POLSKA AKADEMIA NAUK
Instytut Badań Systemowych

ROZMYTOŚĆ I BIPOLARNOŚĆ W INTELIGENTNYM WYSZUKIWANIU INFORMACJI

Sławomir Zadrożny

Warszawa 2013

POLSKA AKADEMIA NAUK
INSTYTUT BADAŃ SYSTEMOWYCH

Seria: BADANIA SYSTEMOWE
Tom 73

Redaktor naukowy:
Prof. dr hab. inż. Jakub Gutenbaum

Warszawa 2013

Rada redakcyjna serii: BADANIA SYSTEMOWE

POLSKA AKADEMIA NAUK
INSTYTUT BADAŃ SYSTEMOWYCH

Sławomir Zadrożny

ROZMYTOŚĆ I BIPOLARNOŚĆ W INTELIGENTNYM WYSZUKIWANIU INFORMACJI

Warszawa 2013

Autorzy:

Dr hab. Sławomir Zadrożny
Instytut Badań Systemowych Polskiej Akademii Nauk
ul. Newelska 6, 01-447 Warszawa
Slawomir.Zadrozny@ibspan.waw.pl

Recenzenci:

dr hab. inż. Maciej Krawczak
dr Marek Reformat

Skład: Aneta M. Pielak

Wydawca:

Instytut Badań Systemowych
Polskiej Akademii Nauk
Newelska 6, 01-447 Warszawa
www.ibspan.waw.pl

ISSN 0208-8029
ISBN 83-894-7551-0

Rozdział 7

Modele IR oparte na logice rozmytej

7.1 Rys historyczny

Bardzo szybko dostrzeżono walory logiki rozmytej jako narzędzia do modelowania procesu wyszukiwania informacji tekstowej. Wśród wczesnych prac należy wymienić prace z przełomu lat 70. i 80. Radeckiego [186], Booksteina [18], Buella [55] czy Krafta i Buella [145]. W dużym stopniu bazują one na naturalnej interpretacji w terminach logiki rozmytej wielu pojęć z zakresu IR, takich jak stopień ważności słowa kluczowego w reprezentacji dokumentu czy stopień dopasowania dokumentu względem zapytania. Późniejsze prace, por. np. [19, 22, 29, 121, 144, 27, 130, 49, 255] dotyczą już bardziej bezpośrednio modelowania nieprecyzyjności i niepewności informacji w procesie wyszukiwania, poprzez użycie terminów lingwistycznych jako stopni ważności słów kluczowych, czy zastosowanie elastycznych operatorów agregacji do cząstkowych stopni spełnienia.

Często punktem wyjścia dla proponowanych "modeli rozmytych" jest klasyczny model logiczny, a w szczególności model boolowski omawiany w p. 6.1.1. Najbardziej oczywistym uogólnieniem modelu boolowskiego z użyciem logiki rozmytej jest zastąpienie klasycznej logiki dwuwartościowej *logiką wielowartościową*. Przy takim podejściu cała struktura modelu boolowskiego może zostać zachowana, a wymagana jest jedynie reinterpretacja jego elementów w terminach logiki wielowartościowej; por. (2.56) na s. 28. W naturalny sposób uzyskuje się tą drogą interpretację ważności słów kluczowych przy reprezentacji dokumentów oraz relewantności dokumentu względem zapytania jako pojęć o charakterze stopniowalnym.

Poszczególnym słowom kluczowym przypisuje się w dokumentach stopnie ważności wyrażone liczbami z przedziału [0,1]. Tak przypisane stopnie ważności definiują wartościowanie (lub zbiór wartościowań) zmiennych zdaniowych skojarzonych z poszczególnymi słowami kluczowymi. Dokument d utożsamiamy więc teraz z wartościowaniem właściwym dla logiki wielowartościowej (2.56). Zapytania w tym przypadku nadal, jak w modelu klasycznym, reprezentowane są przez dowolne formuły q nad alfabetem określonym przez zbiór zmiennych zdaniowych S. Stopień dopasowania dokumentu d i zapytania q jest nadal określony jako *wartość prawdy* formuły q przy wartościowaniu ω_d (por. opis modelu boolowskiego na s. 133). Wartość prawdy takiej formuły jest oczywiście również liczbą z przedziału [0,1]. Obliczanie wartości prawdy q sprowadza się do odpowiedniej interpretacji spójników logicznych. Dla różnych wariantów logiki wielowartościowej interpretacja ta może się różnić. Standardowa ich interpretacja w logice rozmytej polega na zastosowaniu operatorów minimum i maksimum dla, odpowiednio, koniunkcji i alternatywy (por. wzory (2.73)-(2.81)).

