# Systemy wyszukiwania informacji

Wprowadzenie i pojęcia wstępne.

# Definicja systemu informacyjnego

| X\A | A | B | C |
|---|---|---|---|
| $x_1$ | $a_1$ | $b_1$ | $c_1$ |
| $x_2$ | $a_1$ | $b_1$ | $c_2$ |
| $x_3$ | $a_1$ | $b_2$ | $c_3$ |
| $x_4$ | $a_2$ | $b_1$ | $c_4$ |
| $x_5$ | $a_1$ | $b_2$ | $c_1$ |
| $x_6$ | $a_1$ | $b_2$ | $c_2$ |
| $x_7$ | $a_1$ | $b_1$ | $c_1$ |

$S = \langle X, A, V, \rho \rangle$

$X = \{x_1, \ldots, x_7\}$

$A = \{a, b, c\}$

$V_a = \{a_1, a_2\}$

$V_b = \{b_1, b_2\}$

$V_c = \{c_1, c_2, c_3, c_4\}$

$\rho = X \times A \rightarrow V$

$\rho(x, a) = v$ np. $\rho(x_6, b) = b_2$

# Informacja o obiekcie

Funkcja informacji przyporządkowuje każdemu obiektowi i atrybutowi odpowiednią wartość; przedstawiona może być za pomocą tabeli:

| **X\A** | **Wykonawca** | **Gatunek** | **Czas trwania** |
|---|---|---|---|
| $\mathbf{X_1}$ | Lady Pank | Rock | 4:01 |
| $\mathbf{X_2}$ | U2 | Rock | 3:10 |
| $\mathbf{X_3}$ | Verba | Hip Hop | 3:30 |
| $\mathbf{X_4}$ | Bajm | Pop | 4:06 |

**Informacją o obiekcie** x jest funkcja $\rho_x : A \rightarrow V$ taka, że $\rho_x(a) = \rho(x, a)$ dla każdego $a \in A$, $x \in X$.

Jest to zbiór wartości wszystkich atrybutów obiektu w danym systemie.

**Przykład**: $\rho_{x6} = (a, a_1)(b, b_2)(c, c_2)$

# Informacja w systemie

- Informacją w systemie S będzie każda funkcja $\varphi$ o argumentach w zbiorze atrybutów A oraz wartościach należących do V, taka, że $\varphi(a) \in \varphi$ Va.

- Informacja $\varphi$ jest pusta gdy nie odpowiada jej żaden obiekt w systemie: $X_\varphi = \{\emptyset\}$

- Wszystkich możliwych do uzyskania informacji w systemie będzie:

$$\bigcap_{a \in A} card(V_a)$$

# Język deskryptorowy

**Deskryptor** – para atrybut–wartość.

$(a_i, v_{ij})$, gdzie $a_i \in A$; $v_{ij} \in Va_i$;

$Va_i$ – zbiór wartości atrybutu a.

**Dziedzina atrybutu „$a_i$" jest co najmniej dwuelementowa** – każdy atrybut może przyjmować co najmniej jedną z 2 możliwych wartości.

**Opis obiektu** x w systemie S to zbiór deskryptorów wyznaczony przez informację o obiekcie.

Język deskryptorowy (LS) to język systemowy, język opisu w kartotece wyszukiwawczej oraz język pytań i odpowiedzi w systemie. Jest on zdefiniowany jako para:

$LS = \langle A, G \rangle$, gdzie A – alfabet,

G – dwustopniowa gramatyka

# Alfabet

Alfabet – określa zbiór dopuszczalnych symboli jakie występują w języku:

1. 0, 1 – stałe (oznaczenie zbioru pustego i pełnego)
2. A – zbiór nazw atrybutów
3. V – zbiór nazw wartości atrybutów
4. +, ·, ~ – symbole operacji logicznych „lub”, „i”, „nie”
5. ( , ) – nawiasy

