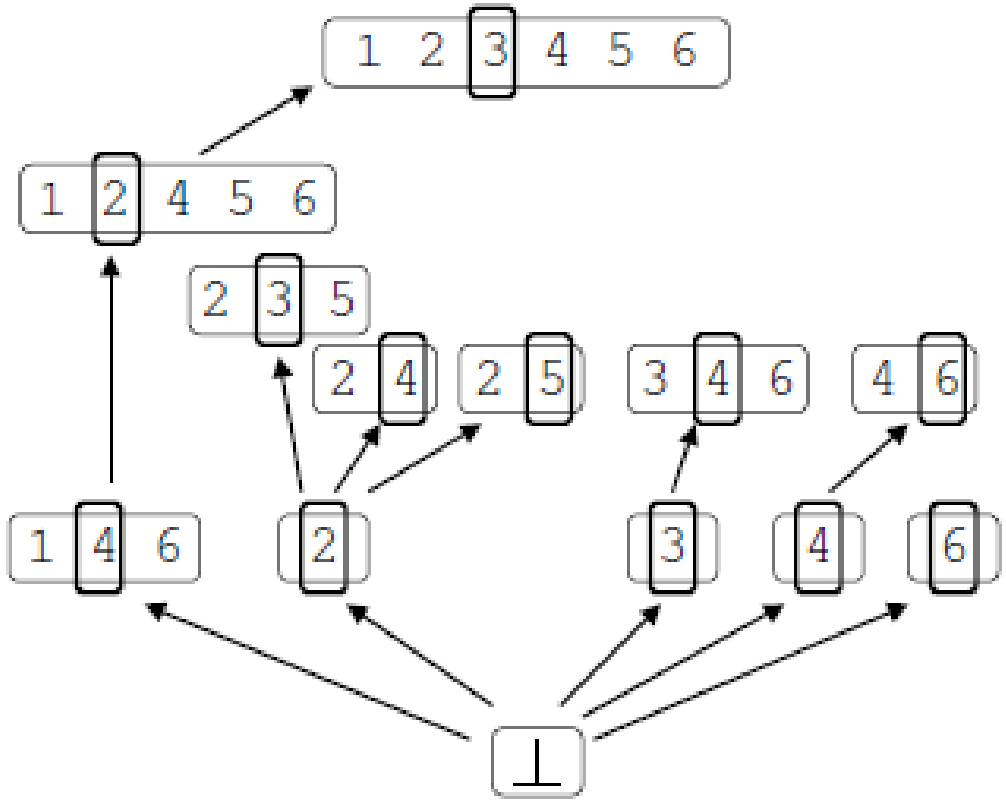
Twierdzenie:

Jeżeli $Q \neq \perp$ jest wzorcem domkniętym, to istnieje tylko jeden wzorzec P taki, że Q jest rozszerzeniem domkniętym zachowującym prefiks.

PRZYKŁAD

transaction database

1	2	3	4	5	6
	2	3		5	
	2			5	
1	2		4	5	6
	2		4		
1			4		6
		3	4		6



LINEAR TIME CLOSED PATTERN MINER

call ENUM_CLOSEDPATTERNS($\perp$);

Procedure ENUM_CLOSEDPATTERNS(P)

if P nie jest częsty then RETURN;

output P;

for $i = core\_i(P) + 1$ to $|E|$

$Q = Clo(P \cup \{i\})$;

