1. KARTOTEKA WTÓRNA

Obiekt	Producent	Model	Poj. Baterii	Wodoodporność	Przek. Ekranu	L.rdzeni proc	Sys. Operacyjny	RAM
x1	Apple	Iphone Xs	mała	Tak	średnia	6	IOS	4
x2	Samsung	Galaxy S10	średnia	Tak	średnia	8	Android	6
x3	Sony	Xperia 1	duża	Tak	duża	8	Android	6
x4	Huawei	P30 PRO	b. duża	Tak	duża	8	Android	6
x5	Xiaomi	MiMix 3	duża	Nie	duża	8	Android	6
x6	Apple	Iphone 8Plus	średnia	Tak	średnia	6	IOS	3
x7	Samsung	Galaxy Note 10	b. duża	Tak	duża	8	Android	8
x8	Sony	Xperia 10 plus	średnia	Tak	duża	8	Android	4
x9	Huawei	P smart Z	b. duża	Nie	duża	8	Android	4
x10	Apple	Iphone 6s	b. mała	Nie	mała	2	IOS	2
x11	Samsung	Galaxy s6	mała	Nie	mała	8	Android	3
x12	Sony	Xperia Z3	średnia	Tak	mała	4	Android	3
x13	Nokia	9 PureView	duża	Tak	średnia	8	Android	6

S = {X ,A, V, P}

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

A = {Producent, Model, Poj. Baterii, Wodoodporność, Przek. Ekranu, L. Rdzeni proc., Sys. Operacyjny, RAM}

$\mathbf{V}_{\text{Producent}}$ = {Apple, Samsung, Sony, Huawei, Xiaomi, Nokia}

$\mathbf{V}_{\text{Model}}$ = { Iphone Xs, Galaxy S10, Xperia 1, P30 PRO, MiMix 3, Iphone 8Plus, Galaxy Note 10, Xperia 10 plus, P smart Z, Iphone 6s, Galaxy s6, Xperia Z3, 9 PureView}

$\mathbf{V}_{\text{Poj. Baterii}}$ = {b. mała < 2000, 2000 ≤ mała < 2800, 2800 ≤ średnia < 3100, 3100 ≤ duża < 3500, 3500 ≤ b. duża < 4200}

$\mathbf{V}_{\text{Wodoodporność}}$ = {TAK , NIE}

$\mathbf{V}_{\text{Przek. Ekranu}}$ = {mała ≤ 5,5; 5,5 < średnia ≤ 6,2; 6,2 < duża}

$\mathbf{V}_{\text{L. rdzeni proc}}$ = {2,4,6,8}

$\mathbf{V}_{\text{Sys. Operacyjny}}$ = {Android, IOS}

$\mathbf{V}_{\text{RAM}}$ = {2,3,4,6,8}

2. USTALENIE PARAMETRÓW

System wyszukiwania informacji zadany jest poprzez określenie zbioru obiektów *X*, zbioru atrybutów, zbioru wartości atrybutów *V*, oraz funkcji informacji. Opisy obiektów pogrupowane są w **bazie danych** w grupy X_i *(i = 1, …, m)*. Każda grupa X_i poprzedzona jest **identyfikatorem grupy**, który nazwany jest **centroidem** (C_i) lub **profilem** (P_i). Po pogrupowaniu kartotekę wyszukiwawczą tworzą grupy poprzedzone reprezentantami (centroid, profil). Ostatecznie struktura kartoteki wyszukiwawczej jest strukturą hierarchiczną **PIEŃ – CENTROID – DOKUMENT**.

3. ALGORYTM ROCCHIA

Potencjalne centrum grupy $x_c - x_2$

(Producent, Samsung)*(Model, Galaxy S10)*(Poj. Baterii, średnia)*(Wodoodporność,Tak) *(Przek. Ekranu, średnia)*(L. Rdzenii proc, 8)*(Sys. Operacyjny, Android)*(RAM, 6).

