Metoda Saltona - wyszukiwanie informacji w strukturach drzewiastych

2 grudnia 2008

1 Główne cechy metody Saltona

- Metoda Saltona - opracowana dla dokumentów i pytań zadawanych w języku naturalnym, dlatego też podstawowy moduł stanowi moduł analizy językowej, którego opracowanie jest niezwykle pracochłonne i wymaga rozwiązania szeregu problemów natury lingwistycznej.
- Zrealizowany system $SMART^*$ oparty na metodzie Saltona zajmuje się wyszukiwaniem dokumentów opisanych w języku angielskim.
- W metodzie Saltona opisy obiektów są tekstami w języku naturalnym. Metoda polega na podziale wszystkich obiektów na grupy o podobnym opisie. Istnieje wiele sposobów takiego grupowania. Każda grupa obiektów jest poprzedzona określonym wektorem pojęć charakterystycznych dla danej grupy (wektor centriodalny, profil).
- Wyszukiwanie odpowiedzi polega na porównaniu pytania z wektorami pojęć charakteryzujących poszczególne grupy obiektów, a następnie wybraniu grup o wektorze najbardziej zbliżonym do pytania. Obiekty występujące w tych grupach stanowią tzw. odpowiedź przybliżoną na pytanie. Następnie dokonuje się przeglądu zupełnego wybranych obiektów dla znalezienia odpowiedzi dokładnej, tzn. obiektów, których opisy dokładnie odpowiadają pytaniu (zawierają identyczne pojęcia jak w pytaniu). W przypadku otrzymania dużej liczby grup w BD stosuje się dalsze ich łączenie w grupy większe, tworząc strukturę drzewiastą. Pojęcia charakteryzujące duże grupy (pnie) zawierają zbiory wektorów pojęć grup, a te dopiero - zbiory obiektów.

$SMART^*$- automatyczny system wyszukiwania dokumentów zaprojektowany na Uniwersytecie Harvarda w latach 1961 - 1964. System przyjmuje dokumenty i żądania usług sformułowane w języku naturalnym, dokonuje automatycznej analizy tekstów przy użyciu jednej z kilkudziesięciu metod analizy językowej, kojarzy przeanalizowane dokumenty z kwerendami i wyszukuje dla użytkownika te pozycje, które uzna za najbardziej odpowiadające zgłoszonym kwerendom.

1.1 PROCES WYSZUKIWANIA

Proces wyszukiwania w systemie Smart można podzielić na 5 etapów:

1. wprowadzenie tekstu drukowanego
2. grupowanie dokumentów dla celów przeszukiwania (wiązanie w grupy)
3. wybranie grupy dokumentów do wyszukiwania
4. przeszukiwanie grupy dokumentów
5. ocena wyszukiwania.

Grupowanie polega na umieszczeniu w tej samej grupie dokumentów zawierających podobne pojęcia, oraz na określeniu dla każdej grupy reprezentatywnej pozycji centralnej (CENTROID). Po utworzeniu kartoteki dokumentów powiązanych w grupy, przeszukiwanie grup polega na uprzednim dobieraniu kwerend do centroidów każdej grupy. Następnie dokonuje się wyboru grup, które prawdopodobnie zawierają najwłaściwsze dokumenty, po czym następuje przeszukiwanie grup przy użyciu normalnej procedury - pozycja za pozycją. Istnieje wiele sposobów grupowania. My poznamy 2 metody:

- wg algorymtu Rocchia
- wg algorytmu Doyle'a

Zarówno proces grupowania, jak i proces porównywania pytania z pniami czy wektorami pojęć odbywa się poprzez znajdowanie współczynników korelacji (podobieństwa) pomiędzy pojęciami występującymi w pytaniu lub pojęciami występującymi w wektorze pojęć danej grupy.

1.2 Miary korelacji (podobieństwa)

Współczynnik korelacji to wartość z przedziału $< 0, 1 >$. Im bardziej podobne są do siebie obiekty tym wyższy jest dla nich współczynnik korelacji.
Jeżeli dwa obiekty są identyczne to współczynnik korelacji = 1.
Dla obiektów w ogóle nie podobnych współczynnik korelacji = 0.
I tak dla dwóch obiektów x1 i x2 poniżej przedstawione są typowe miary korelacji:

$$f_1(x1, x2) = \frac{\overline{\overline{x1 \cap x2}}}{\overline{\overline{x1 \cup x2}}}$$

$$f_2(x1, x2) = \frac{\overline{\overline{x1 \cap x2}}}{\overline{\overline{A}}}$$

$$f_3(xc, x2) = \frac{\overline{\overline{xc \cap x2}}}{\overline{\overline{xc}}}$$

W systemie SMART Saltona istnieją dwie miary korelacji:

- korelacja cosinusowa

$$cos(d,q) = \frac{\sum_{k=1}^{n} d_i * q_i}{\sqrt{\sum_{k=1}^{n}(d_i)^2 * \sum_{k=1}^{n}(q_i)^2}}$$

