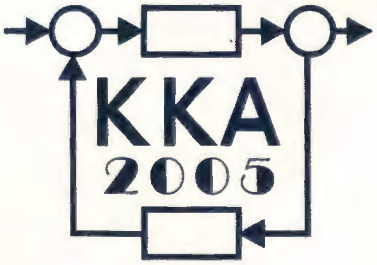
XV Krajowa Konferencja Automatyki

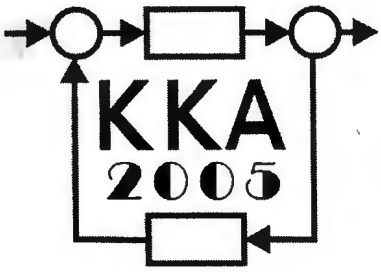
Tom II



Redaktorzy:
Zdzisław Bubnicki
Roman Kulikowski
Janusz Kacprzyk

XV Krajowa Konferencja Automatyki

Tom II



Redaktorzy:
Zdzisław BUBNICKI
Roman KULIKOWSKI
Janusz KACPRZYK

ORGANIZATOR
Komitet Automatyki i Robotyki Polskiej Akademii Nauk
Instytut Badań Systemowych Polskiej Akademii Nauk
WSPÓŁORGANIZATORZY
Politechnika Warszawska
Przemysłowy Instytut Automatyki i Pomiarów
Polskie Stowarzyszenie Pomiarów, Automatyki i Robotyki

ORGANIZATOR

Komitet Automatyki i Robotyki Polskiej Akademii Nauk
Instytut Badań Systemowych Polskiej Akademii Nauk

WSPÓŁORGANIZATORZY

Politechnika Warszawska
Przemysłowy Instytut Automatyki i Pomiarów
Polskie Stowarzyszenie Pomiarów, Automatyki i Robotyki

KOMITET PROGRAMOWY

Przewodniczący	Zdzisław BUBNICKI
Zastępca Przewodniczącego	Roman KULIKOWSKI

CZŁONKOWIE

Stanisław BAŃKA	Michał BIAŁKO
Mikołaj BUSŁOWICZ	Władysław FINDEISEN
Ryszard GESSING	Henryk GÓRECKI
Jakub GUTENBAUM	Jerzy JÓZEFCZYK
Stanisław KACZANOWSKI	Tadeusz KACZOREK
Janusz KACPRZYK	Jerzy KLAMKA
Józef KORBICZ	Zbigniew KOWALSKI
Krzysztof KOZŁOWSKI	Juliusz L. KULIKOWSKI
Krzysztof KUŹMIŃSKI	Kazimierz MALANOWSKI
Krzysztof MALINOWSKI	Wojciech MITKOWSKI
Antoni NIEDERLIŃSKI	Władysław PEŁCZEWSKI
Tadeusz PUCHAŁKA	Leszek RUTKOWSKI
Stanisław SKOCZOWSKI	Roman SŁOWIŃSKI
Jerzy ŚWIĄTEK	Andrzej ŚWIERNIAK
Ryszard TADEUSIEWICZ	Piotr TATJEWSKI
Krzysztof TCHOŃ	Leszek TRYBUS
Jan WĘGLARZ	Andrzej P. WIERZBICKI

KOMITET ORGANIZACYJNY

Przewodniczący	Roman KULIKOWSKI
Zastępcy Przewodniczącego	Janusz KACPRZYK
	Stanisław KACZANOWSKI
	Tadeusz KACZOREK
	Krzysztof MALINOWSKI
Członkowie	Roman OSTROWSKI
	Tadeusz PUCHAŁKA
	Dariusz WAGNER
Sekretarze naukowi	Jan STUDZIŃSKI
	Jan W. OWSIŃSKI

ISBN 83-89475-01-4

Copyright © Instytut Badań Systemowych Polskiej Akademii Nauk
All rights reserved

Druk: ARGRAF, Warszawa

TECHNIKA SYSTEMÓW – DIAGNOSTYKA

PROSTY SYSTEM INDEKSOWANIA I KLASYFIKACJI DOKUMENTÓW W OBRĘBIE OKREŚLONEJ DZIEDZINY†

Jan W. OWSIŃSKI*, Andrzej GUTKIEWICZ**

*Instytut Badań Systemowych Polskiej Akademii Nauk
ul. Newelska 6, 01-447 Warszawa; e-mail: owsinski@ibspan.waw.pl