# Gramatyka

- Zdefiniowana jest dwustopniowo:
  **Syntaktyka** – określa zasady tworzenia słów w danym języku. Słowa w języku deskryptorowym nazywamy **termami** (T – zbiór termów; zbiór słów w języku deskryptorowym):

1. Stałe 0, 1 $\in$ T (są termami).
2. Deskryptor (a, v) $\in$ T, gdzie a $\in$ A, V $\in$ $V_a$ jest termem.
3. Jeżeli dwa termy t, t' należą do języka to słowami języka są również:
   - t + t' $\in$ T
   - t · t' $\in$ T
   - ~t $\in$ T

**Przykłady termów**: (Gatunek, Rock) $\in$ T, ~(Wykonawca, U2) $\in$ T, (Gatunek, Pop) · (Wykonawca, Bajm) $\in$ T, (Gatunek, Rock) · (Wykonawca, U2) + (Wykonawca, Lady Pank) $\in$ T

# Gramatyka c.d.

**Semantyka** – określa znaczenie słów.

Znaczenie słów (termów) określa funkcja σ odwzorowująca zbiór termów T systemu S w zbiór obiektów X tego systemu:

$$\sigma : T \rightarrow P(X),$$

gdzie $P(X)$ oznacza rodzinę podzbiorów zbioru X.

Funkcja σ określona jest w następujący sposób:

1. $\sigma(0) = \emptyset$ $\quad$ $\sigma(1) = X$ (pełny zbiór obiektów)
2. $\sigma(a, v) = \{x \in X, \rho_x(a) = v\}$
3. $\sigma(\sim t) = \{X|_{\sigma(t)}\}$
4. $\sigma(t + t') = \sigma(t) \cup \sigma(t')$
5. $\sigma(t \cdot t') = \sigma(t) \cap \sigma(t')$

# Pytania kierowane do systemu

Gdy pytanie kierowane do systemu jest termem, to znalezienie odpowiedzi na to pytanie polega na nadaniu znaczenia poszczególnym jego termom.

Przykład:
$t = (c, c_3)$
$\sigma(t) = \{x_3\}$

$$t = \underbrace{(c, c_3)}_{t_1} + \underbrace{(b, b_1)}_{t_2}$$

- $\sigma(t_1) = \{x_3\}$
- $\sigma(t_2) = \{x_1, x_2, x_4, x_7\}$
- $\sigma(t) = \sigma(t_1) \cup \sigma(t_2) = \{x_3\} \cup \{x_1, x_2, x_4, x_7\}$
  $= \{x_1, x_2, x_3, x_4, x_7\}$

# Typy termów

- Term **elementarny**.

Term t nazywamy elementarnym jeśli jest postaci:

$$t_e = (a_1, v_{i1}) \cdot (a_2, v_{i2}) \cdot \ldots \cdot (a_n, v_{in})$$

gdzie: $a_i \in A$, $v_{ij} \in V_{aj}$. Jeżeli przez $d_j^i$ oznaczymy deskryptor $(a_j, v_{ij})$ to:

$$t_e = d_1^i \cdot d_2^i \cdot \ \ldots \cdot \ d_n^i$$

- Term **składowy**.

Term t nazywamy składowym jeśli jest postaci:

$$t_s = d_1^i \cdot d_2^i \cdot \ \ldots \cdot \ d_k^i$$

gdzie $k \leq n$.

- Term **normalny**.

Term t nazywamy normalnym jeśli jest on sumą termów elementarnych:

$t_n = t_1 + t_2 + \ldots + t_m$, gdzie $t_i$ to termy elementarne

# Normalizacja

Proces prowadzący od termu składowego do termu elementarnego to normalizacja.

W wyniku normalizacji oczywiście powstaje term normalny.