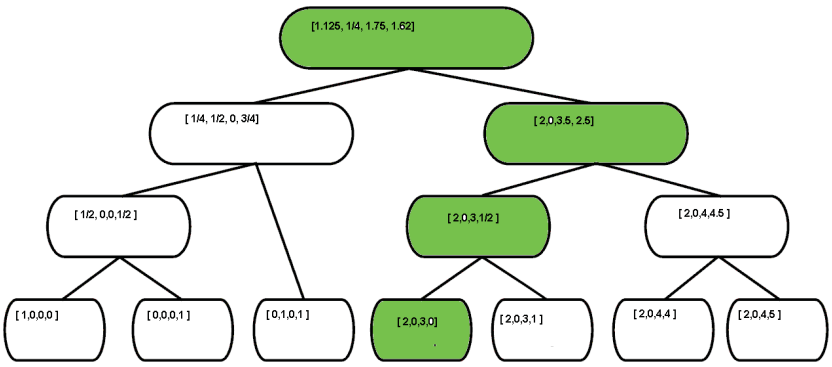
- korelacja nakładania

$$ovlap(d,q) = \frac{\sum_{k=1}^{n} \min(d_i, q_i)}{\min(\sum_{k=1}^{n} d_i, \sum_{k=1}^{n} q_i)}$$

gdzie: d i q to n-wymiarowe wektory terminów reprezentujących analizowaną kwerendę q i analizowany dokument d.

2 STRUKTURA KARTOTEKI

Czyli mamy system $S =< X, A, V, q >$.
Opisy obiektów pogrupowane są w BD w grupy X_i, gdzie $i = 1, .., m$ przy czym spełniony jest warunek: $X = \bigcup_{i=1}^{m} X_i$
Struktura kartoteki ma więc formę drzewiastą (hierarchię) w której dokumenty podobne do siebie łączone są w grupy, dla których tworzy reprezentantów (centroid bądź profil). Jeśli grup tak utworzonych jest dużo, traktowane są one jak dokumenty i ponownie grupowane w grupy a kolejnym poziomie hierarchii (pnie).
Każda grupa X_i poprzedzona jest identyfikatorem grupy, który jest nazywany CENTROIDEM (C_i) lub PROFILEM (P_i): $X_i = (C_i, \{t_{x_i}\})$.
Centroid - C_i to wektor pojęć opisujących dokumenty danej grupy. Stosowany do opisu grupy w algorytmie Rocchio'a.
Profil - P_i to wektor wartości pozycyjnych pojęć opisujących dokumenty danej grupy. Stosowany do opisu grupy w algorytmie Doyle'a.



Rysunek 1: Struktura drzewiasta w systemie Smart Saltona

3 Algorytm Rocchio'a

1. Pobranie opisów obiektów.
2. Ustalenie parametrów:
   $P1, P2, N1, N2$ - dla centrum grupy,
   $P1p, P2p, N1p, N2p$ - dla centroidu.
3. Wybranie potencjalnego centrum grupy: x_c.
4. Przeprowadzamy test gęstości dla centrum grupy x_c,(co najmniej $N1$ dokumentów ma współczynnik większy bądź równy od $P1$, a $N2$ dokumentów ma współczynnik większy bądź równy $P2$). W tym celu obliczamy współczynniki korelacji dokumentów z potencjalnym centrum grupy.
   - Jeżeli założenia nie są spełnione to konieczny jest wybór innego potencjalnego centrum grupy lub zmiana parametrów tesktu gęstości (punkt 3).
   - Jeśli potencjalne centrum grupy przeszło test gęstości: przechodzimy do punktu 5.
5. Określamy rangę obiektów.
6. Wyznaczamy $M1$ (liczebność zbioru obiektów dla których elementy są większe bądź równe P2) , $M2$ (liczebność zbioru obiektów dla których elementy są większe bądź równe P1).
   - Jeśli $M1 = M2$ to Pmin równa się najmniejszemu współczynnikowi korelacji obiektu należącego do M1,przechodzimy do punktu 11.
   - Jeśli $M1 \neq M2$ to:
     - Obliczamy różnicę pomiędzy współczynnikami korelacji obiektów sąsiednich w grupie maksymalnej $M2$,bez obiektów grupy minimalnej $M1$.
     - Określamy największą różnicę.
     - Minimalny współczynnik korelacji Pmin jest równy odjemnej z największej różnicy. Jeśli największa różnica powtarza się to za Pmin przyjmujemy odjemną o większej wartości.
7. Tworzymy wstępną grupę do której należą elementy o współczynniku korelacji większym bądź równym P_{min}.
8. Tworzymy wektor centroidalny, który stanowi sumę opisów obiektów należących do grupy wstępnej.

   II-ga iteracja algorytmu - dla tworzenia tzw. grupy poprawionej.
9. Przeprowadzamy test gęstości dla centroidu, (co najmniej $N1p$ dokumentów ma współczynnik większy bądź równy od $P1p$, a $N2p$ dokumentów ma współczynnik większy bądź równy $P2p$).