** Wyższa Szkoła Informatyki Stosowanej i Zarządzania
ul. Newelska 6, 01-447 Warszawa; e-mail: gutkiewi@wsisiz.edu.pl

Streszczenie: Praca opisuje projekt dotyczący prostego systemu indeksowania i klasyfikacji dokumentów należących do określonej, względnie jednolitej dziedziny. W wyniku przeprowadzonej na materiale empirycznym analizy opracowano zestaw narzędzi służących do wskazanego celu. Zestaw ten pozwala na: (1) przeprowadzenie analizy zbioru dokumentów traktowanego jako „zbiór uczący", (2) opracowanie na tej podstawie konkretnego (dziedzinowego) systemu indeksowania dokumentów, (3) opracowanie procedury klasyfikacji nowych dokumentów, (4) ewentualną adaptację zarówno systemu indeksowania, jak i procedury klasyfikacji. Narzędzia stanowiące ten zestaw zostały zrealizowane w postaci interaktywnego oprogramowania. W opisywanych pracach posługiwano się możliwie prostymi i interpretowalnymi metodami, pochodzącymi z klasycznego repertuaru analizy danych, zwłaszcza analizy skupień.

Słowa kluczowe: Analiza tekstów, indeksacja dokumentów, klasyfikacja dokumentów, analiza skupień.

1. WPROWADZENIE

Zagadnienia związane z wyszukiwanie, kategoryzacją i analizą tekstów i innych obiektów medialnych są obecnie jednym z najpopularniejszych tematów badań naukowych. W dziedzinie tej próbuje się często wykorzystywać nowe metodyki, szczególnie związane z innowacjami statystycznymi, metaheurystykami oraz podejściami „miękkimi" (por. np. [2, 3, 4]).

Opisywana praca polegała natomiast na przetestowaniu zestawu prostych technik, wywodzących się głównie z analizy skupień, jako elementów narzędzia przeznaczonego do analizy („uczącego") zbioru dokumentów, otrzymywania ich indeksów i następnie dokonywania klasyfikacji.

Zasadniczym założeniem była przynależność dokumentów ze zbioru „uczącego" do pewnej dość dobrze określonej dziedziny oraz ograniczone rozmiary tego zbioru, co powinno pozwalać na dokładną analizę zbioru „uczącego" i umożliwiać wyciąganie z tej analizy daleko idących i precyzyjnych wniosków, co do ewentualnych indeksów i klasyfikacji.

Tak więc zakładana procedura, dla której tworzono zestaw narzędzi miałaby następującą postać: **(i)** analiza „uczącego" zbioru dokumentów, prowadząca do otrzymania **(ii)** struktury tego zbioru w postaci hierarchii jego podziałów (skupień dokumentów), oraz **(iii)** struktury indeksów dokumentów z tego zbioru; przy założeniu, że zbiór ten jest „wystarczająco" reprezentatywny dla całej rozpatrywanej dziedziny struktury otrzymane w (ii) i (iii) powinny wystarczyć do **(iv)** klasyfikowania dokumentów spoza zbioru uczącego.

W przypadku powodzenia zakładano możliwość zastosowania otrzymanych narzędzi do takich zbiorów dokumentów jak, np., ściągane z internetu artykuły naukowe z pewnej dziedziny, artykuły prasowe dotyczące pewnego zagadnienia, opisy produktów służących do zbliżonego celu, dokumenty prawne itp., ale także np. wiadomości o charakterze spamów.

2. OZNACZENIA

Oznaczymy przez D „uczący" zbiór n dokumentów ponumerowanych j, j=1,..,n; a przez m_j – (całkowitą) liczbę słów występujących w dokumencie j (ewentualnie po usunięciu stop-listy); przez I_j – zbiór tych słów, o indeksach i. Analogicznie, oznaczymy przez m_D liczbę (różnych) słów występujących (ewentualnie po usunięciu stop-listy) we wszystkich dokumentach ze zbioru D; przez I_D – zbiór tych słów; A będzie oznaczeniem podzbiorów zbioru D, uzupełniane ewentualnie dodatkowymi indeksami itp.; odpowiednio do poprzednich oznaczeń mamy więc także m_A i I_A – liczbę i zbiór (różnych) słów występujących w dokumentach ze zbioru A. Dalej, l_{ij} niech będzie liczbą wystąpień słowa i w dokumencie j, oraz analogicznie – l_{iD}, a także l_{iA}.

†Praca była finansowana w zakresie prac metodycznych w ramach grantu KBN dotyczącego projektu europejskiego TRANSCAT, kontrakt EVK1-CT-2002-00124, Decyzja Przewodniczącego KBN Nr 55/E-82/SPB/5.PR UE/DZ 385/2003-2005 z dnia 16 lipca 2003r.

