INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 11

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2009

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 11

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2009

Wykaz opiniodawców artykułów zamieszczonych w niniejszym tomie:

prof. dr hab. inż. Jerzy HOŁUBIEC
dr inż. Lech KRUŚ
doc. dr hab. inż. Wiesław KRAJEWSKI
doc. dr hab. Jacek MALINOWSKI
dr inż. Edward MICHALEWSKI
prof. dr Adam SKOREK
dr hab. Ryszard SMARZEWSKI
prof. dr hab. inż. Andrzej STRASZAK
dr Dominik ŚLĘZAK
prof. dr hab. inż. Stanisław WALUKIEWICZ
doc. dr hab. Sławomir ZADROŻNY

Warszawa 2009

ISBN 9788389475220

Druk: Zakład Poligraficzny Jerzy Kosiński, Warszawa

INTELIGENTNE DOPASOWANIE DANYCH PRZY UŻYCIU TEORII ZBIORÓW ROZMYTYCH W SYSTEMACH PRZETWARZANIA DANYCH

Łukasz Sosnowski
Studia Doktoranckie IBS PAN

Standaryzacja danych w systemach przetwarzania przeważnie opiera się o metody słownikowe z dopasowaniem pewnym, takim które łatwo można zaimplementować poprzez złączenia wewnętrzne w systemach relacyjnych baz danych. Ten artykuł ma przybliżyć sposób na dopasowanie wynikające z podobieństwa obiektów, ze znalezienia najbardziej podobnego elementu. Do realizacji zostanie użyta teoria zbiorów rozmytych jak również techniki ograniczające błędne dopasowania, np. funkcja aktywująca, zbiór wyjątków.

***Słowa kluczowe**: zbiory rozmyte, relacje rozmyte, standaryzacja danych, systemy przetwarzania danych, metryka Levenshteina*

Wprowadzenie

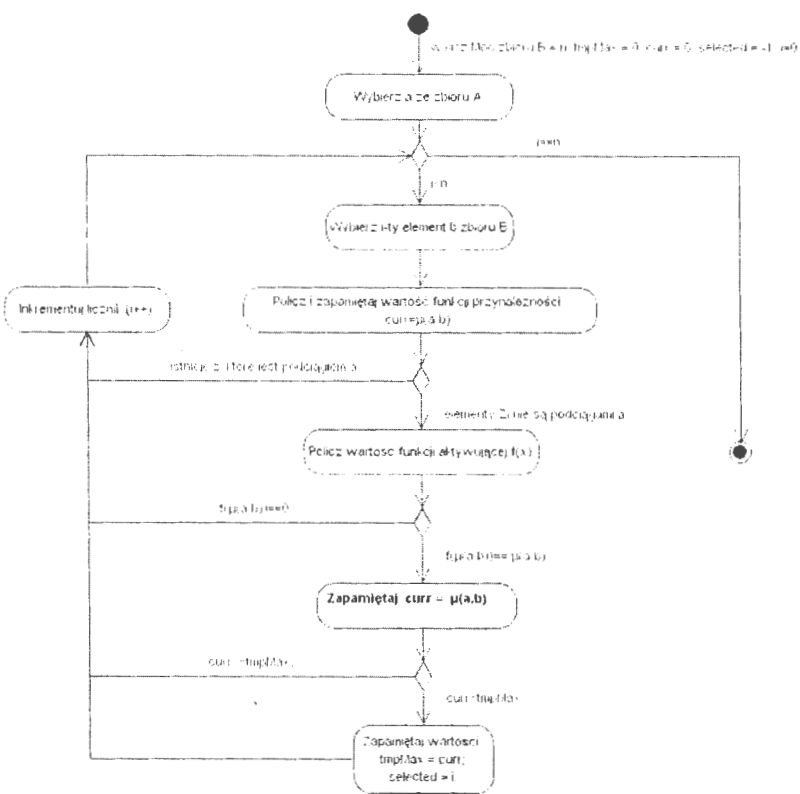
Systemy przetwarzania w ogólności operują na danych. Dane te mogą pochodzić z wielu różnorodnych źródeł, mogą mieć różne formy zapisu pomimo identycznego znaczenia. Systemy takie muszą dokonywać standaryzacji danych wejściowych, aby wszystkie dane były zapisane w tej samej ontologii, tzn. aby poszczególne pojęcia znaczyły to samo dla danych z różnych źródeł.