Kolejnym postulowanym rozszerzeniem klasycznego modelu logicznego jest wprowadzenie stopni ważności słów kluczowych również w zapytaniach. Motywacja jest tu podobna jak w przypadku reprezentacji dokumentów: użytkownik może przywiązywać różną wagę do wystąpienia poszczególnych słów kluczowych w poszukiwanych dokumentach. Uwzględnienie wag w zapytaniach wymaga wyjścia poza składnię logiki klasycznej, czy wielowartościowej. Pewną próbę formalizacji tego podejścia można znaleźć w pracy [130] oraz w pokrewnym kontekście wyszukiwania informacji w bazach danych w [249]. Semantyka stopni ważności w zapytaniach również nie jest jednoznaczna. W literaturze przyjęto rozważać trzy ich semantyczne interpretacje [19, 21, 186, 56]:

1. *względna ważność*: jeśli waga słowa kluczowego w zapytaniu jest wysoka, to jego wystąpienie z wysoką wagą jest wymagane w dokumencie, aby uznać ten dokument za spełniający zapytanie w wysokim stopniu,

2. *waga idealna*: słowo kluczowe powinno mieć w dokumencie przypisany stopień ważności podobny do stopnia mu przypisanego w zapytaniu,

3. *wartość progowa*: słowo kluczowe powinno mieć w dokumencie przypisany stopień ważności co najmniej równy stopniowi ważności przypisanemu mu w zapytaniu.

Poszczególne interpretacje różnią się istotnie między sobą i przyjęcie jednej z nich, ma decydujące znaczenie dla określenia wynikowego uporządkowania dokumentów.

Wszystkie te interpretacje można zamodelować z użyciem wielowartościowego rachunku zdań opisanego na s. 29. W tym celu stosuje się stałe logiczne (2.59). Rozważmy zapytania w formie koniunkcji słów kluczowych (zmiennych zdaniowych im odpowiadających):

$$q = s_1 \wedge \ldots \wedge s_{|q|} \tag{7.1}$$

gdzie $|q|$ oznacza długość zapytania czyli liczbę występujących w nim słów kluczowych.

Załóżmy, że każdemu słowu kluczowemu k_j w zapytaniu q przypisano pewien stopień ważności w_j; $w_j \in [0,1]$ i $\max_j w_j = 1$. Oznaczmy takie rozszerzone zapytanie jako q_w. Można wtedy zastosować ogólne podejście Dubois i Prade'a [92] (por. również Yager [225]) do modelowania ważności elementów poddawanych procesowi agregacji i wyrazić je z użyciem formuły analogicznej do wzoru (2.144) ze s. 52. Zapytanie q_w wyraża się wtedy formułą ϕ_w:

$$q_w \mapsto \phi_w \equiv (\overline{w_1} \rightarrow s_1) \wedge (\overline{w_2} \rightarrow s_2) \wedge \ldots \wedge (\overline{w_{|q|}} \rightarrow s_{|q|}) \tag{7.2}$$

gdzie $\overline{w_i}$ oznacza stałą logiczną odpowiadającą wartości stopnia ważności $w_i \in [0,1]$.

Jak poprzednio, stopień dopasowania dokumentu d względem zapytania q_w obliczany jest jako wartość logiczna formuły ϕ_w przy wartosciowaniu określonym przez dokument d. Zależnie od interpretacji spójnika logicznego implikacji "$\rightarrow$" (por. def. 2.24) uzyskuje się różne semantyki stopni ważności słów kluczowych w zapytaniu.