3. PODSTAWOWE ZALEŻNOŚCI

Mając tego rodzaju charakterystyki dokumentów i potencjalnie ich (pod)zbiorów możemy, naturalnie, skorzystać z podstawowych zależności technik wyszukiwania informacji, a w tym przede wszystkim znanego wyrażenia na wagę w_{ij} słowa i w dokumencie j, czyli (por. np. [1]):

$$w_{ij} = f_{ij} f_i^* \qquad (1)$$

gdzie f_{ij} jest względną częstością wystąpień słów i w dokumentach j, $f_{ij} = l_{ij}/\max_i l_{ij}$, zaś $f_i^* = \log_2(n/n_i)$, przy czym n_i to liczba dokumentów z D, w których występuje słowo i. Zależność (1) pozwala nam, teoretycznie, na wyznaczanie wag automatycznie, bez konieczności określania i usuwania stop-listy. Możemy jednakże obawiać się, że dla zbiorów D o niewielkich rozmiarach „statystyka" n/n_i nie będzie wystarczająca do wyeliminowania słów do niej „obiektywnie" należących.

Dlatego też wprowadzono dodatkową grupę pojęć. Do oznaczeń zbiorów słów dodajemy górny indeks v, oznaczający, że mamy zbiór po odjęciu słów występujących w dokumencie(tach) co najwyżej v razy, np. I_D^2 to zbiór (różnych) słów występujących we wszystkich rozpatrywanych dokumentach po usunięciu (stop-listy oraz innych) słów występujących w nich jeden raz lub dwa razy, zaś m_j^3 to liczba słów występujących w dokumencie j-tym po usunięciu (stop-listy oraz innych) słów występujących w nim co najwyżej trzy razy. Ponieważ dla odpowiednio dużego v zbiory I_A^v są puste, więc możemy wprowadzić pojęcie v_j, v_A, v_D, tj. największej liczby wystąpień słów (liczby wystąpień najczęściej występującego słowa) w dokumencie j, zbiorze dokumentów A, i wszystkich dokumentach ze zbioru D; poza tym, oczywiście, wprowadzimy liczbę wystąpień słowa i w dokumencie j: l_{ij}; pojęcie to i oznaczenie można, naturalnie, odpowiednio rozszerzyć na zbiory A oraz D.

Z punktu widzenia analizy zbioru dokumentów zasadniczym pojęciem jest odległość między dokumentami, bądź podobieństwo dokumentów, określone przede wszystkim dla par dokumentów. Dla wprowadzonych poprzednio oznaczeń możemy sformułować następujące dwie definicje odległości między dokumentami

$$d_{jj'} = \Sigma_i |w_{ij} - w_{ij'}| \qquad (2)$$

oraz

$$d_{jj'}^{\,v} = \frac{\sum_{i \in I_j^v \cup I_{j'}^v} \left| l_{ij} - l_{ij'} \right|}{\sum_{i \in I_j^v \cup I_{j'}^v} \max\{l_{ij}, l_{ij'}\}} \qquad (3)$$

przy czym wartości odległości (3) zawarte są w przedziale [0,1]: 0 dla „takich samych" (przy długości reprezentacji określonej przez v) dokumentów, 1 dla „całkowicie różnych" dokumentów.

Oczywiście, jeśli v=0 i $d_{jj'}^{\,v}$=0, to istnieją podstawy do „podejrzewania", że mamy do czynienia faktycznie z takimi samymi dokumentami (podejrzenie to jest uzasadnione także dla, np., v=1 i „małych" wartości $d_{jj'}^{\,v}$, co daje podstawę do „rozmytej" definicji „tożsamości").

Operowanie na całych reprezentacjach dokumentów jest uzasadnione tym, że analizujemy ograniczony zbiór uczący (np. 200 dokumentów). To uzasadnienie rozciąga się na szerszą analizę, obejmującą analizę dla różnych wartości v.

4. STRUKTURA ZBIORU DOKUMENTÓW

Dla odległości określonych w (2) lub (3) możemy przeprowadzić analizę skupień. Jeśli zastosujemy jeden z algorytmów agregacji hierarchicznej (np. najbliższego lub najdalszego sąsiedztwa), to pierwsze agregacje (zwłaszcza dla v=0) powinny dać w wyniku grupy „takich samych" dokumentów. Następnie będą agregowane dokumenty definiujące – razem – pewne, coraz obszerniejsze, „dziedziny". Wskazane byłoby powiązanie tych agregacji z wartościami wskaźnika hierarchii, np. odległości łączonych skupień. Interpretacja grup przy pomocy wskaźnika hierarchii może być wzmocniona, dla definicji (3), wynikami analizy skupień dla poszczególnych v=0, 1, 2, 3,... .

Aspekt interpretacyjny wiąże się bezpośrednio z zadaniem indeksacji dokumentów i ich grup. Algorytmy hierarchiczne wydają się dostarczać naturalnego narzędzia do tego celu: agregacji dokumentów i ich skupień towarzyszyć może agregacja ich indeksów. Zauważmy, że nie wyklucza to nakładania się indeksów dla różnych grup („dziedzin").

