POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ROZWÓJ I ZASTOSOWANIA TECHNOLOGII I SYSTEMÓW INFORMATYCZNYCH

pod redakcją:
Jana Studzińskiego
Ludosława Drelichowskiego
Olgierda Hryniewicza

ROZWÓJ I ZASTOSOWANIA TECHNOLOGII
I SYSTEMÓW INFORMATYCZNYCH

Polska Akademia Nauk • Instytut Badań Systemowych

Seria: BADANIA SYSTEMOWE
tom 28

Redaktor naukowy:

Prof. dr hab. Jakub Gutenbaum

Warszawa 2001

ROZWÓJ I ZASTOSOWANIA TECHNOLOGII I SYSTEMÓW INFORMATYCZNYCH

pod redakcją
Jana Studzińskiego, Ludosława Drelichowskiego
i Olgierda Hryniewicza

Wydano z wykorzystaniem dotacji KOMITETU BADAŃ NAUKOWYCH

Książka zawiera wybór artykułów poświęconych omówieniu aktualnego stanu badań w kraju w zakresie rozwoju technologii, modeli i systemów informatycznych oraz ich zastosowań w różnych dziedzinach gospodarki narodowej. Wyodrębnioną grupę stanowią artykuły aplikacyjne omawiające wyniki projektów badawczych i celowych KBN.

Recenzenci artykułów:

Dr hab. inż. Ryszard Budziński, prof. US
Prof. dr hab. inż. Janusz Kacprzyk
Dr hab. Adam Kopiński, prof. AE we Wrocławiu
Doc dr hab. inż. Marek Libura
Prof. dr hab. inż. Andrzej Straszak

ISBN 83-85847-59-6
ISSN 0208-8028

Rozdział 5

Modele i systemy wspomagania decyzji w zarządzaniu i technice

SYSTEM AUTOMATYCZNEGO ROZPOZNAWANIA KATEGORII TEMATYCZNYCH DOKUMENTÓW INTERNETOWYCH

Klaudia A. Ławcewicz[#1] *i Sławomir Zadrożny*[#,&]
[#] *Wyższa Szkoła Informatyki Stosowanej i Zarządzania*
[&] *Instytut Badań Systemowych PAN, Warszawa*

The concept and implementation of the software system for the automatic recognition of the topic of an Internet document is proposed. In the training mode the user provides the system with a list of the topics and sets of documents representing each topic (supervised learning). In the recognition mode the system automatically classifies previously unseen document to a topic category. A simple learning algorithm is devised and implemented. The results of the classification are presented to the user in the form of a set of linguistic terms. Some new measures of the correctness of the classification are proposed. The implemented system processes documents in several popular Internet-related formats.

1. Wstęp

Internet jest niewyczerpanym źródłem informacji. Stały rozrost sieci, nieunikniony i pożądany, niesie ze sobą jednak różnorakie wyzwania. Z jednej strony, praktycznie każdy może znaleźć tu dokumenty na interesujący go temat. Z drugiej strony, ich odnalezienie staje się coraz trudniejsze. Z tego względu, jednym z ważnych zagadnień jest określanie tematyki dokumentów dostępnych w Internecie. Najwłaściwszym rozwiązaniem wydawać się może osobiste oznaczanie zawartości dokumentu przez jego autora. Nie jest to jednak rozwiązanie w pełni satysfakcjonujące. Po pierwsze, nie można oczekiwać, że wszystkie dokumenty będą opatrzone stosownym opisem autora. Po drugie, klasyfikacja podana przez autora może nie być zgodna z postrzeganiem zawartości dokumentu przez jego odbiorcę. Innym rozwiązaniem problemu może być opisywanie (klasyfikowanie) dokumentów dostępnych w Internecie przez innego człowieka (grupę osób) posługujących się w tym celu pewnym spójnym zestawem kategorii. Przykładem takiego systemu może być serwis Yahoo. Podejście to częściowo rozwiązuje obydwie niedogodności występujące w pierwszym wariancie. Jego połowiczność zasadza się na tym, że, po pierwsze, nadal odbiorca musi polegać na cudzym systemie klasyfikacji (kategoriach tematycznych) choć tym razem bardziej zobiektywizowanym. Po drugie, liczba dokumentów, które mogą być sklasyfikowane przez grupę osób jest z natury rzeczy ograniczona, niewspółmierna do bogactwa zasobów Internetu.