Dla operatora implikacji $\mathtt{i}_{S-M}$ (Kleene-Dienesa) (2.88) stopnie ważności zachowują sie zgodnie z interpretacją *względnej ważności*. Słowa kluczowe k_j, reprezentowane w formule ϕ_w (7.2) przez składową $\overline{w_j} \rightarrow s_j$ dla $w_j = 0$, nie odgrywają żadnej roli przy określaniu stopnia dopasowania zapytania: taka składowa ma przy dowolnym wartościowaniu ω_d wartość logiczną $\max(1-0, \omega_d(s_j)) = 1$ i nie ma wpływu na łączną wartość koniunkcji występującej w (7.2) (por. własność operatora t-normy $\mathtt{t}(x,1) = x$ na stronie 33). Słowa kluczowe k_j, reprezentowane w formule ϕ_w (7.2) przez składową $\overline{w_j} \rightarrow s_j$ dla $w_j = 1$, są z kolei krytyczne dla wartości stopnia dopasowania zapytania: taka składowa ma przy dowolnym wartościowaniu ω_d wartość logiczną $\max(1-1, \omega_d(s_j)) = \omega_d(s_j)$, więc jeśli słowo kluczowe k_j ma w dokumencie d przypisany stopień ważności

równy 0, to łączna wartość koniunkcji występującej w (7.2) jest również równa 0. Pośrednie stopnie ważności słowa kluczowego k_j w zapytaniu skutkują proporcjonalnym wpływem wystąpienia tego słowa kluczowego z odpowiednio wysokim stopniem ważnosci w dokumencie na wynik dopasowania, co jest zgodne z semantyką względnej wazności.

Dla operatorów implikacji $\mathtt{i}_{R-M}$ (Gödla) (2.91) i $\mathtt{i}_{R-\Pi}$ (Goguena) (2.92) uzyskuje sie interpretację stopni ważnosci jako wartości progowych. Jeśli słowo kluczowe k_j występuje w dokumencie ze stopniem ważności wyższym lub równym stopniowi przypisanemu temu słowu kluczowemu w zapytaniu w_j, to stopień spełnienia składowej $\overline{w_j} \to s_j$ formuły (7.2) wynosi 1. W przeciwnym wypadku stopień spełnienia składowej i w konsekwencji stopień spełnienia całej formuły (7.2) jest mniejszy niż 1. Operator implikacji Goguena zapewnia ciągłe przejście od stopnia dopasowania równego 1 do niższych wartości, w przeciwieństwie do operatora implikacji Gödla.

Wzór (7.2) nie obejmuje interpretacji stopni ważności słów kluczowych w zapytaniu jako *wag idealnych*. W [130] zaproponowano w tym celu pewną modyfikację tej formuły do postaci:

$$\phi_w = (\overline{w_1} \leftrightarrow s_1) \wedge (\overline{w_2} \leftrightarrow s_2) \wedge \ldots \wedge (\overline{w_{|q|}} \leftrightarrow s_{|q|}) \tag{7.3}$$

gdzie spójnik logiczny równoważności "$\leftrightarrow$" interpretowany jest z użyciem następującego operatora (jak wcześniej, używamy tego samego symbolu "$\leftrightarrow$" do oznaczenia spójnika i operatora definiującego jego semantykę zamiast symbolu `e` stosowanego we wzorze (2.103)):

$$x \leftrightarrow y = \min(x \to y, y \to x)$$

i $\to$ oznacza operator implikacji $\mathtt{i}_{R-\Pi}$ (Goguena).

Łatwo stwierdzić, że:

$$x \leftrightarrow y = \begin{cases} 1 & \text{jeśli } x = y \\ x/y & \text{jeśli } y > x \\ y/x & \text{jeśli } y < x \end{cases}$$

co potwierdza, że otrzymuje się w ten sposób adekwatną interpretację stopni ważności jako *wag idealnych*.

Przykład 7.1 ilustruje wspomniane wcześniej różnice pomiędzy poszczególnymi interpretacjami stopni ważności słów kluczowych w zapytaniu.