Dla celów badania przyjmiemy, że indeksy dokumentów są równoważne wektorom $W_j = \{w_{ij}\}_i$ w przypadku wag (1), bądź L_j^v, w drugim z rozpatrywanych opisów (niezależnie od kwestii reprezentacji takich indeksów przy pomocy odpowiednich wektorów binarnych, czyli indeksów werbalnych – zestawów słów kluczowych). Tak określone indeksy będą podlegały operacjom związanym z agregacją dokumentów i ich zbiorów. Długość indeksu jest, oczywiście, proporcjonalna do precyzji opisu zawartości dokumentu lub zbioru dokumentów. Zakładamy, że otrzymana procedura prowadzić będzie do coraz krótszych – coraz ogólniejszych – indeksów.

Najprostszym przyjętym w pracy sposobem operowania na indeksach jest ich agregacja przy pomocy minimum: $W_{j\cup j'} = \{\min(w_{ij}, w_{ij'})\}_i$ i analogicznie dla L_j^v. W ten sposób, oczywiście, otrzymujemy też, że $I_{j\cup j'} = I_j \cap I_{j'}$. Łatwo zauważyć, że po odpowiedniej liczbie takich agregacji otrzymuje się indeksy puste, bądź to w sensie absolutnym bądź względem pewnego przyjętego progu, eliminującego mniejsze wartości w_{ij} lub f_{ij}. W ten sposób jednak możemy również otrzymać dodatkowe kryterium stopu algorytmu (minimalna długość indeksu). Alternatywa jest posługiwanie się przy określaniu $W_{j\cup j'}$ inną funkcją niż minimum.

5. ANALIZA SKUPIEŃ

W ramach prowadzonych badań testowano stosowalność trzech algorytmów analizy skupień: najbliższego

sąsiedztwa, najdalszego sąsiedztwa oraz algorytmu Owsińskiego, [5].

Ponieważ, niezależnie od wyników w postaci dychotomicznej hierarchii, interesuje nas uwarunkowanie tej hierarchii, poświęcimy nieco miejsca możliwości oceny tego uwarunkowania. Zauważmy przy tym jednak, że zagadnienie, które chcemy rozwiązać, nie dotyczy otrzymania „rozwiązania optymalnego" w sensie podziału, nawet, gdyby takie rozwiązanie było do osiągnięcia przy pomocy znanych metod analizy skupień. To „rozwiązanie optymalne" będzie w tym przypadku jedynie wskazówką, co do zatrzymania algorytmu i ustalenia końcowego (najkrótszego) zestawu indeksów (etykiet węzłów hierarchii).

Zaczniemy od kryterium, mogącym służyć do optymalizacji podziału w analizie skupień, a przynajmniej do porównywania poszczególnych podziałów. Podział P jest zbiorem podzbiorów $A_q \subseteq J$, gdzie $q \in \{1,...,p(P)\} = C(P)$, zaś $J=\{1,..,j,..,n\}$ jest zbiorem indeksów obiektów (obserwacji, w tym przypadku dokumentów), $P=\{A_q\}_q$, przy czym $\cup_q A_q = J$ oraz $A_q \cap A_{q'} = \varnothing \ \forall q,q' \in C(P)$.

Chodzi głównie o możliwość porównywania podziałów o różnej krotności (liczności) p, której to porównywalności nie zapewniają w zasadzie istniejące metody analizy skupień. Powinno to się zarazem łączyć z „dopasowaniem" poszukiwanego kryterium do sposobu, w jaki otrzymaliśmy porównywane podziały (chyba, że wygenerowaliśmy je losowo).

Postulowaną ogólną postacią funkcji kryterium jest (dualnie) albo maksymalizowane $Q_S^D = Q_S + Q^D$, gdzie Q_S oznacza funkcję podobieństwa obiektów należących do tych samych skupień, rozciągniętą na cały podział, zaś Q^D – funkcję odległości między obiektami z różnych skupień, też rozciągniętą na cały podział, albo też minimalizowane $Q_D^S = Q_D + Q^S$, przy definicjach składników dualnych do poprzednich. W dalszym ciągu, bez straty ogólności, będziemy się zajmowali wyłącznie formą maksymalizowaną (Q_S^D).

Postulowana postać funkcji kryterium wymaga jednoczesnego operowania odległościami $d_{jj'}$ oraz bliskościami (podobieństwami) $s_{jj'}$ między obiektami ze zbioru indeksów J. W tym celu należy określić przekształcenie $d(s) = d$ i odwrotne $s(d) = s$, między odległością i bliskością, np. dla wartości normalizowanych $d = 1 - s$ oraz $s = 1 - d$. Analogiczne przekształcenia można zdefiniować dla wartości nie normalizowanych (np. $s = d^{max} + d^{min} - d$). Można także narzucić dodatkowe warunki na te przekształcenia, np. warunek równych średnich dla zbiorów wartości $d_{jj'}$ oraz $s_{jj'}$.

