ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 157, 2002

Adam Niewiadomski[*]
Piotr S. Szczepaniak[**]

INTUICJONISTYCZNE RELACJE ROZMYTE W PRZESZUKIWANIU DOMEN E-COMMERCE

Poniższy referat prezentuje koncepcję wykorzystania intuicjonistycznych relacji rozmytych do przeszukiwania domen e-commerce. Zaprezentowana została miara podobieństwa słów i fragmentów tekstów, zakorzeniona w teorii zbiorów rozmytych Zadeha [7]. Następnie na dwóch przykładach wyjaśnione zostały korzyści płynące z zastosowania nowej miary podobieństwa w handlu elektronicznym.

This paper focuses on application of intuitionistic fuzzy relations applied to services available within the e-commerce domains. Firstly, concepts for comparison of natural language words and sentences rooted in the theory of fuzzy sets, and in the concept of intuitionistic fuzzy relation in particular, are presented. Then, on two examples of application to the e-commerce domain, the aspect of the user-friendliness of the approach is demonstrated.

Wprowadzenie

W projektowaniu i zarządzaniu domenami e-commerce ścierają się zasadniczo dwie przeciwne tendencje: potrzeba standaryzacji i – z drugiej strony – wymaganie coraz to większej elastyczności w obsłudze klienta. Jedynym rozsądnym wyjściem z tej sytuacji wydaje się być częściowa standaryzacja lub, innymi słowy, ograniczona elastyczność. Szczególnie widoczne jest to w domenach z zakresu Business-to-Customer (B-to-C), gdyż pozostałe dwa umowne zakresy działalności, Business-to-Administration (B-to-A) and Business-to-Business

[*] Instytut Informatyki, Politechnika Łódzka

[**] Instytut Informatyki, Politechnika Łódzka
Instytut Badań Systemowych, Polska Akademia Nauk

(B-to-B), posługują się zazwyczaj terminologią oraz procedurami łatwymi do sformalizowania.

Poniższa praca skupia się na się na możliwościach wykorzystania nowej koncepcji określania podobieństwa fragmentów tekstów sporządzonych w języku naturalnym w utworzeniu „przyjaznego użytkownikowi" interfejsu. Metoda ta oparta jest na intuicjonistycznych relacjach rozmytych. Umożliwia ona klientowi stosunkowo sprawne orientowanie się pośród olbrzymiej ilości różnych produktów oferowanych poprzez różne usługi w Internecie.

"Rozmytość" w sensie Zadeha [7] jest wystarczająco silnym aparatem matematycznym, aby przeprowadzić porównywanie tekstów częściowo standaryzowanych. W poniższym opracowaniu została ona jednak rozszerzona o elementy teorii intuicjonistycznych zbiorów rozmytych [1, 2] z myślą o jak najwierniejszym spełnianiu ludzkich intuicji językowych.

Podstawowe definicje

Zbiory rozmyte i intuicjonistyczne zbiory rozmyte

Pojęcie zbioru rozmytego pochodzi od Zadeha [7] i zasadza się na rozszerzeniu zbioru wartości funkcji charakterystycznej do całego przedziału [0,1]. Koncepcja ta umożliwia formalizację potocznej konstrukcji językowej, iż jakiś element posiada daną własność „w pewnym stopniu", np. amfibia nie jest tym samym co samochód, chociaż *w pewnym stopniu* posiada jego cechy". Formalnie zbiór rozmyty A w niepustej przestrzeni X przedstawiamy jako zbiór par uporządkowanych

$$A = \{<x, \mu_A(x)>: x \in X\}, \tag{2.1}$$

gdzie: $\mu_A: X \rightarrow [0,1]$ – funkcja przynależności do zbioru rozmytego A.

W 1984 roku Atanassov wystąpił z propozycją rozszerzenia zbioru rozmytego do *intuitionistic fuzzy set* – intuicjonistycznego zbioru rozmytego [1, 2]. Do pary element, stopień przynależności dodana została liczba z zakresu [0,1] oznaczająca *stopień nieprzynależności danego elementu do zbioru.* Analogicznie jak w (2.1), intuicjonistycznym zbiorem rozmytym B w niepustej przestrzeni X nazywa się zbiór trójek uporządkowanych

$$B = \{<x, \mu_B(x), \nu_B(x)>: x \in X\}, \tag{2.2}$$

gdzie: μ_B: $X \rightarrow [0,1]$, ν_B: $X \rightarrow [0,1]$ – odpowiednio funkcje przynależności i nieprzynależności do intuicjonistycznego zbioru rozmytego B, spełniające warunek

$$0 \leq \mu_B(x) + \nu_B(x) \leq 1 \qquad \text{dla każdego } x \in X \tag{2.3}$$