10. Obliczamy współczynniki korelacji dokumentów z centroidem.
11. Określamy rangę obiektów.
12. Wyznaczamy $M1p$ (liczebność zbioru obiektów dla których elementy są większe bądź równe $P2p$) , $M2p$ (liczebność zbioru obiektów dla których elementy są większe bądź równe $P1p$).
    - Jeśli $M1p = M2p$ to $Pmin$ równa się najmniejszemu współczynnikowi korelacji obiektu należącego do M1p,przechodzimy do punktu 19.
    - Jeśli $M1p \neq M2p$ to:
        - Obliczamy różnicę pomiędzy współczynnikami korelacji obiektów sąsiednich w grupie maksymalnej M2p,bez obiektów grupy minimalnej M1p.
        - Określamy największą różnicę.
        - Minimalny współczynnik korelacji Pmin jest równy odjemnej z największej różnicy. Jeśli największa różnica powtarza się to za Pmin przyjmujemy odjemną o większej wartości.
13. Tworzymy grupę poprawioną do której należą elementy o współczynniku korelacji większym bądź równym P_{min}.
14. Tworzymy wektor centroidalny, który stanowi sumę opisów obiektów należących do grupy poprawionej.
15. Obiekty nie należące do grupy poprawionej (swobodne),traktujemy jako wejściowe opisy obiektów i przechodzimy do punktu 3.

3.1 Przykład

Wykorzystując opis (poniżej) algorytmu Rocchia przeprowadź grupowanie 10 obiektów o następujących opisach:
$x1 = a1\ b1\ c1\ d1\ e1$
$x2 = a1\ b1\ c1\ d1\ e2$
$x3 = a1\ b1\ c2\ d1\ e3$
$x4 = a1\ b1\ c3\ d1\ e1$
$x5 = a1\ b1\ c1\ d1\ e3$
$x6 = a2\ b1\ c2\ d1\ e2$
$x7 = a2\ b1\ c3\ d1\ e3$
$x8 = a2\ b2\ c3\ d3\ e3$
$x9 = a3\ b3\ c2\ d2\ e2$
$x10 = a3\ b3\ c2\ d3\ e2$

1. Dla podanego wyżej zbioru obiektów dane są następujące parametry:
   a) Dla centrum grupy: $N1 = 5$, $N2 = 3$, $p1 = 0,2$, $p2 = 0,3$
   b) Dla centroidu: $N1c = 5$, $N2c = 3$, $p1c = 0,25$, $p2c = 0,35$

2. Wybór potencjalnego centrum grupy x_c
   Jako potencjalne centrum grupy 1 przyjmij obiekt – x_1.

3. Wybór miary podobieństwa (korelacji) każdego dokumentu z centrum grupy x_c

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

4. Przeprowadzamy test gęstości dla centrum grupy (x_c). Test ten mówi, że co najmniej $N1$ dokumentów ma współczynnik większy bądź równy od $P1$, a $N2$ dokumentów ma współczynnik większy bądź równy $P2$.

   - W tym celu obliczamy współczynniki korelacji (podobieństwa każdego dokumentu (x_i) z wybranym centrum grupy x_c) stosując wybraną wcześniej miarę korelacji.
     Gdy mamy 10 dokumentów w systemie to po kolei dla każdego dokumentu wyliczamy taki współczynnik:

     $p(x_1, x_c) =?$
     ...
     $p(x_{10}, x_c) =?$

     /*W liczniku podajemy liczbę pojęć wspólnym danego dokumentu z centrum grupy x_c W mianowniku podajemy sumę pojęć, którymi są opisane obydwa dokumenty: dany dokument x_i i dokument stanowiący centrum grupy.
     zatem:
     Aby obliczyć współczynnik korelacji obiektu 1 z centrum grupy – który jest jednocześnie obiektem 1 wykonujemy następujące czynności.
     $x_1 = a1\ b1\ c1\ d1\ e1$
     x_c=$a1\ b1\ c1\ d1\ e1$
     Liczba pojęć wspólnych = 5, bo są to pojęcia: $(a1, b1, c1, d1, e1)$
     Suma wszystkich pojęć = 5, bo są to pojęcia: $(a1, b1, c1, d1, e1)$*/

   - Zatem:
     $p(xc, x1) = 5/5 = 1.0$
     $p(xc, x2) = 4/6 = 0.67$
     $p(xc, x3) = 3/7 = 0.43$
     $p(xc, x4) = 4/6 = 0.67$
     $p(xc, x5) = 4/6 = 0.67$
     $p(xc, x6) = 2/8 = 0.25$
     $p(xc, x7) = 2/8 = 0.25$
     $p(xc, x8) = 0/10 = 0$
     $p(xc, x9) = 0/10 = 0$
     $p(xc, x10) = 0/10 = 0$

- Określamy rangę dokumentów, czyli porządkujemy dokumenty malejąco według obliczonych w kroku 5 współczynników korelacji i nadajemy tak ułożonym wartościom rangi od 1 do n.
  Ranga 1: $p(x1, xc) = 1.0$
  Ranga 2: $p(x2, xc) = 0.67$
  Ranga 3: $p(x4, xc) = 0.67$
  Ranga 4: $p(x5, xc) = 0.67$
  Ranga 5: $p(x3, xc) = 0.43$
  Ranga 6: $p(x6, xc) = 0.25$
  Ranga 7: $p(x7, xc) = 0.25$
  Ranga 8: $p(x8, xc) = 0.0$
  Ranga 9: $p(x9, xc) = 0.0$
  Ranga 10: $p(x10, xc) = 0.0$

- Przeprowadzamy test gęstości – czyli sprawdzamy, czy na pewno:
  N1 dokumentów ma $p >= p1$ i N2 dokumentów ma współczynnik $p >= p2$ Jeśli tak to znaczy, że wybrane centrum grupy przeszedł test gęstości.