if $P(i\text{-}1) = Q(i\text{-}1)$ then

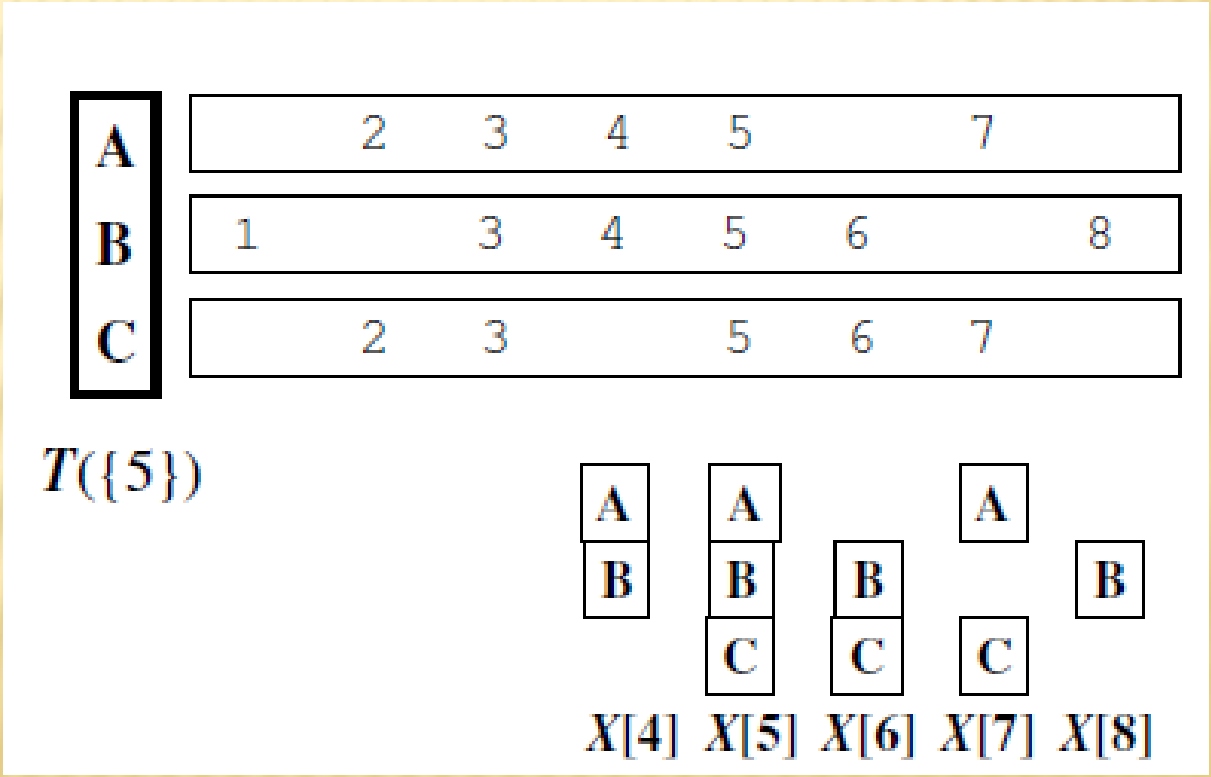
call ENUM_CLOSEDPATTERNS(Q);

end for

ULEPSZENIA

Occurence deliver

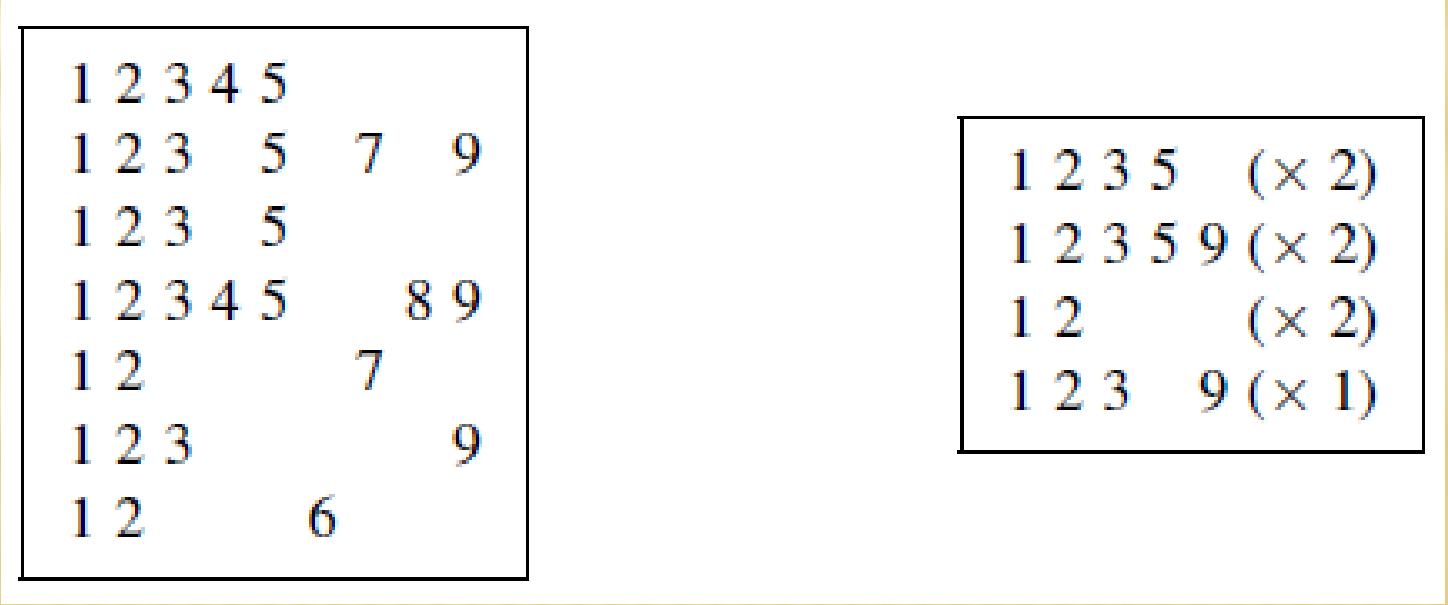
$\rightarrow$ redukuje czas konstrukcji $T(P \cup \{i\})$