Przykład 7.1. *Rozważmy trzy dokumenty indeksowane z użyciem trzech słów kluczowych w następujący sposób:*

	k_1	k_2	k_3
d_1	1.0	0.8	0.7
d_2	0.5	1.0	0.6
d_3	0.3	1.0	0.45

i zapytanie q, w którym poszczególnym słowom kluczowym przypisano następujące stopnie ważności:

$$k_1 \mapsto 0.3, \quad k_2 \mapsto 1.0, \quad k_3 \mapsto 0.5$$

Zależnie od przyjętej interpretacji stopni ważności uzyskuje sie następujące wyniki dopasowania poszczególnych dokumentów d_i względem zapytania q:

interpretacja stopni ważności	wynikowa kolejność dokumentów
względna ważność	$d_1 \succ d_2 \succ d_3$
wartość progowa	$d_2 \succ d_3 \succ d_1$
waga idealna	$d_3 \succ d_2 \succ d_1$

gdzie $\succ$ oznacza relację określającą porządek dokumentów w odpowiedzi na zapytanie q na podstawie wyliczonych stopni dopasowania.

Pokazane wyniki potwierdzają znaczenie przyjętej interpretacji stopni ważności.

Należy stwierdzić, że postać formuły ϕ_w (7.2) jest właściwa wyłącznie dla zapytania w postaci koniunkcji słów kluczowych, takiego jak pokazane w (7.1). Zapytanie w postaci alternatywy słów kluczowych:

$$q = s_1 \vee \ldots \vee s_{|q|} \tag{7.4}$$

można reprezentować z użyciem podobnej formuły przez analogię do operatorów opisanych wzorem (2.145):

$$q_w \mapsto \phi_w = (\neg\overline{w_1} \rightarrow_c s_1) \vee (\neg\overline{w_2} \rightarrow_c s_2) \vee \ldots \vee (\neg\overline{w_{|q|}} \rightarrow_c s_{|q|}) \tag{7.5}$$

gdzie $\rightarrow_c$ jest spójnikiem koimplikacji (por. def. 2.25 na s. 37).

7.2 Model z jawną reprezentacją nieprecyzyjności i niepewności

Większość z omawianych modeli wyszukiwania informacji tekstowej wymaga określenia wartości pewnych parametrów związanych między innymi z reprezentacją dokumentów i zapytań. Wartości te bardzo często dobierane są do pewnego stopnia arbitralnie i z natury są *nieprecyzyjne*. Jednocześnie ich dobór, jak i cały proces wyszukiwania są obciążone *niepewnością*. Źródłem niepewności może być na przykład niepełna wiarygodność eksperta lub algorytmu odpowiedzialnego za indeksowanie zbioru dokumentów. Poszczególne modele w różnym stopniu uwzględniają tę *niedoskonałość* informacji dostępnej w procesie wyszukiwania. W sposób jawny czynią to modele probabilistyczne, ale również one nie uwzględniają łącznie różnych form tej niedoskonałości. W niniejszym punkcie omówimy pokrótce model oparty na logice rozmytej, który pozwala uwzględnić nieprecyzyjny i niepewny charakter stopni ważności.

W omawianym modelu [257, 171, 172, 256, 258] ważność słowa kluczowego k_i traktowana jest jako zmienna lingwistyczna X_i (por. def. 2.27):

$$(X_i, T(X_i), U, G, M) \tag{7.6}$$

gdzie $T(X_i)$= {*około* 0.1, *około* 0.2,..., *ważne*, *bardzo ważne*,...}, $U = [0, 1]$. Należy wspomnieć, że koncepcja potraktowania stopnia ważności jako zmiennej lingwistycznej została najpierw zaproponowana w pracach Bordogni i Pasi ze współautorami [20, 22, 144]. W opisywanym tu modelu koncepcja ta została znacznie rozszerzona i sformalizowana.

Reprezentacja dokumentów. Znaczenie słowa kluczowego k_i dla reprezentacji dokumentu wyraża się z użyciem *wyrażeń z kwalifikatorami pewności* (por. (2.124)):

$$X_i \text{ jest } A_i,\ \alpha \tag{7.7}$$

gdzie $A_i \in T(X_i)$ oznacza termin lingwistyczny.

