Systemy wyszukiwania informacji

Wprowadzenie i pojęcia wstępne.

Definicja systemu informacyjnego

Systemem informacyjnym nazywamy czwórkę

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

$X = \{x_1, \ldots, x_n\}$ – skończony zbiór obiektów (podmiotów) systemu: np. książki w bibliotece, filmy w wypożyczalni, studenci w dziekanacie

$A = \{a_1, \ldots, a_n\}$ – skończony zbiór atrybutów (cech, pojęć) opisujących obiekty w systemie, np. dla książek: wydawnictwo, rok wydania, autor, język;

$V = \bigcup_{a \in A} V_a$, gdzie V_a to zbiór wartości atrybutu a.

$\rho : X \times A \rightarrow V$ – funkcja informacji, która odwzorowuje produkt iloczynu kartezjańskiego zbioru obiektów i zbioru atrybutów w zbiór wartości atrybutów.

Funkcja informacji

Powiemy, że $\rho(x, a) = v \in V_a$ dla każdego $x \in X$, oraz $a \in A$ (funkcja informacji przyporządkowuje każdemu obiektowi i atrybutowi odpowiednią wartość).

Funkcja informacji ρ jest funkcją **całkowitą**, co oznacza, że musi być określona dla wszystkich wartości argumentów x oraz a.

Intuicyjnie rzecz biorąc, musi być zawsze znana wartość każdego atrybutu.

Funkcja informacji cz. 2

Funkcja informacji przedstawiona może być za pomocą tabeli:

X\A	**A**	**B**	**C**
$\mathbf{X_1}$	a_1	b_1	c_1
$\mathbf{X_2}$	a_1	b_2	c_2
$\mathbf{X_3}$	a_2	b_2	c_3
$\mathbf{X_4}$	a_1	b_1	c_2
$\mathbf{X_5}$	a_2	b_2	c_3

Przykład kartoteki wtórnej

Obiekt	Producent	Oznaczenie kodowe	Przekątna ekranu	Technologia matrycy	Czas reakcji	Kontrast	Jasność	Rozdzielczość natywna
x1	Fujitstu-Siemens	P19-2	średnia	S-PVA	krótki	bardzo duży	średnia	1280x1024
x2	Samsung	SM193P+	średnia	S-PVA	krótki	bardzo duży	średnio-niska	1280x1024
x3	Samsung	SM971P	średnia	S-PVA	krótki	bardzo duży	średnio-niska	1280x1024
x4	NEC	90GX2	średnia	TN	krótki	mały	wysoka	1280x1024
x5	NEC	2070NX	duża	S-IPS	długi	mały	niska	1600x1200
x6	Fujitsu-Siemens	S20-1	duża	MVA Premium	krótki	duży	średnio-wysoka	1600x1200
x7	Eizo	L578	mała	PVA	średni	bardzo duży	niska	1280x1024
x8	Eizo	L887	duża	S-IPS	bardzo długi	bardzo mały	niska	1600x1200
x9	Fujitsu-Siemens	P17-2	mała	S-PVA	krótki	bardzo duży	średnia	1280x1024
x10	Eizo	L568	mała	PVA	bardzo długi	bardzo duży	niska	1280x1024
x11	Belinea	101710	mała	TN	krótki	mały	średnio-wysoka	1280x1024
x12	Samsung	SM913TM	średnia	MVA Premium	krótki	bardzo duży	niska	1280x1024
x13	Samsung	SM204TS	duża	S-PVA	długi	mały	średnio-wysoka	1600x1200
x14	NEC	1980SXi	średnia	S-IPS	bardzo długi	bardzo mały	średnio-niska	1280x1024
x15	Eizo	M1700	mała	TN	krótki	mały	średnio-wysoka	1280x1024

$X = \{x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15\}$

$A = \{$Producent, Oznaczenie kodowe, Przekątna ekranu, Technologia matrycy, Czas reakcji, Kontrast, Jasność, Rozdzielczość natywna$\}$

$V_{\text{Producent}} = \{$Belinea, Eizo, Fujitstu-Siemens, NEC, Samsung$\}$

$V_{\text{Oznaczenie kodowe}} = \{$P19-2, SM193P+, SM971P, 90GX2, 2070NX, S20-1, L578, L887, P17-2, L568, 101710, SM913TM, SM204TS, 1980SXi, M1700$\}$