Zakładamy parametry metody N1, N2, p1, p2

N1 =5
N2 = 3
p1 = 0,2
p2 = 0,3

Określamy prawdopodobieństwo wszystkich dokumentów do centrum grupy

$$p(x_c, x_i) = \frac{\overline{\overline{x_c \cap x_i}}}{\overline{\overline{x_c \cup x_i}}}$$

$p(x_c,x_1) = \frac{2}{14} = 0{,}14$

$p(x_c,x_2) = \frac{1}{1} = 1$

$p(x_c,x_3) = \frac{4}{12} = 0{,}33$

$p(x_c,x_4) = \frac{4}{12} = 0{,}33$

$p(x_c,x_5) = \frac{3}{13} = 0{,}23$

$p(x_c,x_6) = \frac{3}{13} = 0{,}23$

$p(x_c,x_7) = \frac{4}{12} = 0{,}33$

$p(x_c,x_8) = \frac{4}{12} = 0{,}33$

$p(x_c,x_9) = \frac{2}{14} = 0{,}14$

$p(x_c,x_{10}) = \frac{0}{16} = 0$

$p(x_c,x_{11}) = \frac{3}{13} = 0{,}23$

$p(x_c,x_{12}) = \frac{3}{13} = 0{,}23$

$p(x_c,x_{13}) = \frac{5}{11} = 0{,}45$

Nasze założenia zostały spełnione (test gęstosci zaliczony), ponieważ 10 obiektów ma współczynnik korelacji większy od 0,2 oraz jednocześnie 6 obiektów ma współczynnik większy od 0,3.

Sortujemy obiekty względem korelacji + nadajemy rangę R

$p(x_c,x_2) = \frac{1}{1} = 1$	R1		$p(x_c,x_6) = \frac{3}{13} = 0{,}23$	R8	
$p(x_c,x_{13}) = \frac{5}{11} = 0{,}45$	R2		$p(x_c,x_{11}) = \frac{3}{13} = 0{,}23$	R9	
$p(x_c,x_3) = \frac{4}{12} = 0{,}33$	R3		$p(x_c,x_{12}) = \frac{3}{13} = 0{,}23$	R10	
$p(x_c,x_4) = \frac{4}{12} = 0{,}33$	R4		$p(x_c,x_1) = \frac{2}{14} = 0{,}14$	R11	
$p(x_c,x_7) = \frac{4}{12} = 0{,}33$	R5		$p(x_c,x_9) = \frac{2}{14} = 0{,}14$	R12	
$p(x_c,x_8) = \frac{4}{12} = 0{,}33$	R6		$p(x_c,x_{10}) = \frac{0}{16} = 0$	R13	
$p(x_c,x_5) = \frac{3}{13} = 0{,}23$	R7				

Wyznaczamy M1, M2 , gdzie
M1 – zbiór elementów $\geqslant p_2$ i M2 – zbiór elementów $\geqslant p_1$

M1 = 6
M2 = 10

M1 ≠ M2

więc
obliczamy różnicę pomiędzy współczynnikami korelacji obiektów sąsiednich w grupie maksymalnej M2 bez grupy minimalnej M1, wybieramy obiekty o rangach:

R3, R4, R5, R6, R7, R8, R9, R10

Wyznaczamy największą różnicę współczynnika korelacji

R6 – R7 = 0,33 - 0,23 = 0,1

P_{min} =0,33

Tworzymy grupę wstępną, do której należą elementy o współczynniku $\geqslant P_{min}$

G_w – grupa wstępna

$G_w = \{x_2, x_3, x_4, x_7, x_8, x_{13}\}$

Tworzymy wektor centroidalny, który stanowi sumę opisów obiektów należących do grupy minimalnej

C_w – wektor centroidalny

C_w – {Samsung, Galaxy S10, średnia, Tak, średnia, 8, Android, 6, Sony, Xperia 1, duża, duża, Huawei, P30 Pro, B. duża, Galaxy Note 10, Xperia 10 Plus, 4, Nokia, 9 PureView}

- *średnia*, *duża* się powtarza, lecz odnosi się do innych obiektów.

Wektor centroidalny opisuje zawartość grupy wstępnej.