ULEPSZENIA

(Anytime) Database Reduction

- wyrzucamy artykuły, których wsparcie jest mniejsze od minimalnego lub występują wszędzie, a następnie scalamy takie same transakcje.



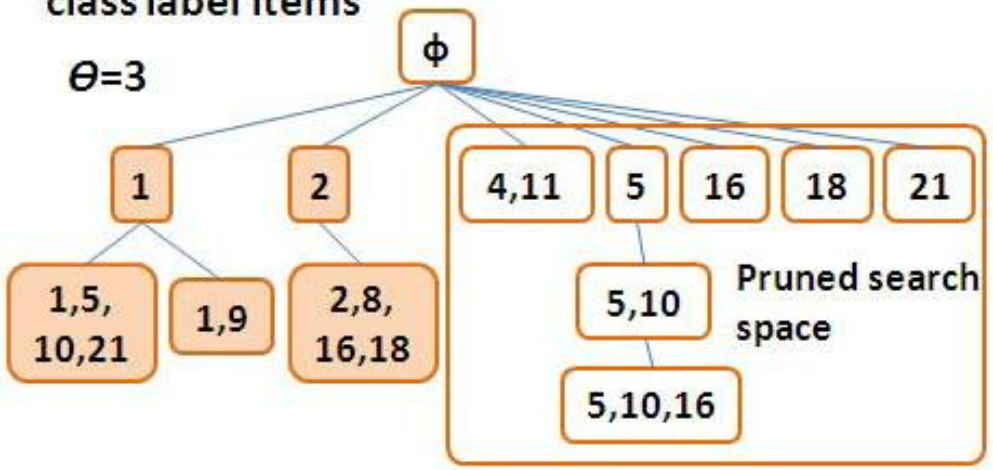
ULEPSZENIA

→ Class label items

Transaction	Class label Item	Item				
A	1	5	8	10	16	21
B	1	5	9	10	14	21
C	1	5	9	10	16	21
D	1	4	9	11	14	18
E	2	5	8	10	16	18
F	2	4	6	11	17	21
G	2	4	8	11	16	18
H	2	5	8	12	16	18