- Jeżeli założenia nie są spełnione: wybieramy inny obiekt jako centrum grupy (x_c).
- Jeżeli założenia są spełnione: przechodzimy do punktu 5.

5. Obliczamy faktyczne rozmiary grupy. Wyznaczamy $M1$ (liczebność zbioru obiektów dla których elementy są większe bądź równe $P2$) , $M2$ (liczebność zbioru obiektów dla których elementy są większe bądź równe $P1$).
   $M1 = 5$ zaś $M2 = 7$

6. Obliczamy minimalny współczynnik korelacji p_{min}:

   - Jeśli $M1 = M2$ to:
     to p_{min} równa się najmniejszemu współczynnikowi korelacji obiektu należącego do $M1$
   - Jeśli $M1 < M2$ to:
     Obliczamy różnicę pomiędzy współczynnikami korelacji obiektów sąsiednich w grupie maksymalnej $M2$,bez obiektów grupy minimalnej $M1$. Wybieramy największą różnicę i obliczamy minimalny współczynnik korelacji p_{min} jako odjemną z tej największej różnicy.
     - Jeśli największa różnica powtarza się to za p_{min} przyjmujemy odjemną o większej wartości.
   - W naszym przypadku: $M1 = 5$ a $M2 = 7$, zatem są to różne wartości, więc, aby obliczyć współczynnik korelacji p_{min} obliczamy różnicę między dokumentami na granicy tych grup.
     5 różnica z 6: 0,43 – 0,25 = 0, 18
     6 różnica z 7: 0,25 – 0,25 = 0
     7 różnica z 8: 0,25 – 0 = 0,25

Minimalny współczynnik korelacji p_{min} jest równy odjemnej z największej różnicy.
$p_{min} = p_7(x_7) = 0,25$

7. Wyznaczamy grupę wstępną X_{1_w}:
Do grupy wstępnej będą należały wszystkie te dokumenty, które miały wyliczony współczynnik korelacji większy lub równy p_{min}.
Są to wszystkie obiekty grupy maksymalnej $M2$:
$x1, x2, x3, x4, x5, x6$ i $x7$.

8. Wyznaczamy wstępnego reprezentanta grupy X_1 – czyli centroid:
Centroid to zbiór wszystkich pojęć, którymi są opisane dokumenty grupy minimalnej M1, czyli...:$C_{w_1} = \{a1, b1, c1, c2, c3, d1, e1, e2, e3\}$

DRUGA ITERACJA

9. Generujemy grupę poprawioną:
W tym celu powtarzamy raz jeszcze cały algorytm, z tym,
że teraz centrum grupy stanowi teraz $CENTROID\ C_1$.

10. Ustalenie parametrów testu gęstości dla centroidu:
$p1c = 0,25$;$p2c = 0,35$;$N1c = 5$;$N2c = 3$

11. Test gęstości dla centroidu:

- W tym celu obliczamy współczynniki korelacji (podobieństwa) dokumentów grupy maksymalnej M2 z centroidem C_1.
$P(x1, c1) = 5/9 = 0.55$
$P(x2, c1) = 5/9 = 0.55$
$P(x3, c1) = 5/9 = 0.55$
$P(x4, c1) = 5/9 = 0.55$
$P(x5, c1) = 5/9 = 0.55$
$P(x6, c1) = 4/10 = 0.4$
$P(x7, c1) = 4/10 = 0.4$
- Określamy rangę dokumentów:
$Ranga1\ p(x1, xc) = 0.55$
$Ranga2\ p(x2, xc) = 0.55$
$Ranga3\ p(x4, xc) = 0.55$
$Ranga4\ p(x5, xc) = 0.55$
$Ranga5\ p(x3, xc) = 0.55$
$Ranga6\ p(x6, xc) = 0.4$
$Ranga7\ p(x7, xc) = 0.4$

- Sprawdzamy, czy na pewno: $N1c$ dokumentów ma $p >= p1c$ i $N2c$ dokumentów ma współczynnik $p >= p2c$
Jeśli tak to znaczy, że wybrane centrum grupy przeszedł test gęstości. Jeśli nie to zmieniamy parametry testu gęstości dla centroidu,

bądź zaczynamy cały algorytm od nowa łącznie z wyborem nowego potencjalnego centrum grupy x_c.

12. Obliczamy faktyczne rozmiary grupy poprawionej:
    Wyznaczamy $M1$ (liczebność zbioru obiektów dla których elementy są większe bądź równe $P2$) , $M2$ (liczebność zbioru obiektów dla których elementy są większe bądź równe $P1$).