$V_{\text{Przekątna ekranu}} = \{$mała, średnia, duża$\}$

$V_{\text{Technologia matrycy}} = \{$MVA Premium, PVA, S-IPS, S-PVA, TN$\}$

$V_{\text{Czas reakcji}} = \{$krótki, średni, długi, bardzo długi$\}$

$V_{\text{Kontrast}} = \{$bardzo mały, mały, duży, bardzo duży$\}$

$V_{\text{Jasność}} = \{$niska, średnio-niska, średnia, średnio-wysoka, wysoka$\}$

$V_{\text{Rozdzielczość natywna}} = \{1280 \times 1024, 1600 \times 1200\}$

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

Ważne pojęcia wprowadzające

- **Baza danych** – zbiór danych zapisanych w ściśle określony sposób w strukturach odpowiadających założonemu modelowi danych (hierarchiczny, relacyjny, obiektowy)
- **Aktualizacja bazy danych** – czynność zapewniająca w każdej chwili właściwy zbiór danych, w bazie danych polega na dodawaniu, usuwaniu dokumentów z bazy danych oraz dokonywaniu zmian w opisach dokumentów.
- **Czas wyszukiwania** – czas upływający od momentu skierowania pytania do systemu, do momentu, gdy system znajdzie odpowiedź w swojej strukturze wewnętrznej.
- **Deskryptor** – para atrybut-wartość [lub jednostka składniowa używana jako podstawowy element (słowo kluczowe) języka informacyjno-wyszukiwawczego w systemie informatycznej analizy treści dokumentu lub automatycznego wyszukiwania dokumentów]:

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

Przykład: (Rok wydania, 2010); (Gatunek, Komedia);
(Moc silnika, 200);

Ważne pojęcia wprowadzające cz. 2

- **Dokument relewantny** – dokument *a* jest relewantny względem pytania *q*, jeżeli w opisie dokumentu *a* występują wszystkie niezaprzeczone deskryptory pytania *q* i w opisie tym nie występuje żaden z deskryptorów zaprzeczonych pytania *q* (o ile *q* zawiera deskryptory zaprzeczone).
- **Dokument wtórny** – dokument opracowany na podstawie dokumentu źródłowego przystosowany do konkretnego systemu informatycznego; dokument gdzie wszystkie informacje z dokumentu źródłowego są kodowane; są to informacje skrócone. Zawiera wyłącznie niezbędne informacje.
- **Kartoteka wtórna** – zbiór dokumentów wtórnych.
- **Dokument wyszukiwawczy** – jest to dokument opracowany na podstawie dokumentu wtórnego; przystosowany do konkretnej metody wyszukiwania informacji.
- **Kartoteka wyszukiwawcza** – zbiór dokumentów w formacie wyszukiwawczym.
- **Dokument źródłowy** – dokument w języku naturalnym lub innym źródłowym powstały w miejscu źródła informacji.
- **Kartoteka źródłowa** – zbiór dokumentów źródłowych.

Ważne pojęcia wprowadzające cz. 3

- **Dyskretyzacja atrybutów** – atrybuty ilościowe (jak np. wiek, wzrost, waga) ulegają rozgraniczeniu na przedziały.
- **Podział dychotomiczny** – podział danej całości (zbioru) na dwie części (grupy) które są wobec siebie przeciwstawne lub wzajemnie się wykluczają. Przykładem takiego podziału jest podział zbioru liczb całkowitych na liczby parzyste i nieparzyste.
- **Cecha dychotomiczna** – atrybut mający dokładnie dwie wartości które się wzajemnie wykluczają. Atrybutem dychotomicznym jest np. **płeć**.

Ćwiczenia

Proszę stworzyć własną kartotekę wtórną. Tematyka dowolna.

System powinien posiadać:

- Od 10 do 20 obiektów.
- Każdy obiekt powinien być opisany od 5 do 7 atrybutami.
- Proszę dokonać dyskretyzacji wszystkich atrybutów numerycznych poza jednym.
- Proszę uwzględnić 1 atrybut dychotomiczny.