W omawianym modelu dokument reprezentowany będzie zazwyczaj jako koniunkcja wyrażeń typu (7.7), czyli w postaci:

$$d : (X_1 \text{ jest } A_1,\ \alpha_1) \wedge (X_2 \text{ jest } A_2,\ \alpha_2) \wedge \ldots \wedge (X_n \text{ jest } A_n,\ \alpha_n) \tag{7.8}$$

gdzie X_i jest zmienną lingwistyczną odpowiadającą ważności słowa kluczowego t_i, A_i jest terminem lingwistycznym typu *około 0.8* czy *ważny*, zaś $\alpha \in [0, 1]$ jest liczbą określającą stopień pewności co do tak określonej ważności słowa kluczowego. Zakłada się, że terminy lingwistyczne

typu *około 0.8* będą zazwyczaj stosowane przy automatycznym indeksowaniu dokumentów, zaś terminy takie, jak *bardzo ważny* mogą znaleźć zastosowanie przy indeksowaniu ręcznym.

Poszczególne terminy lingwistyczne modelowane są przez zbiory rozmyte, określone na przedziale [0,1]. W trakcie obliczania stopnia dopasowania dokumentu względem zapytania, wyrażenia postaci (7.7) są przekształacane do postaci bez kwalifikatora konieczności, zgodnie ze wzorem (2.127). Określają one więc rozkłady możliwości stopni ważności poszczególnych słów kluczowych (por. p. 2.1.2). Łączny rozkład możliwości określony na przestrzeni będącej iloczynem kartezjańskim przedziałów [0,1] definiuje wzór (2.27). W związku z tym jako reprezentację dokumentu można przyjąć wyrażenie:

$$X \text{ jest } A \tag{7.9}$$

przy czym zmienna lingwistyczna X jest *zmienną złożoną*, zaś zbiór rozmyty modelujący A jest określony na $[0,1]^n$, gdzie n jest liczbą rozważanych słów kluczowych. Jednocześnie wyrażenie to będziemy utożsamiać ze związanym z nim rozkładem możliwości π_d:

$$\begin{aligned} \pi_d &: [0,1]^n \longrightarrow [0,1] \\ \pi_d(x) &= \mu_A(x) \end{aligned} \tag{7.10}$$

gdzie A jest określone jak w (7.9).

Zakłada się [257, 171], że istnieje zestaw *szablonów* funkcji przynależności terminów lingwistycznych, które będą stosowane do określania stopni ważności poszczególnych słów kluczowych przy indeksowaniu dokumentów. Jeden z szablonów ma postać trókątnej funkcji przynależności $\mu_A : [0,1] \longrightarrow [0,1]$ (por. rys. 2.1), którą można zapisać w następujący sposób:

$$\mu_A(x) = 1 - |c_d - x| \tag{7.11}$$

gdzie $c_d \in [0,1]$ stanowi parametr tego szablonu i oznacza stopień ważności o najwyższym stopniu przynależności do zbioru A równym 1. Wartość c_d może zostać dobrana na przykład z użyciem jednego ze znormalizowanych schematów ważenia słów kluczowych stosowanych w modelu wektorowym, takich jak tf x IDF (por. p. 6.1.2). W ogólniejszym wypadku, jeśli wyrażona ma być ograniczona pewność co do stopnia ważności ($\alpha < 1$ we wzorze (7.7)), to funkcja przynależności (7.11) przyjmuje postać:

$$\mu_A(x) = \max(1 - |c_d - x|, 1 - \alpha) \tag{7.12}$$

i szablon w takiej ogólniejszej postaci ma dwa parametry: c_d i α.

Drugi z szablonów funkcji przynależności inspirowany jest *logiką posybilistyczną* (por. p. 2.2.1). Przyjmuje się, że ważność słów kluczowych ma charakter binarny, ale może być określana z różnym stopniem przekonania. Tak więc zbiory A_i i B_i we wzorach (7.7) i (7.14) są klasycznymi zbiorami jednoelementowymi $\{0\}$ lub $\{1\}$, rozważanymi w przestrzeni dwuelementowej $\{0, 1\}$. Po uwzględnieniu ograniczonej pewności α i zastosowaniu transformacji opisanej wzorem (2.127), stają się one w ogólnym przypadku zbiorami rozmytymi określonymi w przestrzeni $\{0, 1\}$. Zakłada się więc, że do opisu dokumentu wybiera się tylko takie słowa kluczowe, które uznane są za ważne dla jego reprezentacji (w stopniu 1). Jednocześnie określa się minimalny stopień przekonania w tym względzie. Innymi słowy przyjmuje się, że jest *całkowicie możliwe* (*stopień możliwości wynosi 1*), że słowo kluczowe jest *ważne*, ale jednocześnie dopuszcza się niezerowy stopień możliwości, że jest ono *nieważne*. Szablon ten można opisać następującą funkcją przynależności:

$$\mu_{A_i}(x) = \begin{cases} 1 - \alpha & \text{dla } x = 0 \\ 1 & \text{dla } x = 1 \end{cases} \tag{7.13}$$

Parametrem jest tu α, odpowiadające α z wzoru (7.7), określające dolne ograniczenie przekonania co do ważności słowa kluczowego k_i.