[1] studentka studiów magisterskich na Wydziale Informatyki

Powyższe rozważania prowadzą do wniosku, że jedynie automatyczna bądź półautomatyczna klasyfikacja dokumentów jest w stanie rozwiązać postawione tu zadanie. Wkraczamy tym samym w obszar zagadnień wchodzących w zakres zainteresowań wielu gałęzi nauk matematyczno-informatycznych. Wymienić tu można *rozpoznawanie wzorców* (ang. *pattern recognition*) - w szczególności zagadnienia związane z konstruowaniem *klasyfikatorów* - czy też ogólnie *wyszukiwaniem informacji* (ang. *information retrieval*).

W niniejszej pracy opisujemy projekt i implementację, zaproponowanego przez K. A. Ławcewicz (2001), systemu TCAT, który można wykorzystać do rozwiązania postawionego wcześniej zadania. W p. 2 opisujemy dokładniej warunki rozwiązywanego zadania. W p.3 podajemy algorytmy, które zastosowano w systemie TCAT. Kolejny punkt zawiera szczegóły implementacji systemu w środowisku Internetu. Wreszcie w p. 5 omawiamy przykład obliczeniowy i uzyskane wyniki. Na zakończenie opisujemy planowane dalsze prace nad systemem.

2. Przeznaczenie i zasada działania systemu TCAT

System TCAT automatycznie klasyfikuje dokumenty tekstowe zgodnie z ustalonymi wcześniej kategoriami tematycznymi. System działa w dwóch trybach. W trybie uczenia się analizuje on zadane przez użytkownika dokumenty reprezentujące poszczególne kategorie tematyczne. Mamy tu więc do czynienia z *uczeniem się z nauczycielem* (ang. *supervised learning*). W trybie rozpoznawania system dokonuje automatycznej klasyfikacji zadanego dokumentu do jednej z wcześniej "wyuczonych" kategorii tematycznych. Z tego punktu widzenia system można potraktować jako przykład *klasyfikatora*. Charakterystyczne dla przyjętego rozwiązania cechy to:

- praktyczna, uniwersalna reprezentacja dokumentów za pomocą *tokenów*- ciągów znaków o ustalonej długości
- prosty algorytm uczenia się
- prezentacja wyników z użyciem elementów logiki rozmytej

W trybie uczenia system wyodrębnia z analizowanych dokumentów tokeny, to jest kolejne 5 lub 10 elementowe ciągi znaków i zapamiętuje w bazie danych częstość ich występowania. Na zakończenie fazy uczenia dla każdego tokena, na podstawie częstości jego występowania w dokumentach związanych z każdą z kategorii, obliczany jest jego *stopień przynależności do poszczególnych kategorii*. Wyniki tych obliczeń również zapisywane są w bazie danych.

W trybie rozpoznawania, dokument także dzielony jest na tokeny, z tym że uwzględniane są jedynie wcześniej (w fazie uczenia) napotkane tokeny. Następnie, na podstawie stopni przynależności tokenów do poszczególnych kategorii (wyliczonych w trybie uczenia) obliczany jest *stopień przynależności dokumentu do poszczególnych kategorii*. Ze względu na możliwą niejednoznaczność klasyfikacji jej wyniki prezentowane są w formie wyrażeń lingwistycznych. Dokładniej, stopień przynależności do kategorii traktowany jest jako zmienna lingwistyczna. W rezultacie,

użytkownik otrzymuje słowny opis stopnia przynależności dokumentu do poszczególnych kategorii tematycznych.

Stosowane algorytmy inspirowane są pojęciami teorii zbiorów rozmytych. Ich szczegóły podajemy w następnym punkcie.

System przetwarza dokumenty zapisane w typowych formatach stosowanych w Internecie. W aktualnej wersji system może rozpoznawać pojedyncze dokumenty umieszczone na lokalnym komputerze bądź w Internecie, wskazane poprzez podanie ich URL-a (ang. *Uniform Resource Locator*).