Różnica pomiędzy jednością a sumą wartości funkcji μ_A i ν_A dla dowolnego $x \in X$ interpretowana jest jako „stopień niepewności" (ang. *hesitancy degree, hesitancy margin*) zwany także „indeksem intuicjonistycznym dla elementu x w A" (ang. *intuitionistic index of x in A*). Wskaźnik ten obliczany jest ze wzoru

$$\pi_A(x) = 1-\mu_A(x)-\nu_A(x) \qquad \forall x \in X. \tag{2.4}$$

Oczywiście

$$0 \leq \pi_A(x) \leq 1 \qquad \forall x \in X. \tag{2.5}$$

Intuicjonistyczne relacje rozmyte

Intuicjonistyczna relacja rozmyta na produkcie niepustych przestrzeni X i Y może być zdefiniowana na bazie (2.2), zob. [6], jako zbiór trójek uporządkowanych postaci

$$R = \{<(x, y), \mu_R(x, y), \nu_R(y, x)>: x \in X, y \in Y\}, \tag{2.6}$$

gdzie: μ_R: $X \times Y \rightarrow [0,1]$, ν_R: $X \times Y \rightarrow [0,1]$ – jak w (2.2), (2.3). Zazwyczaj liczba μ interpretowana jest jako „siła powiązania" elementów x i y, zaś liczba ν – jako stopień ich zróżnicowania.

Intuicjonistyczna relacja rozmyta, która jest

a) zwrotna na X wtedy i tylko wtedy, gdy

$$\mu_R(x, x) = 1 \qquad \forall x \in X, \tag{2.7}$$

oraz

b) symetryczna na X wtedy i tylko wtedy, gdy

$$\mu_R(x, y) = \mu_R(y, x) \wedge \nu_R(x, y) = \nu_R(y, x) \qquad \forall x, y \in X, \tag{2.8}$$

zwana jest „relacją sąsiedztwa" i może być interpretowana jako model nieprzechodniej relacji podobieństwa.

Podobieństwo tekstów

Porównywanie słów

Intuicjonistyczne relacje rozmyte opisane w sekcji 2 posłużyć mogą do określania podobieństwa fragmentów tekstów języka naturalnego. Określmy w tym celu na S – zbiorze wszystkich słów – intuicjonistyczną relację rozmytą RS postaci:

$$RS = \{(\langle s_1, s_2 \rangle, \mu_{RS}(s_1, s_2), \nu_{RS}(s_1, s_2)) : s_1, s_2 \in S\} \tag{3.1}$$

o funkcji przynależności μ_{RS}: $S \times S \rightarrow [0,1]$ danej wzorem:

$$\mu_{RS}(s_1, s_2) = \frac{2}{(N^2 + N)} \sum_{i=1}^{N(s_1)} \sum_{j=1}^{N(s_1)-i+1} h(i, j) \qquad \forall s_1, s_2 \in S, \tag{3.2}$$

gdzie: $h(i, j)=1$, jeżeli podciąg i-elementowy liter występujący w słowie s_1 i rozpoczynający się od j-tego miejsca w słowie s_1 występuje conajmniej raz w słowie s_2 (w przeciwnym przypadku $h(i, j)=0$); $N(s_1)$, $N(s_2)$ – liczby liter w słowach s_1 i s_2, odpowiednio, zwane dalej „długościami słów"; $N = \max\{N(s_1), N(s_2)\}$;

oraz o funkcji nieprzynależności ν_{RS}: $S \times S \rightarrow [0,1]$ danej wzorem

$$\nu_{RS}(s_1, s_2) = (\mu_{RS}(s_1, s_2))^{0,5} - \mu_{RS}(s_1, s_2). \tag{3.3}$$

Przykład porównywania słów przy pomocy (3.2) (3.3) opisany został szczegółowo w [4]. Warto wspomnieć, że wybór funkcji nieprzynależności jest sprawą subiektywną, podobnie jak i dla funkcji przynależności.