Zmieniają się parametry

N1 = 5	→	N1c = 5
N2 = 3	→	N2c = 3
p1 = 0,2	→	p1c = 0,25
p2 = 0,3	→	p2c = 0,35

Przeprowadzamy test gęstości dla centoridu + nadajemy nową rangę **R**

$p(x_c,x_2) = \frac{8}{20} = 0,4$		R1
$p(x_c,x_3) = \frac{8}{20} = 0,4$		R2
$p(x_c,x_4) = \frac{8}{20} = 0,4$		R3
$p(x_c,x_7) = \frac{8}{20} = 0,4$		R4
$p(x_c,x_8) = \frac{8}{20} = 0,4$		R5
$p(x_c,x_{13}) = \frac{8}{20} = 0,4$		R6

Test gęstości zaliczony, wyznaczamy M1 oraz M2

$$M1 = 6$$
$$M2 = 6$$
$$M1 == M2$$

G_p – grupa poprawiona
$G_p = \{x_2, x_3, x_4, x_7, x_8, x_{13}\}$

C_p – centroid dla grupy poprawionej

C_p – {Samsung, Galaxy S10, średnia, Tak, średnia, 8, Android, 6, Sony, Xperia 1, duża, duża, Huawei, P30 Pro, B. duża, Galaxy Note 10, Xperia 10 Plus, 4, Nokia, 9 PureView}

Pozostałe obiekty to obiekty swobodne

L – grupa obiektów spowodnych
$L = \{ x_1, x_5, x_6, x_9, x_{10}, x_{11,} x_{12}\}$

4. CZAS WYSZUKIWANIA

Czas wyszukiwania odpowiedzi na pytanie składowe ti w metodzie Saltona (strukturalnej) jest sumą czasów obliczania współczynników podobieństwa z centroidami i czasu przeglądu opisów obiektów wybranej grupy.

$$\tau = \tau w + \tau p$$

Dla pytania składowego ogólnego (pojedynczy deskryptor) na ogół jest bardzo duży czas τ p czyli czas przeglądu obiektów. Dla pytania szczegółowego czas przeglądu jest znacznie mniejszy. Dlatego metoda Saltona charakteryzuje się szybkim wyszukiwaniem odpowiedzi na pytania szczegółowe i długim czasem przy pytaniach ogólnych.

5. PROCES WYSZUKIWANIA

Wyszukaj wszystkie telefony producenta *Samsung,* które są *wodoodporne* lub producenta *Huawei* o *dużej* przekątnej ekranu.

t = (Producent, Samsung) * (Wodoodporność, TAK) + (Producent, Huawei) *
(Przek. Ekranu, duża)

Dzielimy pytanie na 2 części

t_1 = (Producent, Samsung) * (Wodoodporność, TAK)
t_2 = (Producent, Huawei) * (Przek. Ekranu, duża)

Korzystając ze wzoru

$$p(x_c, x_i) = \frac{\overline{\overline{x_c \cap x_i}}}{\overline{\overline{x_c \cup x_i}}}$$

Dla grupy poprawionej współczynnik podobieństwa wynosi

$$\frac{2}{20} = 0{,}1$$

$\sigma(t_1) = \{ x_2, x_7 \}$

Analogicznie postępujemy z drugim pytaniem, współczynnik również wynosi

$$\frac{2}{20} = 0{,}1$$

$\sigma(t_2) = \{ x_4 \}$

Oprócz tych odpowiedzi w grupie elementów swobodnych znajdują się **x_9** oraz **x_{11}**

Ostatecznie

$$\sigma(t) = \{ x_2, x_4, x_9, x_{11} \}$$

6. REDUNDANCJA

Korzystając ze wzoru

$$R = \frac{\sum_{i=1}^{m} card\, X_i - cardX}{cardX}$$

Gdzie

card X_i - Liczba obiektów w i-tej grupie

cardX - Liczba obiektów w bazie danych

m – liczba grup

Redundancja w naszej bazie danych ze względu na uzyskanie tylko jednej grupy będzie wynosiła:

$$\frac{6}{13}$$

7. AKTUALIZACJA

Przeprowadzimy aktualizację dodając obiekt do grupy obiektów swobodnych.