    - Jeśli $M1 = M2$ to: to p_{min} równa się najmniejszemu współczynnikowi korelacji obiektu należącego do $M1$ czyli $pmin = p7(x7) = 0,4$ $m1 = m2 = 7$

13. Wyznaczamy grupę poprawioną X_1
    Do tej grupy będą należały wszystkie te dokumenty, które miały wyliczony współczynnik korelacji większy lub równy p_{min}.
    Są to wszystkie obiekty grupy maksymalnej M2:
    $X_1 = \{x1, x2, x3, x4, x5, x6, x7\}$

14. Wyznaczamy reprezentanta grupy X_1 – czyli centroid
    Centroid to zbiór wszystkich pojęć, którymi są opisane wszystkie dokumenty grupy X_1, czyli...
    $C_1 = \{a1, a2, b1, c1, c2, c3, d1, e1, e2, e3\}$
    KONIEC GENEROWANIA PIERWSZEJ GRUPY.

Zatem jedna iteracja algorytmu doprowadziła do powstania grupy:
$X_1 = \{X1, x2, x3, x4, x5, x6, x7\}$
Na jej czele stoi centroid
$C1 = \{a1, a2, b1, c1, c2, c3, d1, e1, e2, e3\}$
Co dalej ???????
Z dokumentów pozostałych $X – X1 = \{x8, x9, x10\}$ powinniśmy tworzyć kolejne grupy. Jednakże jak łatwo zauważyć patrząc na ustalone na początku parametry testu gęstości nie możliwe będzie utworzenie następnych grup, gdyż test ten wymaga aby...grupa maksymalna liczyła co najmniej $N2 = 5$ dokumentów...a nam zostały już tylko 3 .... Zatem na tym kończy się algorytm.

4 Algorytm Doyle'a

Zakładamy następujące wartości:
m - liczba grup
T - wartość progowa
a - współczynnik skalujący z przedziału - $< 0, 1 >$

1. Dokonujemy wstępnego podziału zbioru dokumentów na m grup
2. dla każdej grupy wyznaczamy:
   - 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 przyporządkowanych pojęciom grupy
   - Wektor P_j - wektor wartości pozycyjnych (PROFIL) gdzie: $p_i = (b - r_i)$ * wcześniej wyznaczamy wartość bazową "b".
3. dla każdego d_i wyliczamy wartość funkcji punktującej $g(d_i, P_j)$ w każdej grupie zawierającej wszystkie pojęcia opisujące obiekt d_i.
   * wybieramy wartość maksymalną !!!
4. Na podstawie wyznaczonych wartości funkcji punktującej dokonaj wstępnego podziału dokumentów do grup tak, że:
   $S_j = \{d_i : g(d_i, P_j) \geqslant Tj\}$

   $$T_j = \begin{cases} T & \text{dla } H_j \leqslant T \\ H_j - a(H_j - T) & \text{dla } H_j > T \end{cases}$$

   Gdzie: $H_j = \max(g(d_i, P_j))$
   * z reguły powstaje $m + 1$ grup (bo m grup + grupa dokumentów swobodnych)
5. Jeśli podział w $i + 1$-ej iteracji jest identyczny jak w i-tej to KONIEC algorytmu.

REZULTAT:
m- grup dokumentów (na czele każdej grupy stoi PROFIL) i ewentualnie grupa dokumentów swobodnych (L).

4.1 Przykład algorytmu Doyle'a

Dla podanego zbioru obiektów przeprowadź jedną iterację grupowania algorytmem Doyle'a przy założeniach: liczba grup wynosi $m = 3$, współczynnik $a = 0,5$.
t_{x1} = (Pł,K)(TY,DR)(SP,5)(OZ,c)
t_{x2} = (Pł,M)(TY,PR)(SP,2)(OZ,b)
t_{x3} = (Pł,M)(TY,MGR)(SP,5)(OZ,c)
t_{x4} = (Pł,M)(TY,MGR)(SP,2)(OZ,a)
t_{x5} = (Pł,M)(TY,PR)(SP,12)(OZ,d)
t_{x6} = (Pł,M)(TY,DR)(SP,5)(OZ,b)
t_{x7} = (Pł,K)(TY,DR)(SP,2)(OZ,b)
t_{x8} = (Pł,M)(TY,MGR)(SP,12)(OZ,c)
t_{x9} = (Pł,M)(TY,PR)(SP,5)(OZ,d)
t_{x10} =(Pł,K)(TY,PR)(SP,2)(OZ,d)
I ITERACJA

1. Tworzymy wektory opisujące każdą grupę:

X1	C1	F1	R1	P1	X2	C2	F2	R2	P2	X3	C3	F3	R3	P3
x1	k	1	3	8	x5	M	2	1	10	x8	M	2	1	10
x2	M	3	1	10	x6	K	1	2	9	x9	K	1	2	9
x3	DR	1	3	8	x7	PR	1	2	9	x10	MGR	1	2	9
x4	PR	1	3	8		DR	2	1	10		PR	2	1	10
	MGR	2	2	9		12	1	2	9		2	1	2	9
	2	2	2	9		2	2	1	10		5	1	2	9
	5	2	2	9		b	2	1	10		12	1	2	9
	a	1	3	8		d	1	2	9		c	1	2	9
	b	1	3	8							d	2	1	10
	c	2	2	9										

2. Obliczamy wartość funkcji punktującej $g(d_i, P_j)$ dla każdego dokumentu d_i i profilu P_j:

$g(xi, Pj)$	$P1$	$P2$	$P3$
x1	34	-	-
x2	35	39	-
x3	37	-	37
x4	36	-	-
x5	-	37	39
x6	35	40	-
x7	33	39	-
x8	-	-	37
x9	-	-	39
x10	-	37	38

3. Dla każdej grupy ustalamy wartość progową T_j, którą muszą spełnić dokumenty aby wejść do danej grupy. Wartość progową obliczamy wg jednego z poniższych wzorów:
   - $T = \frac{min(g(d_i,P_j))+max(g(d_i,P_j))}{2} = 37$,

- $T = \overline{g(d_i, P_j)} = 36,5.$

Przyjmijmy więc, że $T = 37$.