Porównywanie zdań

Relacje podobne do opisanych w sekcji 3.1 można zastosować także do porównywania zdań. Ustalmy na Z – zbiorze wszystkich zdań – intuicjonistyczną relację rozmytą RZ postaci

$$RZ = \{(<z_1, z_2>, \mu_{RZ}(z_1, z_2), \nu_{RZ}(z_1, z_2)): z_1, z_2 \in Z\} \tag{3.4}$$

o funkcji przynależności μ_{RZ}: $X \rightarrow [0,1]$ danej wzorem

$$\mu_{RZ}(z_1, z_2) = \frac{1}{N} \sum_{i=1}^{N(z_1)} \max_{j \in \{1,\ldots,N(z_2)\}} \mu_{RS}(s_i, s_j) \tag{3.5}$$

gdzie: s_i – i-te słowo w zdaniu z_1;
s_j – j-te słowo w zdaniu z_2;
μ_{RS} –funkcja podobieństwa słów dana w (2.2);
$N(z_1)$, $N(z_2)$ – liczba słów odpowiednio w zdaniach z_1 i z_2 („długości zdań");
$N = \max\{N(z_1), N(z_2)\}$;

oraz o funkcji nieprzynależności ν_{RZ}: $Z \times Z \rightarrow [0,1]$ danej wzorem:

$$\nu_{RZ}(z_1, z_2) = (\mu_{RZ}(z_1, z_2))^{0,5} - \mu_{RZ}(z_1, z_2) . \tag{3.6}$$

Uwaga: Funkcja μ_{RZ}(3.5) nie uwzględnia różnic, które występują pomiędzy zdaniami złożonymi z tych samych, lecz ustawionych w różnej kolejności wyrazów. Umożliwia to porównywanie zwłaszcza zbiorów a nie tylko ciągów słów (np. zbiorów słów kluczowych).

Przykłady porównywania zdań – patrz [4].

Obliczony na podstawie (2.4) indeks intuicjonistyczny dla dowolnej pary słów lub zdań może nieść informację niezwykle ważną z punktu widzenia użytkownika Internetu – ma on interpretację „stopnia niepewności" dla dokonanego porównania (lub też stopnia niezgodności wyniku z intuicjami ludzkimi). Innymi słowy jest to określenie marginesu błędu dla porównania.

Zastosowanie miar podobieństwa w domenach *e-commerce*

Księgarnie internetowe

Przykładowe zastosowanie metody określania podobieństwa tekstów dotyczyć może interfejsu bazy danych księgarni internetowej (ewentualnie internetowej bazy danych dużej biblioteki lub dowolnego innego zbioru danych, w którym duża ilość informacji przechowywana jest w postaci czysto tekstowej).

Zbiór opisów wszystkich książek (czyli tekstów „częściowo standaryzowanych") dostępnych w sprzedaży w danej księgarni internetowej może być zgromadzony w operacyjnej tekstowej bazie danych o przykładowej postaci przedstawionej w tabeli 1

Tabela 1 Przykładowa operacyjna baza danych księgarni internetowej.

Tytuł	Autor	Tytuł serii
Paragraf 22	Joseph Heller	Klub Interesującej Książki
Namaluj to	Joseph Heller	Biblioteka Mistrzów
Cyberiada	Stanisław Lem	-

Załóżmy, że użytkownik poszukujący książki kieruje do bazy danych zapytanie postaci: {„Paragraf 22", „J. Heller", „Interesująca Książka"}. Porównywanie kolejnych pól formularza z polami poszczególnych rekordów (wierszy) w bazie danych (tabeli) przebiegać może według wzoru (4.1):

$$\mu(z,r_j) = \frac{\sum_{i=1}^{n} w_i \cdot \mu_i(z_i, r_{ji})}{\sum_{i=1}^{n} w_i} \tag{4.1}$$

gdzie: z – zapytanie użytkownika;
z_i – i-te pole zapytania użytkownika (w tym przypadku: 1 – Tytuł, 2 – Autor, 3 – tytuł serii);
r_j – j-ty rekord w tabeli bazy danych;
r_{ji} – i-te pole j-tego rekordu bazy (analogicznie jak w opisie z_i);
w_i – waga i-tego pola rekordu oraz zapytania;
μ_i – miara podobieństwa zawartości i-tych pól formularza i zapytania (3.2) lub (3.5).