Reprezentacja zapytań Zapytanie q, podobnie jak w klasycznym modelu logicznym, reprezentowane jest w postaci dowolnej kombinacji elementów podstawowych - w przypadku klasycznego modelu są to zmienne zdaniowe s_i, zaś w przypadku omawianego tu modelu są to wyrażenia lingwistyczne typu (7.7). Można je zapisać następująco:

$$X_i \text{ jest } B_i,\ \alpha \tag{7.14}$$

a zagregowaną postać zapytania można, podobnie jak dla dokumentów, wyrazić jako:

$$X \text{ jest } B \tag{7.15}$$

Ważność słów kluczowych w zapytaniu może być zatem określona z użyciem terminów lingwistycznych takich, jak *ważne*, *bardzo ważne* itp., kóre modelowane są z użyciem zbiorów rozmytych określonych na przedziale [0,1]. Pierwszy z tych terminów może być modelowany z użyciem

na przykład zbioru rozmytego o następującej funkcji przynależności:

$$\mu_{ważny}(x) = x \tag{7.16}$$

Wzór (7.14) przewiduje, podobnie jak w wypadku dokumentów, wyrażenie ograniczonej pewności co do podanego stopnia ważności. W ten sposób użytkownik uzyskuje możliwość wyrażenia stopni nie tylko w postaci nieprecyzyjnej (*ważne*, *bardzo ważne*), ale również z jawnym wskazaniem stopnia swego przekonania co do poprawności tej wartości.

Wzór (7.15) przedstawia finalną formę reprezentacji zapytania jako zbioru rozmytego $B \in \mathcal{F}([0,1]^n)$. Konstruuje się go na podstawie podanej jako zapytanie kombinacji wyrażeń (7.14) stosując przekształcenie (2.127) i standardowe operacje na zbiorach rozmytych opisane w rozdziale 2.

Podobnie jak w wypadku dokumentów przyjmuje się, że stopnie ważności słów kluczowych w zapytaniach określane są na podstawie analogicznych szablonów. Poza dwoma typami szablonów wspomnianych przy okazji reprezentacji dokumentów, rozważa się dla zapytań jeszcze trzeci szablon. Przyjmuje się w nim, że ważność słów kluczowych w zapytaniu określona jest terminem lingwistycznym *ważne* wraz z wyrażeniem stopnia pewności α. Zakłada się, że użytkownik podaje w zapytaniu wyłącznie słowa kluczowe, które uważa za *ważne* dla reprezentacji jego potrzeb informacyjnych, ale jednocześnie użytkownik może nie być całkowicie pewien, czy wybrane przez niego słowa kluczowe faktycznie dobrze charakteryzują jego potrzeby informacyjne. Stąd bazowa postać wyrażenia (7.14) użytego do określenia ważności słowa kluczowego k_i w zapytaniu to:

$$X_i \text{ jest } \textit{ważne}$$

przy czym:

$$\mu_{ważne}(x) = x$$

czyli im wyższa jest liczbowa ocena ważności słowa kluczowego, tym bardziej jest ona zgodna z oceną użytkownika. Na to nakłada się niepewność użytkownika, wyrażona parametrem α – im jego wartość jest wyższa, tym bardziej użytkownik jest pewny swojego wyboru. Ostatecznie, funkcja przynależności zbioru rozmytego B_i reprezentującego wartość stopnia ważności ma następującą postać:

$$\mu_{B_i}(x) = \max(x, 1-\alpha) \tag{7.17}$$

Szczegóły dotyczące doboru par szablonów reprezentujących stopnie ważności słów kluczowych w dokumentach i w zapytaniach można znaleźć w [257, 171].