Dodamy obiekt x14

x14 = (Producent, Goophone)*(Model, X)*(Poj. Baterii, średni)*(Wodoodporność, Nie)*(Przek. Ekranu, średnia)*(L. Rdzeni proc, 6)*(Sys. Operacyjny, Android)*(RAM,1)

Obiekt	Producent	Model	Poj. Baterii	odoodporno	rzek. Ekranu	L.rdzeni proc	ys. Operacyjr	RAM
x12	Sony	Xperia Z3	średnia	Tak	mała	4	Android	3
x13	Nokia	9 PureView	duża	Tak	średnia	8	Android	6
x14	Goophone	X	średnia	Nie	średnia	6	Android	1

Tak więc nasza grupa poprawiona pozostanie bez zmian, obiekt x14 zostanie natomiast dodany do grupy obiektów swobodnych. Chcąc zmienić opis nowo dodanego obiektu x16 należy usunąć go z bazy, a następnie dodać wraz z uaktualnionym opisem. W związku z tym, że dodaliśmy obiekt do grupy obiektów swobodnych, a nie dokonywaliśmy powiązań z istniejącymi grupami, proces ten będzie znacznie szybszy.

8. ALGORYTM DOYLE'A

Wstępny dowolny podział

$$X_1 = \{x_1, x_2, x_3, x_4\}$$
$$X_2 = \{x_5, x_6, x_7, x_8\}$$
$$X_3 = \{x_9, x_{10}, x_{11}, x_{12}, x_{13}\}$$

Wyznaczamy wektory

S1	ATRYBUT	C1	F1	R1	P1
X1	PRODUCENT	Apple	1	4	22
X2		Samsung	1	4	22
X3		Sony	1	4	22
X4		Huawei	1	4	22
	MODEL	Iphone Xs	1	4	22
		Galaxy S10	1	4	22
		Xperia 1	1	4	22
		P30 Pro	1	4	22
	POJ BATERII	Mała	1	4	22
		Średnia	1	4	22
		Duża	1	4	22
		Bardzo duża	1	4	22
	WODOODPORNOŚĆ	Tak	4	1	25
	PRZEK. EKRANU	Średnia	2	3	23
		Duża	2	3	23
	LICZBA RDZENII	6	1	4	22
		8	3	2	24
	SYS. OPERACYJNY	IOS	1	4	22
		Android	3	2	24
	RAM	4	1	4	22
		6	3	2	24

S2	ATRYBUT	C2	F2	R2	P2
X5	PRODUCENT	Xiaomi	1	4	20
X6		Apple	1	4	20
X7		Samsung	1	4	20
X8		Sony	1	4	20
	MODEL	Mi Mix 3	1	4	20
		8 Plus	1	4	20
		Galaxy Note 10	1	4	20
		Xperia 10 Plus	1	4	20
	POJ BATERII	Duża	1	4	20
		Średnia	2	3	21
		Bardzo duża	1	4	20
	WODOODPORNOŚĆ	Tak	3	2	22
		Nie	1	4	20
	PRZEK. EKRANU	Duża	3	2	22
		Średnia	1	4	20
	LICZBA RDZENII	8	3	2	22
		6	1	4	20
	SYS. OPERACYJNY	Android	3	2	22
		IOS	1	4	20
	RAM	6	1	4	20
		3	1	4	20
		8	1	4	20
		4	1	4	20