3. Algorytm uczenia się i rozpoznawania

Niech $D=\{d_j\}_{j\in[1,M]}$ i $T=\{t_i\}_{i\in[1,N]}$ oznaczają, odpowiednio, zbiór wszystkich rozważanych dokumentów i zbiór tokenów wyodrębnionych z dokumentów w trybie uczenia się systemu. Każdy dokument reprezentowany jest jako wektor $d_i=(d_{i1},...,d_{iN})$, gdzie d_{ij} oznacza liczbę wystąpień tokena t_j w dokumencie d_i. Otrzymujemy w ten sposób klasyczną, wektorową reprezentację klasyfikowanych obiektów zakładaną zwykle w literaturze dotyczącej konstruowania i zastosowania klasyfikatorów. Tradycyjne narzędzie stanowią tu metody statystyczne, jednak również zastosowania teorii zbiorów rozmytych rozwijają się w tej dziedzinie bardzo burzliwie. Obydwa nurty badań związanych z klasyfikacją są interesująco przedstawione w Kuncheva (2000). Podejście zastosowane przy konstrukcji systemu TCAT jest w pewnym stopniu rozwiązaniem hybrydowym łączącym pewne elementy metod statystycznych i rozmytych.

Podstawowym rezultatem działania systemu TCAT w trybie uczenia się jest wyliczenie dla każdego tokena jego *stopni przynależności do poszczególnych kategorii tematycznych*. Stopień ten jest wskaźnikiem na ile dany token jest charakterystyczny dla ustalonej kategorii. Zakładamy następujące własności tego wskaźnika:

1) jest wprost proporcjonalny do tego, jak często dany token występuje w dokumentach należących do danej kategorii
2) jest odwrotnie proporcjonalny do tego jak często dany token występuje w dokumentach dotyczących innych kategorii
3) zależność pierwsza jest silniejsza niż druga,

Wprowadźmy następujące oznaczenia:

SP_k^t - wskaźnik stopnia przynależności tokena t do kategorii k,

K - zbiór wszystkich kategorii,

n^t, n_k^t - liczba wystąpień tokena t, odpowiednio, we wszystkich dokumentach użytych w fazie uczenia systemu i w dokumentach należących do kategorii k.

$\Delta_+^t = \frac{1}{n^t}$ oraz $\Delta_-^t = \frac{1}{4n^t}$ - stałe pomocnicze

Niech w_k^t oznacza pomocniczy wskaźnik "premiujący" token t proporcjonalnie do liczby jego wystąpień w dokumentach należących do kategorii k i "karzący" token t proporcjonalnie do liczby jego wystąpień w dokumentach należących do innych kategorii

$$w_k^t = v + \left(n_k^t \cdot \Delta_+^t\right) - \sum_{m \in K \setminus \{k\}} \left(n_m^t \cdot \Delta_-^t\right) \tag{1}$$

gdzie v stanowi wartość początkową współczynnika w_k^t (standardowo $v = 0.1$).

Można łatwo wykazać, że tak zdefiniowany wskaźnik w_k^t spełnia warunki 1)-3). Dodatkowo chcemy, żeby wartości poszukiwanego wskaźnika należały do przedziału [0,1]. W tym celu zastosujemy następujące przekształcenie, które określa ostateczny wzór na poszukiwany wskaźnik SP_k^t:

$$SP_k^t = \begin{cases} 0, & \text{dla } w_k^t \le 0 \\ -{w_k^t}^3 + {w_k^t}^2 + w_k^t, & \text{dla } 0 < w_k^t < 1 \\ 1, & \text{dla } w_k^t > 1 \end{cases} \tag{2}$$