Formalnie rzecz biorąc, jeśli dane jest przekształcenie $d \leftrightarrow s$, to postać funkcji celu można sprowadzić do wyrażenia wyłącznie względem d bądź wyłącznie względem s, jednak dla jasności prezentacji, a także ponieważ taki zabieg ma tylko charakter formalny, pozostaniemy przy wyjściowej postaci Q_S^D.

Najważniejsza przy doborze („projektowaniu") konkretnych postaci Q_S oraz Q^D jest (**i**) sensowność merytoryczna, tj. „prawidłowe" odzwierciedlenie ich treści werbalnej. Poza tym, (**ii**), należy zadbać o „zrównoważenie" (w granicach sensowności merytorycznej) składników Q_S oraz Q^D tak, by otrzymywane najlepsze wyniki nie były w istocie wyznaczone przez skalę lub normalizację przyjętych funkcji. I wreszcie, (**iii**) dobrze jest się przyjrzeć zależności Q_S oraz Q^D od krotności podziału (zaleca się, by jedna wartość rosła wraz z p, a druga malała). Do porównywania jakości podziału wystarczy spełnienie pierwszych dwóch postulatów, trzeci zmierza do projektowania efektywnego algorytmu znajdowania podziałów najlepszych w sensie przyjętej funkcji celu. Pamiętajmy także (**iv**) o postulacie „dopasowania" postaci funkcji celu do metody, przy pomocy której znaleziono porównywane podziały. W sumie, konieczne jest przeprowadzenie choćby pobieżnej analizy wstępnej proponowanej postaci funkcji celu.

Przykładem funkcji celu, spełniającym wszystkie powyższe postulaty (pod warunkiem zastosowania odpowiedniego przekształcenia $d \leftrightarrow s$) jest

$$Q_S^D = Q_S + Q^D = \Sigma_q \Sigma_{j,j' \in A_q} s_{jj'} + \Sigma_{q,q' \in C(P)} \Sigma_{j \in A_q, j' \in A_{q'}} d_{jj'} \qquad (4)$$

czyli sumy wszystkich bliskości i odległości, odpowiednio, wewnątrz skupień i pomiędzy skupieniami. Dla tej postaci można sformułować prosty algorytm sub-optymalizacji.

Innym przykładem, na pierwszy rzut oka sensownym merytorycznie (warunek (i)), może być:

$$Q_S^D = Q_S + Q^D = \Sigma_q s^{q\min} + \Sigma_{q<q'} d^{qq'\min} \qquad (5)$$

gdzie $s^{q\min} = \min_{j,j' \in A_q} s_{jj'}$, zaś $d^{qq'\min} = \min_{j \in A_q, j' \in A_{q'}} d_{jj'}$, czyli sumy najmniejszych podobieństw wewnątrz poszczególnych skupień i najmniejszych odległości pomiędzy poszczególnymi parami skupień. W tym przypadku należy zauważyć, że (i) liczby składników Q_S i Q^D są zasadniczo różne, co pociąga za sobą konsekwencje dotyczące potencjalnych wartości tych dwóch składników, por. poniższa tabela:

Charakter podziału skrajnego	Q_S	Q^D	Q_S^D
$P=\{J\}$: wszystkie obiekty w jednym skupieniu - liczba składników - wartość funkcji	$p=1$ $s^{J\min} =$ $\min_{j,j' \in J} s_{jj'}$	$p(p-1)/2$ $=$ 0 0	1 $s^{J\min}$
$P=J$: wszystkie obiekty osobno - liczba składników - wartość funkcji	$p=n$ ns^{MAX}*	$p(p-1)/2$ $=n(n-1)/2$ $D_J = \Sigma_{j,j' \in J} d_{jj'}$	$n(n+1)/2$ $ns^{MAX} + D_J$

* ponieważ zakładamy $d_{jj'}=0$, więc dla $s_{jj'}$ ustalamy możliwie największą wartość, np. $s_{jj'} = s^{Imax} + s^{Imin} = s^{MAX}$

Łatwo zauważyć, że dla powyższych założeń najlepszym podziałem będzie zawsze $P=J$ i że to właśnie

definicja $s^{q\min}$ dla $A_q=q$, czy też $s_{ji'}$ jest kluczem do sensowności tego sformułowania. Jeśli bowiem założymy, że zbiór bliskości, w którym dokonujemy minimalizacji, żeby otrzymać $s^{q\min}$ dla $A_q = q$ jest w istocie pusty, to możemy przyjąć – wręcz przeciwnie – że dla $A_q = q$ mamy $s^{q\min}=0$ i wówczas optimum podziału nie musi przypadać na $P = J$.