Porównanie podanego zapytania użytkownika oraz pierwszego rekordu-wiersza tabeli 1 r_1 przedstawia się następująco:

$z = \{ z_1 =$ Paragraf 22, $z_2 =$ Józef Heller, $z_3 =$ Interesująca książka $\}$

$r_1 = \{ r_{11} =$ Paragraf 22, $r_{12} =$ Joseph Heller, $r_{13} =$ Klub Interesującej Książki $\}$

$w_1 = w_2 = w_3 = 1$,

a zatem:

$$\mu(z, r_1) = \frac{\sum_{i=1}^{3} w_i \cdot \mu_i(z_i, r_{ji})}{3} = 0{,}831 \qquad (4.2)$$

przy czym: $\mu_1(z_1, r_{11}) = 1{,}0$;
$\mu_2(z_2, r_{12}) = 0{,}722$;
$\mu_3(z_3, r_{13}) = 0{,}772$.

Stopień nieprzynależności dla tak opisanej pary argumentów relacji wynosi – *via* (3.3) lub (3.6) – 0,167.

Odpowiedzią na zapytanie użytkownika może być np. ranking dziesięciu rekordów z bazy najbardziej podobnych do zapytania.

Frequently Asked Questions

Autorzy pracy [3] problematykę *Frequently Asked Questions* stanowiący tytuł tej sekcji. „Często zadawane pytania" umieszczają w obszarze zainteresowania systemów *Case-Based Reasoning*. Jest to podejście trafne z tego względu, iż każde pytanie czy wątpliwość pochodzącą od użytkownika potraktować można jako oddzielny przypadek-wektor i zapisać go przy pomocy zbioru atrybutów, następnie znaleźć dlań przypadki podobne i w końcu zaaplikować rozwiązanie, zob. [5]. Dotychczasowe rozwiązania opierają się jednak o struktury z góry przewidzianych pytań.

Proponowane unowocześnienie polega na umożliwieniu użytkownikowi zadawania pytań w zupełnie dowolny – jak najbardziej naturalny – sposób. Potraktujmy zatem pytanie od użytkownika jako zdanie naturalne, niekoniecznie sformułowane według ścisłych zasad trybu pytającego, np.:

„gdzie szukać informacji o HP DeskJet 690 ?"

albo:

„drukarka nie drukuje, brak tonera"

Integralną częścią systemu *FAQ* winna być baza przypadków (analogia do *CBR*), w której jednak zawarte są rekordy nie z danymi numerycznymi bądź symbolicznymi, ale zwykłe zdania sformułowane przy pomocy naturalnego (nawet potocznego) języka, wyrażające najczęstsze pytania i wątpliwości odnośnie tematyki danego serwisu WWW. Ponieważ baza danych przypadków może również przechowywać informacje dotyczące rozwiązań stosowanych w wypadku pojawienia się danego problemu, można nadać tym rozwiązaniom formę odnośnika URL do strony WWW, na której znajdują się stosowne instruktaże. Szczegółowo rzecz ujmując pojedynczy zapis w bazie *FAQ* powinien mieć postać n-tki uporządkowanej:

$$\{ <s_1, ..., s_{n-1}>, URL\}$$

gdzie: $s_1, ..., s_{n-1}$ – słowa naturalne opisujące przypadek (problem);
URL – odnośnik (hiperłącze) do strony WWW.

Przykładowa baza często zadawanych pytań może mieć postać jak w tabeli 2:

Tabela 4.2. Przykładowa baza pytań systemu *FAQ*.

ID	Pytanie	URL
1	Blady, niewyraźny wydruk, zamazane litery.	www.mysite.com/toner.htm
2	Jak ustawić marginesy w dokumencie ?	www.mysite.com/wydruk.htm
3	Instalacja drukarki Hewlett Packard	www.hp.com/first_install.asp

System *FAQ* może pytanie pobrane od użytkownika porównywać z przypadkami w bazie na podstawie wzoru:

$$\mu(\{s_1,..,s_n\},\{p_1,..,p_k\}) = \frac{1}{n}\sum_{i=1}^{n} \max_{j\in\{1,..,k\}} g(s_i, p_j) \qquad (4.3)$$

gdzie: $s_1, ..., s_n$ – zbiór słów opisujących pytanie użytkownika;
$p_1, ..., p_k$ – zbiór słów opisujących problem zawarty w bazie;
μ_i – miara podobieństwa słów i zapytania (3.2)

Przykładowe porównanie zapytania o postaci { BLADE LITERY, ZAMAZANY WYDRUK } z polem „Pytanie" w pierwszym wierszu tabeli 4.2 przedstawia się następująco:

$$\mu(\{s_1,..,s_4\},\{p_1,..,p_5\}) = \frac{1}{4}\sum_{i=1}^{4} \max_{j\in\{1,...,5\}} \mu_{DIL(RS)}(s_i, p_j) = 0{,}80 \qquad (4.4)$$

Stopień nieprzynależności dla tak opisanej pary argumentów relacji wynosi – *via* (3.3) lub (3.6) – 0,094.