4. Nowy podział na grupy ustalamy zgodnie ze wzorem podanym poniżej. Do nowych grup będą należeć obiekty, których wartości funkcji punktującej będą $\geqslant T_j$, czyli większe bądź równe od wartości progowej j-tej grupy.

$$T_j = \begin{cases} T & \text{dla } H_j \leqslant T \\ H_j - a(H_j - T) & \text{dla } H_j > T \end{cases}$$

Gdzie: $H_j = \max(g(d_i, P_j))$
Wyznaczamy maksymalną wartość funkcji punktującej j-tej grupy: $H_j = \max(g(d_i, P_j))$
$H_1 = 37, H_2 = 40, H_3 = 39$
Następnie wartości progowe danych grup (T_j), przy założeniu, że $a = 0.5$.
$T_1 = H_1 - a(H_1 - T) = 37$
$T_2 = H_2 - a(H_2 - T) = 40 - 0,5 * (40 - 37) = 38,5$
$T_3 = H_3 - a(H_3 - T) = 39 - 0,5 * (39 - 37) = 38$

5. OTRZYMANE GRUPY:
Porównując wartości funkcji punktującej z wartościami progowymi według wzoru
Otrzymujemy nowe grupy których jest m+1 ponieważ tworzy się jeszcze jedna grupa , grupa obiektów swobodnych (niesklasyfikowanych).
$X_1 = \{x3\}$, $X_2 = \{x2, x7\}$, $X_3 = \{x5, x9, x10\}$, Grupa obiektów swobodnych: $L = \{x1, x4, x6, x8\}$

II ITERACJA
Aby wykonać kolejną iterację algorytmu przyporządkujemy obiekty swobodne do grup ale innych niż występowały w poprzedniej iteracji, wtedy otrzymujemy nowy podział grup. Cała operacja kolejnych iteracji się kończy, kiedy otrzymujemy po raz kolejny ten sam podział.

1. Tworzymy wektory opisujące każdą grupę:

X1	C1	F1	R1	P1	X2	C2	F2	R2	P2	X3	C3	F3	R3	P3

2. Obliczamy wartość funkcji punktującej $g(d_i, P_j)$ dla każdego dokumentu d_i i profilu P_j:

$g(xi, Pj)$	$P1$	$P2$	$P3$
x1			
x2			
x3			
x4			
x5			
x6			
x7			
x8			
x9			
x10			

Wartośc progowa: $T = \frac{min(g(di,Pj))+max(g(di,Pj))}{2} =$
$H_1 =$, $H_2 =$,$H_3 =$
Wartości progowe grup:
$T_1 =$
$T_2 =$
$T_3 =$
OTRZYMANE GRUPY:
$X_1 = \{$
$X_2 = \{$
,$X_3 = \{$
, oraz grupa obiektów swobodnych:.$L = \{$

5 Wyszukiwanie w systemie Saltona

Proces wyszukiwania składa się z czterech etapów.
Najpierw formułujemy kwerendę, posługując się oryginalnym żądaniem autora albo jego modyfikacją w postaci numerycznego wektora pojęć. Jedną z najważniejszych metod modyfikacji kwerend źródłowych jest korzystanie z dokumentów, które autor ocenił jako relewantne. Z chwilą sformułowania kwerendy selekcjonuje się zbiór dokumentów, które będą z nią korelowane.
Wyróżnia się 2 metody wyszukiwania:

- sekwencyjna - pełna (full search)
- strukturalna (tree search)

5.1 Metoda sekwencyjna

Metoda ta jest niezależna od klasyfikacji dokumentów w grupie. Polega ona na tym, że pytanie kierowane do systemu jest korelowane z każdym dokumentem. Jest liczony współczynnik korelacji - podobieństwa pytania z każdym dokumentem. Wybiera się te dokumenty, w których współczynnik jest większy od założonej wartości progowej (p_{min}). Dla wszystkich dokumentów robiony jest przegląd zupełny. Czyli nie grupujemy dokumentów. Odpowiedź na zadane pytanie otrzymujemy przez przegląd wszystkich, po kolei dokumentów znajdujących się w kartotece wyszukiwawczej. Im więcej będzie dokumentów tym dłuższy będzie czas obliczenia współczynników korelacji. Wada: bardzo wiele zależy od przyjętego współczynnika progowego, im on będzie mniejszy tym więcej obiektów zaliczymy do grupy będącej odpowiedzią na pytanie. Jeśli będzie za wysoki - to może się okazać, ze mało dokumentów spełni warunek wymagalny (tzn. mało będzie miało współczynnik korelacji z pytaniem $\geqslant$ temu założonemu współczynnikowi progowemu).