Narzędzia służące do wykonywania wstępnego przetwarzania danych (a zatem i standaryzacji) nazywane są narzędziami ETL (ang. Extract Transform Load) [Microsoft Corporation (2000), Agosta L. (2000)]. Dlatego też wszędzie tam gdzie napotkamy ETL, możemy spodziewać się trudności związanych ze standaryzacją wykonywaną tylko przy pomocy komparatorów opartych na sprawdzaniu identyczności.

W wielu przypadkach mamy do czynienia z danymi bardzo podobnymi lecz jednak nie identycznymi. Dla algorytmów wspomnianych wcześniej takie dane stanowią problem. W dalszej części artykułu przybliżę jeden ze sposobów rozwiązujących tego typu zagadnienia.

1. Opis rozwiązania

Do rozwiązania postawionego problemu potrzebna będzie metryka za pomocą której można dokonać pomiaru podobieństwa dwóch ciągów znaków, jednego ze zbioru standaryzowanego A, a drugiego ze zbioru referencyjnego B. Zakłada się że elementami B są niepuste ciągi znaków. Jako metrykę określającą różnicę między ciągami znaków przyjmujemy metrykę Levenshteina [Levenshtein V., (1966)]. W praktyce należy zbadać stopień podobieństwa danego $a \in A$ elementu do każdego $b \in B$.



Rys. 1. Diagram aktywności algorytmu inteligentnej standaryzacji

Zdefiniujmy relację rozmytą R = {(μ(a,b),(a,b))}, [Kacprzyk J. (2001)] dla funkcji przynależności μ(a,b) : A x B → [0,1] zdefiniowanej jako

$$\mu(a,b) = 1 - \frac{D(a,b)}{\max(L(a),L(b))}$$

Wzór 1 - funkcja przynależności

dla max(L(a),L(b)) ≠ 0, gdzie: D(a,b) to odległość Levensteina, zaś L(a) oraz L(b) są długościami ciągów znaków a i b.

W celu uniknięcia zbyt słabych rozwiązań wprowadza się funkcję aktywacji, która dana jest wzorem:

$$f(x) = \begin{cases} 0 & dla\ x < p \\ x & dla\ x \geq p \end{cases}, \text{ gdzie } x \in [0,1] \text{ i } p \in [0,1]$$

Wzór 2 - funkcja aktywacji

Ustalając „p" definiujemy przedział [p,1], z którego akceptujemy rozwiązania. W celu wyboru najlepszego rozwiązania dla danego elementu $a \in A$ zastosujemy funkcję maximum do wszystkich wyliczonych wartości funkcji przynależności do relacji rozmytej R, spełniających $\mu(a,b) \geq p$.

W przypadkach szczególnych gdy mamy dodatkowe informacje dotyczące przetwarzanych danych można wprowadzić tzw. słownik wyjątków, tzn. rodzinę zbiorów Z_i elementów powiązanych z danym elementem referencyjnym $b \in B$, gdzie „i" jest i'tym elementem zbioru B. Elementy zbioru Z_i to zakazane ciągi znaków, których nie może zawierać $a \in A$, który w wyniku standaryzacji miałby zostać dopasowany z $b \in B$. Oznacza to, że nie istnieje takie $z \in Z_i$, że dla $a \in A$ i $b \in B$ oraz μ(a,b) maksymalnego spośród wszystkich spełniona jest relacja R_w, gdzie relacja R_w = {(a,z)} jest spełniona wtedy i tylko wtedy, gdy ciąg znaków „z" jest podciągiem „a".

Przykład 1.

Załóżmy, że są dane zbiory A={a_1}, B={b_1, b_2} oraz Z_1={z_{11}} i Z_2={z_{21}}, takie że:

a_1 - „Witamina C, 100mg, FirmaX, tabletki", b_1 - „Vitaminum E 100 mg FirmaX, tabletki", b_2 - „Vitaminum C 100 mg FirmaX tabletki", z_{11} - „ C ", z_{21} - „ E " oraz dany jest współczynnik p = 0.6.