Generalnie, udzielanie odpowiedzi na pytania użytkowników w systemie *FAQ* przy zastosowaniu wyżej opisanych miar podobieństwa przebiegać może według następującego algorytmu:

1. Pobierz od użytkownika zapytanie – zbiór słów { s_1, ... , s_n }.
2. Pobierz od użytkownika wymagany stopień dokładności porównania μ_0.
3. Porównaj zapytanie z polami „Pytanie" w kolejnych rekordach w k-elementowej bazie pytań, odnotowując stopnie podobieństwa μ_1, μ_2, ..., μ_k.
4. **IF** przynajmniej jeden rekord w bazie jest podobny w stopniu $\mu_i \geq \mu_0$
   THEN przejdź do 5.

   ELSE i. prześlij pytanie *via* e-mail do administratora serwisu
   ii. przejdź do 6 .
5. Zaproponuj odnośniki *URL* do (np. trzech) stron z odpowiedziami.
6. **STOP**

Podsumowanie

Korzyści płynące ze stosowania nowych miar podobieństwa wyrażeń językowych w systemie *FAQ* są następujące:

- możliwość formułowania przez użytkownika pytań bez konieczności używania terminów technicznych i specjalistycznego słownictwa;
- używanie swobodnego, zbliżonego do naturalnego języka w trakcie wymiany informacji;
- wygodna i stwarzająca „poczucie bezpieczeństwa" forma kierowania pytań ze strony klienta – interfejs na stronie WWW (formularze);
- odpowiedzi od systemu w przyjaznej dla użytkownika formie – strony WWW o nieskomplikowane strukturze (podobieństwo do tradycyjnych systemów pomocy HELP).

Opisana metoda obsługi klienta *via* Internet jest w dużym stopniu oparta na stylu działania systemów *CBR*. Ma ona duże znaczenie dla handlu elektronicznego, a w szczególności dla usług typu *Business-to-Customer*.

Wnioski

Z obserwacji autorów wynika, iż stosowane dotąd metody „inteligentnego" przeszukiwania zbiorów dokumentów tekstowych opierały się zazwyczaj na metodzie ścisłego porównywania fragmentów tekstów. Zastosowanie teorii intuicjonistycznych zbiorów rozmytych może usprawnić metody wyszukiwania informacji tekstowych.

Co więcej, porównywanie rozmyte sprawia, iż metoda ta nie jest czuła na błędy gramatyczne i ortograficzne; nie istnieje zatem możliwość wygenerowania błędu porównania tylko z powodu prostej pomyłki przy wpisywaniu zapytania. Słowa różniące się od siebie zaledwie jedną literą sklasyfikowane zostaną jako bardzo podobne, niemalże identyczne.

Cechy te pozwalają autorom żywić nadzieję na liczne internetowe i bazodanowe zastosowania metody porównywania tekstów.

Źródła

1. Atanassov K. (1984). *Intuitionistic fuzzy sets*. Fuzzy Sets and Systems, 20 (1986), ss. 87-96.
2. Atanassov K. (1999).*Intuitionistic fuzzy sets, Theory and Applications*. Springer Verlag.
3. Lenz M., Hübner A., Kunze M. (1998). *Textual CBR* . In: Lenz M., Bartsch-Spörl B., Burkhard H.-D., Wess S. (Eds.) (1998): *Case-Based Reasoning Technology. From Foundations to Applications*. Springer Verlag, Berlin, Heidelberg.
4. Niewiadomski A., Szczepaniak P.S.: *Intuicjonistyczne relacje rozmyte w przybliżonym porównywaniu tekstów*. W: „Zbiory Rozmyte i Ich Zastosowania", Praca zbiorowa pod redakcja Jana Chojcana i Jacka Łęskiego. Silesian University Press 2001.
5. Pal S.K., Dillon T.S., Yeung D.S. (Eds.) (2001): *Soft Computing in Case Based Reasoning*. Springer-Verlag, London.
6. Pedrycz W., Gomide F. (1998): *An Introduction to Fuzzy Sets; Analysis and Design*. A Bradford Book, The MIT Press, Cambridge, Massachusetts and London, England.
7. Zadeh L.A. (1965). *Fuzzy sets*. Information and Control, 8, pp. 338-353.