5.2 Metoda strukturalna

Ta metoda jest ściśle związana ze strukturą bazy danych. Polega na obliczeniu współczynnika korelacji pytania z pniami i wybór pni najbardziej obiecujących, czyli tych o najwyższych współczynnikach korelacji. Wybrane pnie zostają usunięte i następuje obliczanie współczynników korelacji pytania z centroidami (w tych wybranych grupach). Ponownie wybiera się poziomy najbardziej obiecujące na poziomie centroidów i dla tych centroidów, usuwamy je i liczymy współczynniki korelacji dokumentów (tzn. pytania z dokumentami zbioru). Ostatecznie odpowiedzią na pytanie jest zbiór dokumentów, dla których współczynniki korelacji są większe od założonego pmin.

6 PARAMETRY EFEKTYWNOŚCI SYSTEMÓW INFORMACYJNYCH

> Dokument jest relewantny względem pytania Q wtedy i tylko wtedy jeżeli w opisie dokumentu występują wszystkie niezaprzeczone deskryptory pytania Q i w opisie tym nie występuje żaden z deskryptorów zaprzeczonych pytaniem.

Kompletność określa zdolność systemu do wyszukiwania wszystkich dokumentów, które mogą okazać się relewantnymi. Dokładność określa zdolność systemu do nie wyznaczania dokumentów nierelewantnych względem danego pytania Q.
Kompletność:

$$K = \frac{a}{a+c}$$

gdzie:
a - liczba dokumentów relewantnych wyszukanych
c - liczba dokumentów relewantnych niewyszukanych
Dokładność:

$$D = \frac{a}{a+b}$$

gdzie:
a - liczba dokumentów relewantnych wyszukanych
b - liczba dokumentów nierelewantnych wyszukanych.

6.1 Pozostałe parametry efektywności

parametr	wzór	opis
Obcięcie	$O = \frac{a+b}{a+b+c+d}$	prawdopodobieństwo, że dokument wogóle zostal wyszukany Ile jest wyszukanych w stosunku do wszystkich.
Ogólność	$G = \frac{a+c}{a+b+c+d}$	prawdopodobieństwo, że dokument jest relewantny
Odrzut (szum)	$F = \frac{b}{b+d}$	prawdopodobieństwo, że dokument nierelewantny jest wyszukany Ile nieprawidłowych odpowiedzi zostało podanych jako relewantne.

gdzie:
a - liczba dokumentów relewantnych wyszukanych
b - liczba dokumentów nierelewantnych wyszukanych
c - liczba dokumentów relewantnych niewyszukanych
d - liczba dokumentów nierelewantnych niewyszukanych.

6.2 Przykład

W systemie zorganizowanym zgodnie z metodą Saltona występują dokumenty o następujących opisach: $d1 : abe$,$d2 : acef$,$d3 : abec$,$d4 : ab$,$d5 : cde$,$d6 : def$,$d7 : aef$,$d8 : f$,$d9 : efg$,$d10 : ceg$.
Do systemu zadano pytanie $t = ab + f$, na które odpowiedź systemu była nastepująca: $\{d1, d2, d7, d9\}$. Określ wartości parametrów efektywności wyszukiwania.
Rozwiązanie:
Pytanie: $t = ab + f$, zatem $t = t_1 + t_2$, gdzie $t_1 = ab$ zaś $t_2 = f$.
$\sigma(t) = \sigma(t_1) \cup \sigma(t_2)$ Przegląd zupełny wszystkich obiektów powoduje, iż odpowiedzią na pytanie t_1 są dokumenty: $d1, d3, d4$ zaś odpowiedzią na pytanie t_2 są dokumenty: $d2, d6, d7, d8, d9$. Zatem:
$\sigma(t) = \sigma(t_1) \cup \sigma(t_2) = \{d1, d3, d4\} \cup \{d2, d6, d7, d8, d9\} = \{d1, d2, d3, d4, d6, d7, d8, d9\}$.
Jednak system, który podrupował wcześniej te dokumenty zgodnie z jednym z algorytmów grupowania, wykazał, że odpowiedzią na tak zadane pytanie są jedynie dokumenty: $\{d1, d2, d7, d9\}$.
Parametry oceny efektywności wyszukiwania takiego systemu kształtują się zatem następująco:

- dokumenty relewantne:$\{d1, d2, d3, d4, d6, d7, d8, d9\}$ w tym:
  - wyszukane: $\{d1, d2, d7, d9\}$
  - nie wyszukane: $\{d3, d4, d6, d8\}$
- dokumenty nierelewantne: $\{d5, d10\}$ w tym:
  - wyszukane:$\{\emptyset\}$
  - nie wyszukane: $\{d5, d10\}$

Zatem:

$$K = \frac{a}{a+c} = \frac{4}{4+4} = \frac{1}{2}$$

gdzie:
a - liczba dokumentów relewantnych wyszukanych
c - liczba dokumentów relewantnych niewyszukanych
Dokładność:

$$D = \frac{a}{a+b} = \frac{4}{4} = 1.0$$

Uzyskaliśmy pełną dokładność (D), gdyż nie wyszukano nierelewantnych dokumentów. Kompletność wyniosła jedynie 0.5 gdyż spośród 8 relewantnych dokumentów znaleziono jedynie połowę.