Rozpatrzmy podobieństwo elementu a_1 z elementami b_1 i b_2, przy elementach zbiorów wyjątków z_{11} i z_{21}. W pierwszej kolejności liczymy odległość Levenshteina dla odpowiednich par i stąd mamy: $D(a_1, b_1) = 7$, $D(a_1, b_2) = 7$, następnie liczymy maksimum z długości ciągów znaków co daje odpowiednio: $\max(L(a_1), L(b_1)) = 35$ oraz $\max(L(a_1), L(b_2)) = 35$. Podstawiając do wzoru funkcji przynaeżności otrzymujemy wartości $\mu(a_1, b_1) = 0.8$ oraz $\mu(a_1, b_2) = 0.8$. Jak widać pod względem stopnia przynależności (podobieństwa) obie pary są równoważne. Pozostał nam jeszcze sprawdzenie kryterium wyjątków. Zatem dla pary (a_1, b_1), sprawdzamy czy a_1 nie zawiera podciągu znaków zdefiniowanego przez z_{11} (czy nie spełnia relacji R_w). W tym przypadku widać, że relacja jest spełniona, a więc element b_1 musi zostać odrzucony dla elementu a_1. Pozostaje jeszcze sprawdzić spełnienie realcji R_w dla pary (a_1, b_2). W tym przypadku widać, iż ciąg znaków reprezentowany przez a_1 nie zawiera ciągu reprezentowanego przez z_{21} („ E "). Zgodnie z podanym algorytmem w dalszej części bierzemy pod uwagę jedynie parę (a_1, b_2). Na koniec sprawdzamy wartość funkcji aktywacji.
Dla podanej pary wartość funkcji wynosi 0.8 co mieści się w przedziale [0.6,1], który wyznacza parametr „p". Pobieramy maksymalną wartość funkcji przynależności. W tym przypadku jest to 0.8. Zatem w podanym przykładzie wynikiem standaryzacji przy pomocy podanego algorytmu jest dopasowanie elementu a_1 z elementem b_2.

Jak wynika z powyższego opisu, algorytm może nie zwrócić żadnego wyniku, jeśli badane pary elementów „a" i „b" nie spełnią wymaganych warunków (funkcja aktywacji, wyjątki) oraz może zwrócić również wiele wartości, jeśli jest wiele elementów które w jednakowym stopniu są podobne z badanym elementem „a".

Przedstawiony algorytm ma zastosowanie w systemach, funkcjonujących na rynku komercyjnym. Przykładem może być system monitoringu rynku farmaceutycznego [Sosnowski Ł. – praca magisterska] funkcjonujący dla transakcji hurtowych produktów farmaceutycznych. Innymi przykładami zastosowań mogą być różnego rodzaju systemy moderacyjne, których zadaniem jest określenie czy dany

obiekt narusza zdefiniowane reguły. Przykładowym przypadkiem użycia, może być internetowe forum, które przekazuje wpisany tekst przez internautę do systemu moderacji, gdzie na podstawie zastosowania opisywanego algorytmu określane jest czy i w jakim stopniu przekazany tekst zawiera frazy zakazane.

Podsumowanie

Niniejszy artykuł przedstawia jedno z rozwiązań suboptymalnych przedstawionego zagadnienia. Definiuje drogę, która niewątpliwie może być modyfikowana i optymalizowana w zależności od konkretnego zastosowania. Jednym z pól do optymalizacji jest zamiana globalnego współczynnika „p" dla funkcji aktywującej na rodzinę współczynników i przyporządkowaniu każdemu elementowi z B indywidualnego współczynnika aktywacji „p".

Innym wariantem poszukiwań lepszych rozwiązań jest zamiana metryki na metrykę bardziej adekwatną do rozwiązania konkretnego problemu.

Literatura

[1]. Kacprzyk J. (2001): Wieloetapowe sterowanie rozmyte. **2**, s. 39-67.

[2]. Rutkowski L. (2005): Metody i techniki sztucznej inteligencji. **4**, s. 52-94.

[3]. Levenshtein V. (1966): Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**, s.707–710.

[4]. Sosnowski Ł. – praca magisterska „Przetwarzanie danych w rozproszonym systemie monitoringu rynku farmaceutycznego".

[5]. Microsoft Corporation (2000): Dokumentacja techniczna do MS SQL SERVER 2000.

[6]. Agosta L. (2000): The Essential Guide to Data Warehousing.

ISBN 9788389475220