W trybie rozpoznawania system wylicza dla podanego dokumentu d jego *stopień przynależności do każdej z kategorii k*, SP_d^k. Podobnie jak w trybie uczenia się system wyodrębnia z dokumentu d występujące w nim tokeny (tym razem uwzględniamy jedynie tokeny wcześniej napotkane w trybie uczenia się). Tym razem dokument utożsamiamy jest ze zbiorem zawartych w nim tokenów: $d=\{t_i\}$. Na podstawie obliczonych w fazie uczenia stopni przynależności tych tokenów do poszczególnych kategorii obliczany jest stopień przynależności dokumentu d do poszczególnych kategorii k, SP_d^k, według następującego wzoru:

$$SP_k^d = \frac{\sum_{t \in d} SP_k^t}{n_d} \tag{3}$$

gdzie n_d i SP_k^t oznaczają, odpowiednio, liczbę tokenów reprezentujących dokument d, i -stopień przynależności tokena t do kategorii k (obliczony w fazie uczenia)

Wyliczone wartości wskaźników SP_k^d dla ustalonego d i każdej kategorii k stanowią podstawę do zakłasyfikowania analizowanego dokumentu do którejś z kategorii. Naturalnym jest wskazanie tej kategorii dla której SP_k^d przyjmuje war-

tość maksymalną. Może się jednak zdarzyć, że wartości tego wskaźnika dla różnych kategorii będą bardzo bliskie i trudno będzie zdecydować jednoznacznie, która kategoria jest właściwa. Użytkownik jest więc informowany w jakim stopniu, według systemu, dokument należy do poszczególnych kategorii. System nie prezentuje jednak użytkownikowi surowych danych liczbowych, lecz przedstawia je w formie bardziej przyjaznej posługując się pojęciem zmiennej lingwistycznej (patrz np. Kacprzyk (1986)).

Zmienna lingwistyczna jest to zmienna, której wartościami nie są liczby, lecz wyrażenia w języku naturalnym, utożsamiane najczęściej w sensie semantycznym z określonymi zbiorami (liczbami) rozmytymi. Określa się ją jako uporządkowaną piątkę (H, $T(H)$, U, G, M), gdzie H jest nazwą zmiennej; $T(H)$ zbiorem wartości zmiennej (terminów lingwistycznych); $U = \{u\}$ określa obszar rozważań, do którego odnosi się dana zmienna lingwistyczna (rozmyte zbiory określone na U stanowią interpretację poszczególnych terminów lingwistycznych z $T(H)$), G jest regułą generującą wartości danej zmiennej lingwistycznej (jeśli $T(H)$ jest skończone to G może sprowadzać się do prostego wyliczenia terminów lingwistycznych); M jest regułą semantyczną przypisująca każdej wartości $l \in T(H)$ jej znaczenie $M(l) \subsetneq U$. Przykładowo, traktując *wiek* jako zmienną lingwistyczną można przyjąć: T("wiek")={"bardzo młody", "młody", "w średnim wieku", "stary", "bardzo stary"}, U=[1,100], M przypisuje poszczególnym elementom T("wiek") liczby rozmyte określone na przedziale [0,100] intuicyjnie odpowiadające poszczególnym określeniom wieku. Przykładowo dla terminu "młody" właściwa może być *trapezoidalna liczba rozmyta* (0;0;25;35). Taką trapezoidalną liczbę rozmytą, zdefiniowaną za pomocą czterech parametrów (tu: 0,0,25,35) należy rozumieć następująco: liczby z przedziału określonego przez drugi i trzeci parametr (tu: [0,25]) należą do definiowanego zbioru rozmytego w stopniu 1.0; liczby mniejsze od pierwszego parametru (tu: 0) oraz większe od parametru czwartego (tu: 35) należą do tego zbioru w stopniu 0; pozostałe liczby należą do zbioru do pewnego stopnia (z przedziału [0,1]) określonego przez funkcje liniową zachowującą odpowiednie warunki brzegowe.

Traktując przynależność dokumentu do danej kategorii tematycznej jako zmienną lingwistyczną otrzymujemy następującą interpretację poszczególnych składowych definicji zmiennej lingwistycznej: zmienną lingwistyczną nazwiemy „związane z kategorią k" (H); obszarem rozważań jest $U = [0,1]$ (czyli zakres możliwych wartości wskaźnika SP_k^d); jako zbiór terminów lingwistycznych można przyjąć

$T(H)$ = {"nie", "lekko", "średnio", "mocno", "bardzo mocno"}. (4)

Reguła semantyczna M przypisuje poszczególnym terminom lingwistycznym liczby rozmyte określone na przedziale możliwych wartości wskaźnika SP_k^d. Przykładowo, termin "bardzo mocno" można reprezentować za pomocą trapezoidalnej liczby rozmytej (0.85,0.95,1.0,1.0).