7 Analiza algorytmów Doyle'a i Rocchio'a

Dla algorytmu Doyle'a i Rocchio'a scharakteryzuj podane parametry według przedstawionego schematu. Dodaj informacje o parametrach brakujących.

7.1 Algorytm Rocchio'a

Symbol (nazwa)	Opis	zadanie
....	Parametry testu gęstości	Ich wartości wpływają na ....
p_{min}	....	Sposób obliczania ....
....	Stoi na czele grupy (identyfikator)	Tworzą go: ....
....	Współczynnik korelacji	Jeden ze sposobów obliczania ....

7.2 Algorytm Doyle'a

Symbol (nazwa)	Opis	zadanie
....	Zakładana liczba grup	Kiedy ustala się ten parametr? ....
a(albo alfa) T	Zakres wartości dla a:	jak a i T wpływają na efektywność wyszukiwania ....
.....	Stoi na czele grupy (identyfikator)	Tworzą go. ....
T_i	(wartość progowa dla danej grupy)	Sposób obliczania ....

8 Zadania egzaminacyjne

8.1

Przedstaw graficznie reprezentację pni i grup w wyszukiwaniu strukturalnym. Masz do dyspozycji:
P1 = [0,5,5,1,3,1]
P2 = [5,0,0,4,1,5]
P3 = [0,0,0,6,5,1]
Oraz:
G11 = [0,3,3,1,1,2]
G12 = [0,4,4,0,2,0]
G21 = [0,0,1,6,0,3]
G22 = [0,1,1,3,1,2]
G31 = [0,0,1,4,5,0]
G32 = [1,0,1,4,5,0]
G33 = [1,2,1,5,3,0]
Omów sposób wyszukiwania dla pytania Q=[2,1,0,5,3,0]. Wykorzystaj w tym celu wzór na korelację w omawianym procesie wyszukiwania.

8.2

Dany jest zbiór obiektów $X = \{x1...X10\}$, które są opisane pojęciami:
X1: adfg
X2: bcdfhij
X3: aeij
X4:defgh
X5: cehij
X6:adf
X7: bcgj
X8:afghi
X9:dejf
X10:ghij
Dla w/w zbioru obiektów dokonać podziału na 3 grupy algorytmem Doyle'a. Przedstaw proces wyszukiwania dla pytania $t = ab + f$.

8.3

Dla podanego zbioru obiektów przeprowadź jedną iterację grupowania algorytmem Doyle'a przy założeniach: liczba grup wynosi k=2, współczynnik a= 0,8.

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
a1	a3	a1	a1	a3	a4	a4	a4	a2	a1
b2	b2	b2	b2	b2	b2	b1	b2	b1	b1

Zaproponuj wstępny podział na grupy.

8.4

Przedstaw graficznie reprezentację pni i grup w wyszukiwaniu strukturalnym. Masz do dyspozycji: P1=[0, 10, 11, 2, 7, 3, 4, 0, 1, 0]
P2=[10, 0, 0, 9, 2, 11, 2, 10, 7, 0] poziom pni
P3=[1, 0, 1, 12, 10, 3, 10, 4, 6, 11]
Oraz
G11=[0, 6, 7, 2, 3, 4, 0, 1, 3, 2]
G12=[1, 8, 9, 1, 4, 0, 2, 3, 2, 2, 1]
G21=[0, 1, 3, 12, 1, 6, 0, 9, 10, 1]
G22=[1, 3, 2, 6, 2, 5, 0, 10, 9, 3] poziom grup
G31=[1, 0, 3, 9, 10, 0, 12, 2, 4, 0]
G32=[2, 1, 3, 10, 8, 1, 13, 1, 2, 1]
G33=[3, 4, 2, 11, 7, 0, 12, 2, 1, 3]
Omów sposób wyszukiwania dla pytania:
Q=[5, 3, 0, 10, 6, 1, 13, 2, 0, 1]
Wykorzystaj odpowiedni wzór na korelację (f. podobieństwa) w omawianym procesie wyszukiwania.

8.5

Zdefiniuj takie pojęcia jak: Dokładność, Kompletność, Ogólność, Odrzut, Obcięcie. W jakim systemie pojęcia te są pomocne.

8.6

Dany jest zbiór 10 różnych typów chipsów spotykanych na polskim rynku: LAYS, LAYS MAX, CHEETOS, TWISTOS, CRUNCHIPS, SOMBREROS, CURLY, CHIO, MACZUGI i SUPERCHIPS. Przedstaw strukturę poszczególnych grup utworzonych zgodnie z metodą grupowania- algorytmem Doyle'a wykorzystując cechę SMAK (np. orzech, papryka, cebula i ser), sugerującą 4 grupy. Proszę zwrócić uwagę, że nie muszą to być grupy rozłączne. Pozostałe cechy (pojęcia opisujące chipsy) istotne dla sprzedawcy zaproponuj sam . Następnie przeprowadź I krok algorytmu Doyle'a dla tak przygotowanego wstępnego podziału na grupy.