**Przykład**

Załóżmy, że szukamy odpowiedzi na pytanie wyrażone termem składowym $t = (a,a_1)\cdot(b,b_1)$. Wartość atrybutu c jest wówczas nieistotna. Zatem należy znaleźć odpowiedź na pytanie:

$$t = (a,a_1)\cdot(b,b_1)\cdot[(c,c_1)+(c,c_2)+(c,c_3)+(c,c_4)] =$$
$$(a,a_1)(b,b_1)(c,c_1) + (a,a_1)(b,b_1)(c,c_2) +$$
$$(a,a_1)(b,b_1)(c,c_3) + (a,a_1)(b,b_1)(c,c_4)$$

# Własności termów elementarnych

1. Wszystkie termy elementarne są rozłączne.

$$\bigwedge_{t,t' \in T_e} \sigma(t) \cap \sigma(t') = \emptyset$$

2. Suma odpowiedzi na termy elementarne daje cały zbiór obiektów.

$$\bigcup_{t \in T_e} \sigma(t) = X$$

# Równość i zawieranie się termów

- Zawieranie się termów

**Term t jest zawarty w t'** wtedy i tylko wtedy, gdy zbiór obiektów odpowiadający wartości termu t' jest zawarty w zbiorze obiektów odpowiadających wartości termu t:

$$t \leq t' \leftrightarrow \sigma(t') \subseteq \sigma(\mathrm{t})$$

- Równość termów

**Termy t i t' są równe** w systemie S wtedy i tylko wtedy, gdy wartości tych termów są równe:

$$t = t' \leftrightarrow \sigma(t) = \sigma(t')$$

# Zawieranie się termów – przykład

$t = (a, a_1)$

$t' = (a, a_1)(b, b_1)$

$\sigma(t) = \{x_1, x_2, x_3, x_5, x_6, x_7\}$

$\sigma(t') = \{x_1, x_2, x_7\}$

**ODP**: **Term t zawiera się w t', gdyż:**

$$\sigma(t') \subseteq \sigma(t) \leftrightarrow t \leq t'$$

| X\A | a | b | c |
|---|---|---|---|
| $x_1$ | $a_1$ | $b_1$ | $c_1$ |
| $x_2$ | $a_1$ | $b_1$ | $c_2$ |
| $x_3$ | $a_1$ | $b_2$ | $c_3$ |
| $x_4$ | $a_2$ | $b_1$ | $c_4$ |
| $x_5$ | $a_1$ | $b_2$ | $c_1$ |
| $x_6$ | $a_1$ | $b_2$ | $c_2$ |
| $x_7$ | $a_1$ | $b_1$ | $c_1$ |

# Właściwości systemu

- System jest **kompletny** wtedy i tylko wtedy, gdy każdej informacji odpowiada co najmniej jeden obiekt (tj. gdy każda informacja jest niepusta).

- Nasz przykładowy system nie jest kompletny bo istnieje informacja której nie odpowiada żaden obiekt w systemie – tj. $\sigma(a, a_1)(b, b_1)(c, c_4) = \{\emptyset\}$.

- System jest **selektywny** wtedy i tylko wtedy, gdy każdej informacji odpowiada co najwyżej jeden obiekt.

- Nasz przykładowy system nie jest selektywny bo istnieje informacja $(a, a_1)(b, b_1)(c, c_1)$ której odpowiadają dwa obiekty $x_1$ oraz $x_7$ – tj. $\sigma(a, a_1)(b, b_1)(c, c_1) = \{x_1, x_7\}$

# Redundancja

- **Liczba informacji** (L) w systemie wyznaczania jest ze wzoru:

  $L = m_1 \cdot m_2 \ldots m_n$,
  gdzie $m_i = \text{card } V_{ai}$
  (liczebność konkretnego zbioru wartości)

  **Przykład**:

  $m_1 = 2$, $m_2 = 2$, $m_3 = 4$ stąd $L = 2 \cdot 2 \cdot 4 = 16$.

- **Redundancja** – nadmiar informacji w stosunku do przyjętego minimum.