Stopnie przynależności danego dokumentu do poszczególnych kategorii traktujemy więc jako zmienne lingwistyczne. Pozostaje określić, jak wybieramy

termin lingwistyczny do reprezentowania wyliczonej wartości wskaźnika SP_k^d. Oczywistym rozwiązaniem jest tu przyjęcie takiego terminu lingwistycznego, dla którego wyliczona wartość wskaźnika należy do zbioru (liczby) rozmytej reprezentującej ten termin w stopniu maksymalnym. Zapiszemy to następująco:

$$SP_k^d = u \rightarrow l :- l = \arg\max_l \mu_{M(l)}(u) \tag{5}$$

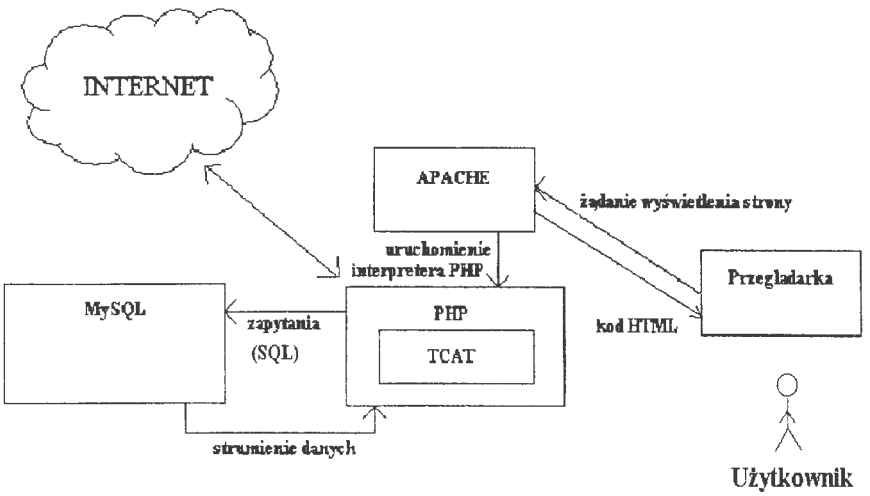
czyli wartości u naszego wskaźnika SP_k^d przypisujemy termin lingwistyczny l, taki że u należy w maksymalnym stopniu do zbioru rozmytego $M(l)$ określającego znaczenie tego terminu.

Dzięki przyjętej definicji wskaźnika SP_k^t token często występujący w dokumentach należących do różnych kategorii uzyska niską wartość stopnia przynależności do wszystkich kategorii. W konsekwencji nie będzie on miał wpływu na działanie systemu w trybie rozpoznawania. Uzyskujemy w ten sposób efekt zbliżony do eliminacji tak zwanych '*stop*' *words* (patrz, np., van Rijsbergen (1979)). Z kolei, token występujący tylko w dokumentach jednej kategorii uzyska wysoki stopień przynależności do tej kategorii. Intuicyjnie, jego wystąpienie w rozpoznawanym dokumencie silnie wskazuje na przynależność dokumentu do tej kategorii. Intuicję tę formalizuje wzór (3).

4. Implementacja systemu

System TCAT jest systemem sensu stricte internetowym. Dokumenty które przetwarza pobierane są zazwyczaj z Internetu. Sam system jest zaimplementowany jako aplikacja internetowa działająca w środowisku usługi WWW. System może być postrzegany w szerszym sensie jako zestaw współpracujących ze sobą: modułów napisanych w PHP, serwera WWW (Apache-a) wraz z interpreterem PHP, serwera bazy danych MySQL oraz przeglądarki WWW. W węższym sensie utożsamiamy TCAT z pierwszym z wymienionych elementów, to jest zestawem napisanych w PHP modułów realizujących zasadnicze funkcje systemu. W dalszym ciągu pomijamy to rozróżnienie jako że kontekst zwykle wyraźnie rozstrzyga, który sposób rozumienia systemu TCAT jest stosowany.