(a) Transaction DB

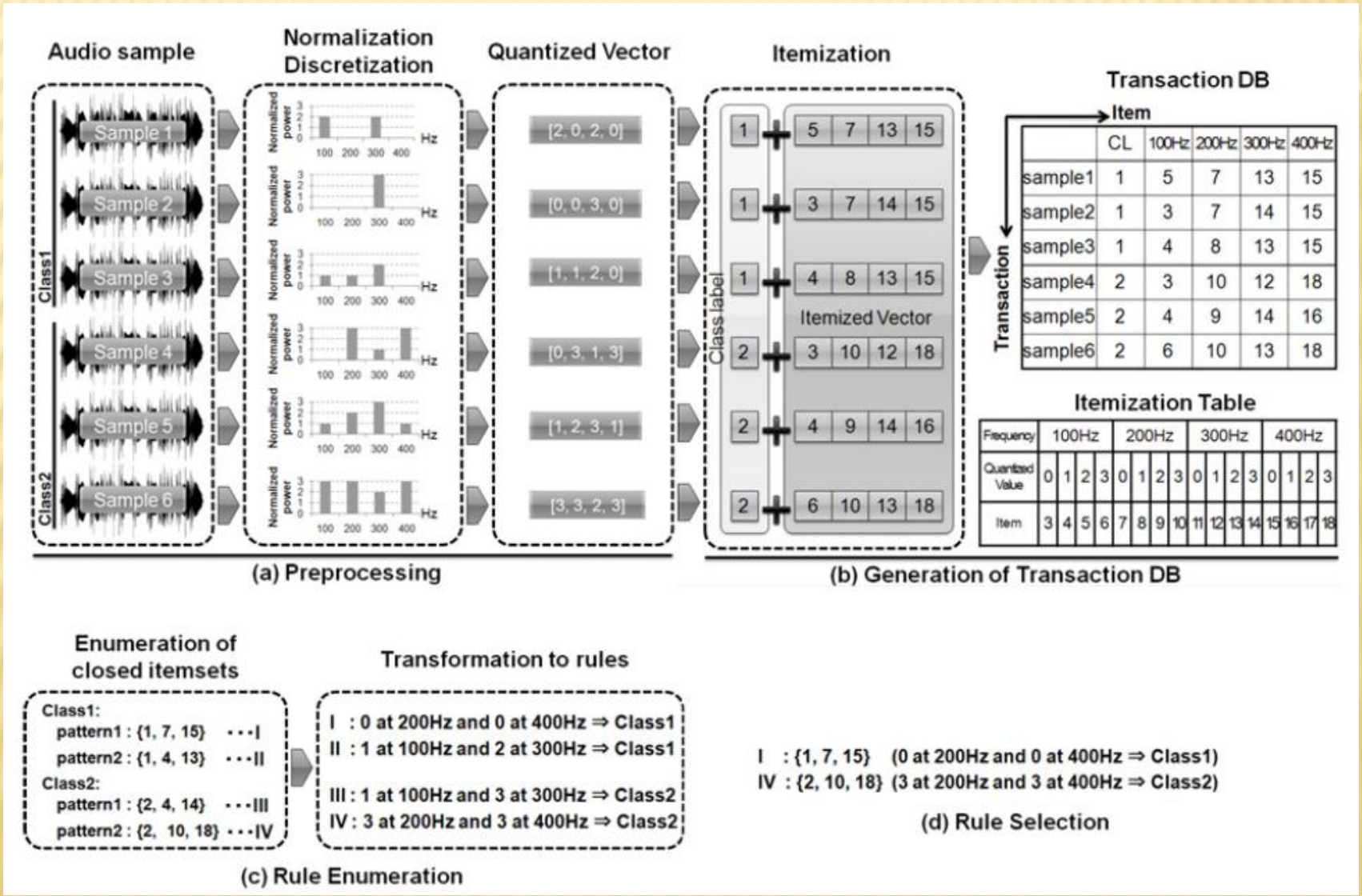
(b) *ppc* extension and the pruning operation by class label items



(c) Obtained rules

{5,10,21} → Class 1, {1,9} → Class 1,
{8,16,18} → Class 2

PRZYKŁAD ZASTOSOWANIA



ŹRÓDŁA:

1. **„Audio classification based on a closed itemset mining algorithm”,** Yoshifumi Okada, Takahiro Tada, Kentarou Fukuta, Tomomasa Nagishima, *2010 International Conference on Computer Information Systems and Industrial Management Applications*
2. **„An efficient algorithm for enumerating closed patterns in transaction databases”**, Takeaki Uno, Tatsuya Asai, Yuzo Uchida, Hiroki Arimura, *Lecture Notes in Artificial Intelligence, 2004.*