Dla postaci Q_S i Q^D, dla których istnieje „przeciwstawna monotoniczność względem p" (jeden składnik rośnie gdy drugi maleje) warto jest rozważyć zmodyfikowaną postać funkcji celu. I tak, zamiast, jak dotąd, rozważać

$$Q_S^D(P) = Q_S(P) + Q^D(P) \qquad (6)$$

możemy rozważać funkcję

$$Q_S^D(P,r) = (1-r)Q_S(P) + rQ^D(P), \qquad (7)$$

w której $r\in[0,1]$.

Zauważmy, że $Q_S^D(P,1) = Q^D(P)$, zaś $Q_S^D(P,0) = Q_S(P)$, natomiast dla „prawidłowo" sformułowanej funkcji optimum podziału poszukiwać będziemy dla $r= ½$. W ogólności, takie poszukiwanie jest trudne – brak jest numerycznie efektywnych algorytmów – ale dla założenia o „przeciwstawnej monotoniczności" możemy próbować następującego postępowania algorytmicznego (zakładając, że $Q^D(P)$ rośnie, a przynajmniej nie maleje, wraz z $p(P)$, zaś $Q_S(P)$ – odwrotnie):

Ustalmy numer kroku algorytmu $t = 0$ i $r^0 = 1$. Dla tej wartości $r=r^0$ wyznaczmy $P^*(r^t)=P^*(r^0)$. Wiemy już, że $Q_S^D(P,r^0) = Q_S^D(P,1) = Q^D(P)$, a wobec założenia o monotoniczności, maksimum $Q^D(P)$ otrzymujemy dla $p(P^*)=n$, czyli $P^*(r^0)=J$.

Obniżmy teraz wartość r od $r^0=1$ o wartość Δr (czyli $r = r^t - \Delta r$). Mamy zatem

$$Q_S^D(P,r) = \Delta r\, Q_S(P) + (1-\Delta r)Q^D(P),$$

a dla $P^*(1)$: $Q_S^D(P^*(1),r) = \Delta r\cdot 0 + (1-\Delta r)Q^D(J)$, a więc o $\Delta r Q^D(J)$ mniej niż dla r^0. Rozważmy pewien alternatywny podział $P^*(\Delta r)$ powstały z $P^*(1)=J$ przez połączenie dwóch obiektów, dla których $d_{ji'} = d^{J\min}$. Różnica wartości funkcji celu to

$$Q_S^D(P^*(1),r) - Q_S^D(P^*(\Delta r),r) =$$

$$= (1-\Delta r)Q^D(J) - \Delta r\, Q_S(P^*(\Delta r)) - (1-\Delta r)Q^D(P^*(\Delta r)) =$$

$$= (1-\Delta r)(Q^D(J)-Q^D(P^*(\Delta r))) - \Delta r\, Q_S(P^*(\Delta r)).$$

Pierwszy z tych dwóch składników jest dodatni (założenie o monotoniczności), ale jego „udział" $(1-\Delta r)$ maleje wraz z Δr, podczas gdy drugi, ujemny składnik ma nie tylko rosnącą co wartości absolutnej wartość (założenie o monotoniczności), ale i jego „udział" (Δr) też rośnie wraz z Δr. Istnieje więc taka wartość, $r^{t+1} = 1- \Delta r$ (ogólniej: $r^t-\Delta r$) dla której

$$(1-\Delta r)(Q^D(J)-Q^D(P^*(\Delta r))) - \Delta r\, Q_S(P^*(\Delta r)) = 0,$$

czyli $P^*(1)$ (bądź $P^*(r^t)$) przestaje być optymalne. Wartość tę wyznaczymy z powyższego równania:

$$r^1=1-\Delta r=Q_S(P^*(\Delta r)) / (Q^D(J)-Q^D(P^*(\Delta r))+Q_S(P^*(\Delta r)))$$

Podstawiamy $t = t +1$ i odpowiadające mu r^t, otrzymane z powyższego wyrażenia, jak również $P^*(\Delta r)=P(r^t)$. Powtarzamy operację aż do momentu, w którym $r^{t\#} \geq ½ \geq r^{t\#+1}$ i przyjmujemy jako rozwiązanie podział $P(r^{t\#})$.

Tak więc otrzymaliśmy, po pierwsze, ogólną funkcję kryterium, pozwalającą na porównywanie wyników otrzymanych przy pomocy różnych metod analizy skupień, a po drugie, dla pewnej klasy takich funkcji, efektywny algorytm suboptymalizacji.

6. WYNIKI BADAŃ EMPIRYCZNYCH

Przedstawimy obecnie dość skrótową ilustrację wyników badań empirycznych, prowadzonych, jak wspomniano, na zbiorze o charakterze „nigeryjskiego spamu", mających na celu sprawdzenie przydatności poszczególnych proponowanych rozwiązań technicznych.