Ocena relewantności Ocena relewantności dokumentu d względem zapytania q w omawianym modelu ma charakter analogiczny do podejścia stosowanego w klasycznym modelu logicznym i polega na określeniu stopnia prawdziwości wyrażenia (7.15) reprezentującego zapytanie, przy założeniu że prawdziwe jest wyrażenie (7.9) reprezentujące dokument:

$$X \text{ jest } A \text{ jest prawdziwe } \longmapsto X \text{ jest } B \text{ jest ?} \tag{7.18}$$

Zgodnie z zasadami operowania na wyrażeniach lingwistycznych z kwalifikatorami (por. s. 45) ten stopień prawdziwości przyjmuje postać rozmytej wartość prawdy τ:

$$\mu_\tau(y) = \begin{cases} \sup_x\{\mu_A(x) \mid \mu_B(x) = y\} & \text{jeśli } \{x \mid \mu_B(x) = y\} \neq \emptyset \\ 0 & \text{wpp} \end{cases} \tag{7.19}$$

Rozmyta wartość prawdy wyraża jednocześnie stopniowalną naturę pojęcia relewantności oraz niepewność co do relewantności [93]. Wyraża ona ocenę na ile możliwe jest, że dokument jest relewantny względem zapytania w określonym stopniu.

Posługiwanie się rozmytymi wartościami prawdy może być jednak niewygodne. W związku z tym używa się miar możliwości i konieczności (2.29)-(2.33), które traktuje się w tym kontekście jako pewne uproszczenie rozmytych wartości prawdy. Jako ocenę relewantności dokumentu d reprezentowanego przez rozkład możliwości (7.10) skojarzony z wyrażeniem X jest A (7.9), względem zapytania q reprezentowanego przez wyrażenie X jest B (7.15), przyjmuje się więc parę wartości miar możliwości i konieczności:

$$\Pi_A(B) = \sup_{x\in[0,1]^m} \min(\pi_A(x), \mu_B(x)) \tag{7.20}$$

$$N_A(B) = \inf_{x\in[0,1]^m} \max(1 - \pi_A(x), \mu_B(x)) \tag{7.21}$$

gdzie m jest liczbą słów kluczowych użytych w zapytaniu.

Ostatecznym wynikiem wyszukiwania zgodnie z omawianym modelem jest uporządkowanie dokumentów według oceny ich relewantności względem zapytania. W związku z tym, że ocena relewantności wyrażona jest przez parę liczb $(\Pi_A(B), N_A(B))$ należy przyjąć pewien sposób uporządkowania dokumentów, na przykład porządek leksykograficzny.

Dla najpopularniejszej postaci dokumentów i zapytań w formie koniunkcji wyrażeń, odpowiednio, (7.7) i (7.14), przy przyjęciu założenia o *nieinteraktywności* pomiędzy stopniami ważności słów kluczowych,

otrzymuje się następującą uproszczoną postać wzorów (7.20)-(7.21):

$$\Pi_A(B) = \Pi_{A_1 \times \ldots \times A_m}(B_1 \times \ldots \times B_m) = \min(\Pi_{A_1}(B_1), \ldots, \Pi_{A_m}(B_m)) \tag{7.22}$$

$$N_A(B) = N_{A_1 \times \ldots \times A_m}(B_1 \times \ldots \times B_m) = \min(N_{A_1}(B_1), \ldots, N_{A_m}(B_m)) \tag{7.23}$$

gdzie każde wyrażenie X_i jest A_i w reprezentacji dokumentu generuje rozkład możliwości π_{A_i} i związaną z nim miarę możliwości Π_{A_i}, które łącznie określają miarę Π_A.

Dla poszczególnych par omawianych wcześniej szablonów funkcji przynależności zbiorów rozmytych A_i i B_i reprezentujących ważność słowa kluczowego w dokumencie i w zapytaniu, wyprowadza się znacznie uproszczone wersje wzorów (7.22)-(7.23). Szczegóły można znaleźć w [257, 171].

ISSN 0208-8029
ISBN 83-894-7551-0

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK
tel.: (+48) 22 3810246 / 22 3810277 / 22 3810241 / 22 3810273
e-mail: biblioteka@ibspan.waw.pl