Sposób funkcjonowania i współdziałania poszczególnych elementów systemu TCAT, zilustrowany na Rys. 1, można najkrócej opisać następująco. Po zgłoszeniu przez klienta żądania strony wchodzącej w skład aplikacji TCAT serwer WWW (Apache) uruchamia interpreter PHP. Odpowiedni moduł PHP, stanowiący część systemu TCAT, pobierając potrzebne dane z bazy i Internetu (dokument w fazie uczenia bądź rozpoznawania) generuje kod HTML uzupełniony funkcjami JavaScript. Komunikacja z bazą MySQL w systemie TCAT odbywa się przy pomocy standardowego modułu PHP udostępniającego interfejs do współpracy z tą bazą danych. Kod HTML jest interpretowany przez przeglądarkę, która w rezultacie wyświetla poszczególne ekrany składające się na interfejs użytkownika.



Rys. 1. *Model współdziałania poszczególnych elementów systemu TCAT.*

Zasady interakcji z systemem w obydwu trybach są proste. Interfejs użytkownika składa się z połączonych ze sobą logicznie stron WWW, które zawierają proste formularze.

Przetwarzanie dokumentów składa się z kolejnych kroków. Najpierw zawartość pliku pobierana jest z danej lokalizacji (dysku lokalnego lub wprost z Internetu) do bufora. Następnie z pobranego tekstu usuwane są wszystkie elementy kodu HTML, DHTML, XHTML, XML, PHP oraz ASP. Realizowane jest to na zasadzie prostego skanera leksykalnego, który pobiera kolejne znaki z łańcucha i analizuje je, usuwając fragmenty, które nie niosą ze sobą treści znaczeniowej. Operacja ta nie ogranicza się do mechanicznego usuwania całych znaczników HTML-owych czy wstawek ze skryptami. Ich zawartość jest również analizowana i odzyskiwane są łańcuchy znaków, które mogą być użyteczne do klasyfikowania dokumentów Łańcuchy te, jeśli nie odpowiadają słowom kluczowym (np. nbsp, IMG) dołączane są do bufora i podlegają dalszej analizie. Przykładowo, uwzględniane są nazwy zmiennych oraz teksty zastępcze dla obrazów (to jest, wartości atrybutu ALT używane w znacznikach HTML-owych IMG). Podczas analizy leksykalnej ustalane są również adresy dokumentów powiązanych z danym dokumentem (linki, części składowe ramek) i tekst znaczeniowy pozyskiwany jest również z tych dokumentów. Podczas dalszej analizy leksykalnej usuwane są również wszystkie znaki nie będące literami i spacjami. Następnie wszystkie spacje zostają zamienione na znak podkreślenia '_', a wszystkie duże litery na małe. Kodowanie wszystkich liter właściwych dla polskiego alfabetu: ą, ę, ł, ń, ó, ś, ź, ż ujednolicone zostaje do formatu ISO 8859-2. Tak przetworzony tekst przekazywany jest do kolejnego etapu analizy: uczenia się bądź rozpoznawania znaczenia tekstu.

Baza danych systemu TCAT służy głownie do przechowywania opisu poszczególnych tokenów wyodrębnionych z dokumentów wskazanych systemowi w

fazie uczenia. Najważniejszym elementem opisu tokena są wyliczone przez system stopnie jego przynależności do poszczególnych, zadanych kategorii.