Poniżej podano przykładowe statystyki odległości liczonych według wzoru (3) dla v=0,...,5.

v	Całość tekstu	Bez stop-listy
v=0	Średnia: 0,7637 Odch. stand.: 0,063 Min: 0 Max: 0,9138	0,8514 0,0591 0 0,9518
v=1	Średnia: 0,6625 Odch. stand.: 0,075 Min: 0 Max: 0,8869	0,7833 0,0683 0 0,9303
v=2	Średnia: 0,5978 Odch. stand.: 0,087 Min: 0 Max: 0,8802	0,7426 0,080 0 0,9523
v=3	Średnia: 0,5530 Odch. stand.: 0,098 Min: 0 Max:0,8747	0,7132 0,095 0 1,00
v=4	Średnia: 0,5175 Odch. stand.: 0,107 Min: 0 Max: 0,8679	-*
v=5	Średnia: 0,4902 Odch. stand.: 0,114 Min: 0 Max: 0,8731	-*

* po odjęciu stop-listy niemożliwe było policzenie tych odległości dla zbioru dokumentów

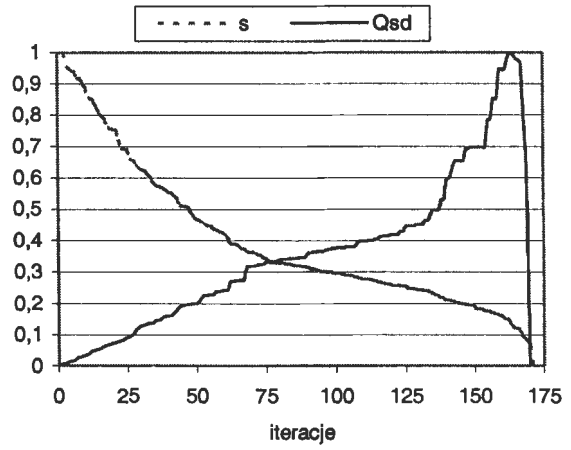
Zbiór uczący zawiera 171 dokumentów, o długościach nie przekraczających 150-200 słów. Dokumenty były analizowane po odrzuceniu nagłówków poczty elektronicznej, z usuwaniem stop-listy i bez tej operacji.

Tabela odległości pokazuje wyraźnie kilka ważnych zjawisk, a przede wszystkim malenie średniej wartości

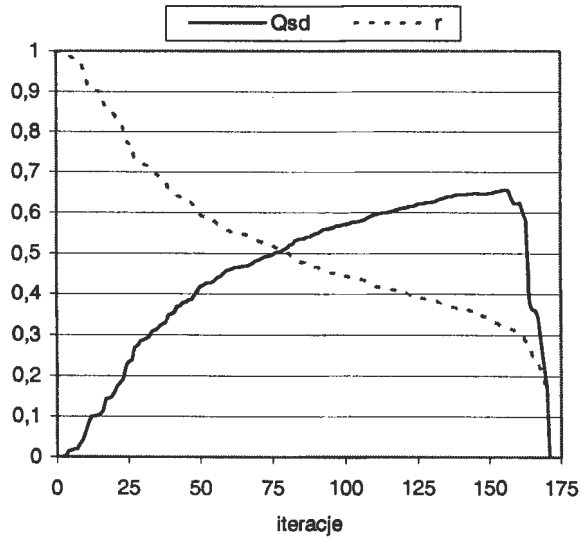
odległości przy wzroście v, co jest związane z jednolitością tematyczną zbioru uczącego.

Rys. 1 pokazuje przebieg wartości Q_S^D według (5) dla algorytmu najdalszego sąsiada („complete linkage"), wraz ze znormalizowanymi wartościami odległości łączonych skupień. Natomiast rys. 2 pokazuje przebieg Q_S^D według (5) dla algorytmu Owsińskiego oraz odpowiadające mu wartości wskaźnika r.

Rys. 1. Przebieg wartości Q_S^D i znormalizowanych odległości łączonych skupień dla algorytmu najdalszego sąsiedztwa



Rys. 2. Przebieg wartości Q_S^D i r dla algorytmu Owsińskiego



Podkreślmy raz jeszcze, że w naszym zagadnieniu nie jest najważniejsze osiągnięcie rozwiązania optymalnego, ale otrzymanie możliwie najlepszej struktury hierarchicznej. Biorąc pod uwagę fakt, że do momentu otrzymania podziałów bliskich „optymalnemu" łączone są coraz odleglejsze skupienia można jedynie założyć, że dalsze agregacje, poza okolicę podziałów „optymalnych", nie mają sensu. Jednocześnie, wspomniane już zjawisko otrzymywania pustych indeksów następowało po wykonaniu wystarczającej liczby kroków (np. 10-15), aby móc utworzyć sensowne „częściowe dziedziny" badanego zbioru dokumentów. Podana poniżej tabela zawiera dość charakterystyczną ilustrację – drogę w drzewie hierarchii rozpoczynającą się od jednego liścia-dokumentu i kończącą na korzeniu zawierającym wszystkie dokumenty ze zbioru.