S3	ATRYBUT	C3	F3	R3	P3
X9	PRODUCENT	Huawei	1	5	25
X10		Apple	1	5	25
X11		Samsung	1	5	25
X12		Sony	1	5	25
X13		Nokia	1	5	25
	MODEL	P Smart Z	1	5	25
		Iphone 6s	1	5	25
		Galaxy S6	1	5	25
		Xperia Z3	1	5	25
		9 PureView	1	5	25
	POJ BATERII	Bardzo mała	1	5	25
		Mała	1	5	25
		Średnia	1	5	25
		Duża	1	5	25
		Bardzo duża	1	5	25
	WODOODPORNOŚĆ	Nie	3	3	27
		Tak	2	4	26
	PRZEK. EKRANU	Mała	3	3	27
		Średnia	1	5	25
		Duża	1	5	25
	LICZBA RDZENII	2	1	5	25
		4	1	5	25
		8	3	3	27
	SYS. OPERACYJNY	Android	4	2	28
		IOS	1	5	25
	RAM	2	1	5	25
		3	2	4	26
		4	1	5	25
		6	1	5	25

Wstępne założenia przyjęte przy naszym algorytmie:

m – liczba grup = 3

a – współczynnik skalujący z przedziału <0, 1> = 0,5

T – wartość progowa = 175

wektor S_j – wektor dokumentów

wektor C_j – wektor pojęć występujących w j-tej grupie

wektor F_j - wektor częstości występowania pojęć

wektor R_j – wektor rang przyporządkowanym pojęciom grupy

wektor P_j – wektor wartości pozycyjnych, gdzie pi = (b -r_1)

b – wcześniej wyznaczona wartość bazowa

b_1 = 25, bo C1.length() + 1 = 25 (b_1 to wartość bazowa dla X_1)

b_2 = 23, bo C2.length() + 1 = 23 (b_2 to wartość bazowa dla X_2)

b_3 = 29, bo C3.length() + 1 = 29 (b_3 to wartość bazowa dla X_3)

Wyliczamy wartość funkcji punktującej:

g(d1, Pj)- wartość funkcji punktującej

	P1	P2	P3
X1	180	-	-
X2	186	-	-
X3	186	-	-
X4	186	-	-
X5	-	166	-
X6	-	163	-
X7	-	168	-
X8	-	169	-
X9	-	-	207
X10	-	-	204
X11	-	-	210
X12	-	-	207
X13	-	-	206

Grupujemy wstępnie

$$Hj = \max(g(di, Pj))$$

$$H1 = 186$$

$$H2 = 169$$

$$H3 = 210$$

$$T1 = 186 – 0,5 * (186\text{-}175) = 180,5$$

$$T2 = 169 – 0,5 * (175 – 169) = 166$$

$$T3 = 210 – 0,5 * (210 – 175) = 192,5$$

Zatem grupy są nastepujące

$$X_1 = \{x_2, x_3, x_4\}$$

$$X_2 = \{x_1, x_5, x_7, x_8\}$$

$$X_3 = \{x_9, x_{10}, x_{11}, x_{12}, x_{13}\}$$

Dokonaliśmy podziału na *m+1* grup, gdyż powstała 4 grupa – dokumentów swobodnych

$$L = \{ x_6 \}$$

Można zauważyć, że kolejne iteracje nie zmienią grup, ponieważ dokumentów swobodnych nie da się powiązać z żadną inną grupą.

Wektory pojęć są nastepujące

Dla P1:

- wektor pojęć D1

(Producent, Samsung) * (Producent, Sony) * (Producent, Huawei) * (Model, Galaxy S10) * (Model, Xperia 1) * (Model, P30 Pro) * (Poj. Baterii, średnia) * (Poj. Baterii, duża) * (Poj. Baterii, b. duża) * (Wodoodporność, Tak) * (Przek. Ekranu, średnia) * (Przek. Ekranu, duża) * (L. Rdzen. Proc, 8) * (Sys. Operacyjny, Android) * (RAM, 6)

$$X_1 = \{x_2, x_3, x_4\}$$

Dla P2:

- wektor pojęć D2

(Producent, Apple) * (Producent, Xiaomi) * (Producent, Samsung) * (Producent, Sony) * (Model, Iphone Xs) * (Model, MiMix 3) * (Model, Galaxy Note10) * (Model, Xperia 10 Plus) * (Poj. Baterii, mała) * (Poj. Baterii, duża) * (Poj. Baterii, b. duża) * (Poj. Baterii, średnia) * (Wodoodporność, Tak) * (Wodoodporność, Nie) * (Przek. Ekranu, średnia) * (Przek. Ekranu, duża) *(L. Rdzen. Proc, 6) *(L. Rdzen. Proc, 8) * (Sys. Operacyjny, IOS) * (Sys. Operacyjny, Android) * (RAM, 4) * (RAM, 6) * (RAM, 8)

$$X_2 = \{x_1, x_5, x_7, x_8\}$$

Dla P3:

- wektor pojęć D3

(Producent, Huawei) * (Producent, Samsung) * (Producent, Sony) * (Producent, Apple) * (Producent, Nokia) * (Model, P Smart Z) * (Model, Iphone 6s) * (Model, Galaxy S6) * (Model,Xperia Z3) * (Model, 9 Pure View) * (Poj. Baterii, b. mała) * (Poj. Baterii, mała) * (Poj. Baterii, średnia) * (Poj. Baterii, duża) * (Poj. Baterii, b.duża) * (Wodoodporność, Tak) * (Wodoodporność, Nie) * (Przek. Ekranu, mała) * (Przek. Ekranu, średnia) * (Przek. Ekranu, duża) * (L. Rdzen. Proc, 2) * (L. Rdzen. Proc, 4) * (L. Rdzen. Proc, 8) * (Sys. Operacyjny, Android) * (Sys. Operacyjny, IOS) * (RAM, 2) * (RAM, 3) * (RAM, 4) * (RAM, 6)

$$X_3 = \{x_9, x_{10}, x_{11}, x_{12}, x_{13}\}$$

- grupa swobodna

(Producent, Apple) * (Model, Iphone 8 Plus) * (Poj. Baterii, średnia) * (Wodoodporność,Tak) * (Przek. Ekranu, średnia) * (L. Rdzen. Proc, 6) * (Sys. Operacyjny, IOS) * (RAM, 3)

$$L = \{ x_6 \}$$

9. PROCES WYSZUKIWANIA

Wyszukaj wszystkie telefony producenta *Samsung*, które są *wodoodporne* lub producenta *Huawei* o *dużej* przekątnej ekranu.

t = (Producent, Samsung) * (Wodoodporność, TAK) + (Producent, Huawei) *

(Przek. Ekranu, duża)

Dzielimy pytanie na 2 części

t_1 = (Producent, Samsung) * (Wodoodporność, TAK)

t_2 = (Producent, Huawei) * (Przek. Ekranu, duża)

Znajdziemy najpierw odpowiedź na pytanie składowe t_1

Porównujemy deskryptory z wektorami pojęć grup, wybierając grupy stanowiące przybliżoną odpowiedź na nasze pytanie. Uzyskujemy odpowiedź przybliżoną σ(t1) = {X1 ∪ X2}. Następnie metodą przeglądu zupełnego sprawdzamy opisy wektorów pojęć 1 oraz 2 oraz opisy obiektów swobodnych znajdując odpowiedź dokładną t1.

Ostatecznie

$$\sigma(t_1) = \{ x_2, x_7\}$$

Analogicznie postępujemy w przypadku szukania odpowiedzi na pytanie zawarte w termie składowym t2

Porównując deskryptor z wektorami pojęć uzyskujemy odpowiedź przybliżoną σ(t1) = {X1 U X3 }. Następnie metodą przeglądu zupełnego sprawdzamy opisy wektorów tej grupy jak i opisy obiektów swobodnych w poszukiwaniu odpowiedzi dokładnej.

$$\sigma(t_2) = \{ x_4, x_9\}$$

Tak więc odpowiedź na nasze pytanie p zawarte w postaci termu t

$$\sigma(t_1) = \{ x_2, x_7, x_4, x_9\}$$

10. REDUNDANCJA

W przypadku algorytmu Doyle'a jeżeli grupowanie obiektów odbywa się z założeniem grupowania tylko obiektów nie związanych, to metoda nie wnosi redundancji. Założenie to zwykle jest przyjmowane przy stosowaniu tego algorytmu.