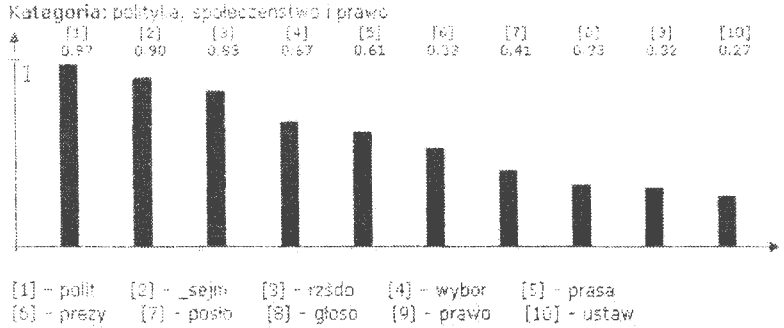
5. Przykład obliczeniowy

Działanie systemu TCAT przetestowano na zbiorze dokumentów tekstowych reprezentujących sześć kategorii tematycznych. Wybór zakresu tematycznego podyktowany był popularnością danych kategorii w zasobach czołowych polskich portali internetowych. Na tej podstawie wybrano następujące kategorie tematyczne:

- przyroda i ekologia,
- ekonomia i gospodarka,
- film i kino,
- komputery i Internet,
- motoryzacja,
- polityka, społeczeństwo i prawo.

W trybie uczenia systemu TCAT użyto po 20 dokumentów z każdej kategorii. Wyniki ilustruje Rys. 2, na którym pokazano wyliczone stopnie przynależności wybranych tokenów do kategorii "polityka, społeczeństwo i prawo".

Dokumenty wykorzystane do przetestowania skuteczności rozpoznawania tematyki tekstów przez system TCAT pobrane zostały ze stron Wielkiej Internetowej Encyklopedii Multimedialnej znajdującej się pod adresem http://wiem.onet.pl/wiem/. Użyto po 10 dokumentów z każdej kategorii. Każdy testowany dokument jest opisem pewnego hasła z tej encyklopedii. Przy wyborze haseł/dokumentów wykorzystano wyszukiwarkę udostępnioną przez encyklopedię. Hasła te zostały wskazane przez przeglądarkę jako "najlepiej" odpowiadające opisom poszczególnych wybranych wcześniej sześciu kategorii tematycznych.



Rys. 2. *Stopnie przynależności tokenów do wybranej kategorii*

Oceniając otrzymane wyniki musimy przyjąć pewną formą ich "defuzyfikacji". Jak to wcześniej opisano, wynikiem działania systemu TCAT w trybie rozpoznawania jest określenie stopnia przynależności zadanego dokumentu do poszczególnych (w tym wypadku sześciu) kategorii tematycznych. Z drugiej strony, dyspo-

nujemy z założenia dokładną informacją o faktycznej przynależności każdego z testowych dokumentów do jednej i tylko jednej kategorii tematycznej. Interesujące może być zadanie, w którym zakładamy, że każdy z dokumentów faktycznie reprezentuje, w różnym stopniu, kilka kategorii tematycznych. Należałoby wtedy zmienić zarówno algorytm uczenia się systemu jak i przyjęte wskaźniki jakości. W niniejszej pracy zakładamy jednak jednoznaczne przypisanie dokumentu do jednej kategorii. Oceniając czy dana rozmyta odpowiedź jest poprawna zastosowaliśmy dwa podejścia. W pierwszym z nich uznajemy, że rozmyta odpowiedź jest poprawna jeśli faktyczna kategoria dokumentu znajduje się wśród kategorii dla których TCAT wyliczył najwyższy stopień przynależności tego dokumentu. W drugim podejściu wymagamy więcej: żądamy aby faktyczna kategoria została wskazana jednoznacznie. Formalnie zapisujemy to następująco:

Podejście I (prosta poprawność)

$$P1 = 100 * \frac{\sum_{i=1}^{M} \phi(d_i)}{M},$$

gdzie M jest licznością zbioru $D=\{d_i\}$ dokumentów użytych do testowania,

$$\phi(d_i) = \begin{cases} 1 & \text{gdy } SP_{k_*}^{d_i} \geq SP_{k_j}^{d_i} \ \forall j \\ 0 & \text{wpp} \end{cases}$$

k_* jest prawdziwą kategorią dokumentu d_i

Podejście II (silna poprawność)

$$P2 = 100 * \frac{\sum_{i=1}^{M} p2(d_i)}{M}$$

$$p2(d) = 100 * \frac{p(d,k_*)}{\sum_{j}^{|K|} p(d,k_j)}$$

gdzie $p(d, k_j)$ przyjmuje wartość liczbową zależnie od tego jakim terminem lingwistycznym (patrz (4)) system określił przynależność dokumentu d do kategorii k_j. Dla terminów "nie", "lekko", "średnio", "mocno" i "bardzo mocno" przypisane są odpowiednio liczby 0,25,50, 75 i 100. Jak poprzednio, k_* oznacza prawdziwą kategorię dokumentu d, a M liczność zbioru $D=\{d_i\}$ dokumentów użytych do testowania.