Nr iteracji, t; n=171	Liczności łączonych skupień		Średnia liczność skupień
13	**1**	2	1,08
27	**3**	1	1,19
51	**4**	1	1,43
125	**5**	3	3,72
144	**8**	13	6,33
152	**21**	24	9,00
157	**45**	16	12,21
162	**61**	3	19,00
164	**64**	19	24,43
165	**83**	43	28,50
169	**126**	43	85,50
170	***169***	2	170,00

Wytłuszczono liczności skupień zawierających śledzony dokument (liść); podkreślono liczności skupień nie zawierających słów znaczących w indeksie; kursywą oznaczono skupienia o pustych indeksach. Po iteracji t dla liczby dokumentów n średnia liczba dokumentów w skupieniu wynosi $n/(n-t)$.

Łatwo zauważyć, że istotnie „nieznaczące" skupienia zaczynają dominować dopiero dla dość wysokich numerów iteracji, powyżej 150 dla tego przykładu, natomiast algorytm pozwala na otrzymywanie skupień dokumentów „znaczących" o licznościach do około 10 i powyżej, co jest wystarczające z punktu widzenia przydatności praktycznej.

Zaznaczmy, że opisane tutaj elementy metodyczne zostały zaimplementowane w postaci oprogramowania, które obejmuje, m.in., także możliwość ustalania tzw. stop-listy (listy słów nie niosących informacji merytorycznej) przez użytkownika. W tym celu użytkownikowi przedstawiana jest progresywnie lista słów występujących w zbiorze dokumentów, według wartości W_j bądź L_j^0, przy czym, jako dodatkowa informacja, podawane są wariancje $V(w_{ij})$ bądź $V(f_{ij})$ w zbiorze indeksów dokumentów, z sugestią, że niższe wariancje wskazują (dla podobnych wartości W_j bądź L_j^0) słowa o niższej zawartości informacyjnej. Ten element podejścia nie został jednak zautomatyzowany.

7. WNIOSKI I PODSUMOWANIE

Przedstawiony zestaw prostych technik, wywodzących się z podstaw analizy danych, okazał się dość skuteczny w rozwiązywaniu postawionego zadania, czyli konstrukcji prostego systemu analizy i klasyfikacji dokumentów z pewnego ograniczonego zbioru. Istnieje, naturalnie, konieczność odpowiedniego doboru szeregu parametrów metodyki, a więc np. transformacji $s \leftrightarrow d$, wagi składnika związanego z bliskościami i odległościami, warunku stopu algorytmu itp., co może być przedmiotem dalszych prac metodycznych, ale także

może być dobierane empirycznie dla rozważanych konkretnych zbiorów dokumentów uczących.

A SIMPLE SYSTEM FOR DOCUMENT CATEGORISATION

Abstract: The paper presents the methodological basis for and the empirical study leading to the development of a simple system of document analysis and categorisation. The study aimed at limited document repositories containing documents from a well-defined domain (scientific articles from a narrow discipline, homogeneous product descriptions etc.). One such repository is treated as a "training set" for future classification and labeling. The empirical part of the study, based on the repository of the "Nigerian spam", allowed to test several technical alternatives, dealing with document description, distances between documents, clustering algorithms used, and the manner of producing joint indexes. The results show that a set of simple, well-known techniques can serve the purpose indicated sufficiently well, provided that judicious choices for particular techniques and parameters are made.

[1] Baeza-Yates R., Ribeiro-Neto B. (1999) *Modern Information Retrieval.* Addison-Wesley, Harlow.

[2] Furnas G. F. (1994) High dimensional representations and information retrieval. W: E. Diday et al., eds., New Approaches in Classification and Data Analysis. Springer-Verlag, Berlin-Heidelberg-New York, 559-568.

[3] Kinderman J, Edda L. (2002) Classification of texts with support vector machines: an examination of the efficiency of kernels and data transformations. W: W. Gaul, G. Ritter, eds., Classification, Automation and New Media. Springer Verlag, Berlin-Heidelberg-New York, 189-198.

[4] Mehler A. (2002) Text mining with the help of cohesion trees. W: W. Gaul, G. Ritter, eds., Classification, Automation and New Media. Springer Verlag, Berlin-Heidelberg-New York, 199-206.

[5] Owsiński J. W. (1990) On a naturally indexed quick clustering method with a global objective function. *Applied Stochastic Models and Data Analysis*, 6, 157-171.

Instytut Badań Systemowych
Polskiej Akademii Nauk

ISBN 83-89475-01-4