  **Wady**:
  - zwiększa zajętość pamięci
  - utrudnia aktualizację

# Zależność między atrybutami

Powiemy, że atrybut **b jest zależy od atrybutu a** wtedy i tylko wtedy, gdy klasa równoważności (nierozróżnialności) dla atrybutu a **zawiera się** w klasie równoważności dla atrybutu b.

$$a \rightarrow b \;\leftrightarrow\; \tilde{a} \subset \tilde{b}$$

Klasa równoważności – zbiór obiektów nierozróżnialnych ze względu na wartość danego atrybutu.

# Zależność między atrybutami c.d.

- Atrybuty a i b są niezależne gdy nie zachodzi żadna z relacji:

$$\tilde{a} \subset \tilde{b} \text{ ani } \tilde{b} \subset \tilde{a}$$

- Atrybuty a i b są równoważne w systemie S gdy:

$$\tilde{b} = \tilde{a}$$

# Zależność między atrybutami – przykład

Klasy równoważności:

$$\tilde{a} = \{\{x_1, x_2, x_3, x_5, x_6, x_7\}, \{x_4\}\}$$
$$\tilde{b} = \{\{x_1, x_2, x_4, x_7\}, \{x_3, x_5, x_6\}\}$$
$$\tilde{c} = \{\{x_1, x_5, x_7\}, \{x_2, x_6\}, \{x_3\}, \{x_4\}\}$$

Wynika z tego że:

$$c \rightarrow a \leftrightarrow \tilde{c} \subset \tilde{a}$$

Atrybut a jest atrybutem zależnym od atrybutu c, zatem jest on zbędny bo:

$$c_1 \rightarrow a_1$$
$$c_2 \rightarrow a_1$$
$$c_3 \rightarrow a_1$$
$$c_4 \rightarrow a_2$$

| X\A | a | b | c |
|---|---|---|---|
| $x_1$ | $a_1$ | $b_1$ | $c_1$ |
| $x_2$ | $a_1$ | $b_1$ | $c_2$ |
| $x_3$ | $a_1$ | $b_2$ | $c_3$ |
| $x_4$ | $a_2$ | $b_1$ | $c_4$ |
| $x_5$ | $a_1$ | $b_2$ | $c_1$ |
| $x_6$ | $a_1$ | $b_2$ | $c_2$ |
| $x_7$ | $a_1$ | $b_1$ | $c_1$ |

# Nierozróżnialność obiektów w systemie

Powiemy, że obiekty x, y $\in$ X są **nierozróżnialne w systemie S** **ze względu na atrybut** **a** $\in$ **A** wtedy i tylko wtedy, gdy

$$\rho_x(a) = \rho_y(a).$$

Nierozróżnialność obiektów x, y ze względu na atrybut a zapisujemy jako $x \tilde{}_a y$.

**Przykład**: Obiekty $\mathbf{x_1}$ i $\mathbf{x_2}$ są nierozróżnialne w systemie ze względu na atrybut **b**.

Obiekty x, y $\in$ X nazywamy **nierozróżnialnymi w systemie S** ($x \tilde{}_S y$) wtedy i tylko wtedy, gdy

$$\rho_x = \rho_y.$$

**Przykład**: Obiekty $\mathbf{x_1}$ i $\mathbf{x_7}$ są nierozróżnialne w systemie.

# Czy to wszystko?

- Nie
- Proszę zapoznać się również dostępnymi materiałami wykładowymi!

# Ćwiczenia

Na podstawie kartoteki wtórnej i opisu systemu sporządzonego w domu:

1. Zapisz po 3 termy elementarne oraz składowe.
2. Dokonaj normalizacji wybranego termu składowego.
3. Sprawdź czy (i uzasadnij dlaczego) system jest selektywny i kompletny.
4. Policz liczbę informacji w systemie.
5. Zbadaj zależności między atrybutami w systemie.