Uwzględniając obydwa podejścia uzyskaliśmy następujące wyniki dla danych testowych z rozbiciem na kategorie:

Tabela 1 *Zestawienie wyników dla danych testowych*

	P1: prosta poprawność	**P2: silna poprawność**
Kategoria 1 (przyroda i ekologia)	90 %	79.5 %
Kategoria 2 (ekonomia i gospodarka)	90 %	69.75 %

Kategoria 3 (film i kino)	80 %	79.00 %
Kategoria 4 (komputery i Internet)	90 %	75.50 %
Kategoria 5 (motoryzacja)	100 %	66.80 %
Kategoria 6 (polityka, społeczeństwo i prawo)	70 %	57.50 %
Ogółem	87 %	71%

W zdecydowanej większości wypadków system TCAT wskazał najsilniejszy związek zawartości testowanych dokumentów z właściwą dla każdego z nich kategorią.

6. Uwagi końcowe

W pracy przedstawiliśmy koncepcję i implementację klasyfikatora dokumentów internetowych. Zaproponowane podejście odwołuje się zarówno do pojęć typowych dla statystyki jak i logiki rozmytej. Dalsze prace nad systemem podążać będą dwutorowo. Po pierwsze, chcemy dokładniej przetestować praktyczną przydatność systemu. Niewielkie modyfikacje mogą przekształcić go w rodzaj *inteligentnego agenta* samodzielnie przeglądającego Internet w poszukiwaniu dokumentów dotyczących zadanego, "wyuczonego" obszaru tematycznego. Po drugie, przeprowadzimy porównanie jego efektywności z innymi klasycznymi klasyfikatorami, poczynając od najprostszych takich jak np. tzw. "naive Bayes". Jednocześnie chcemy zaproponować dodatkowe wskaźniki poprawności klasyfikacji dokumentów. Przykładowo, stopień jednoznaczność klasyfikacji dokumentu można utożsamić ze wskaźnikiem specyficzności zbioru rozmytego zaproponowanym przez Yagera (za Dubois i Prade(1987)). Z drugiej strony, należałoby uwzględnić również dość powszechne zjawisko, że tematyka rozpoznawanych dokumentów faktycznie obejmuje, w różnym stopniu, więcej niż jedną kategorię tematyczną.

Literatura

Dubois, D. i H. Prade (1987) Properties of measures of information in evidence and possibility theories. *Fuzzy Sets and Systems*, 24, 161-182.

Kacprzyk, J. (1986) *Zbiory rozmyte w analizie systemowej*, PWN, Warszawa.

Ławcewicz, K.A. (2001) *System automatycznego rozpoznawania tematyki dokumentów tekstowych*. Praca inżynierska. Wyższa Szkoła Informatyki Stosowanej i Zarządzania, Warszawa.

Kuncheva, L.I. (2000) *Fuzzy Classifier Design*. Physica-Verlag, Heidelberg New York.

Nowakowski, M. (2001) *PHP4 i MySQL dla webmastera*, Translator.

van Rijsbergen C. J. (1979) *Information Retrieval*. Butterworths, London.

Rutkowska, D., M. Piliński i L. Rutkowski (1999) *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, PWN, Warszawa.

ISSN 0208-8028
ISBN 83-85847-59-6

W celu uzyskania bliższych informacji i zakupu dodatkowych egzemplarzy
prosimy o kontakt z Instytutem Badań Systemowych PAN
ul. Newelska 6, 01-447 Warszawa
tel. 837-35-78 w. 241 e-mail: bibliote@ibspan.waw.pl